Performance of prediction models on survival outcomes of colorectal cancer with surgical resection

Surgical Oncology 29 (2019) 196–202

Contents lists available at ScienceDirect

Surgical Oncology

journal homepage: www.elsevier.com/locate/suronc

Performance of prediction models on survival outcomes of colorectal cancer with surgical resection: A systematic review and meta-analysis

Yazhou He[a,b], Yuhan Ong[c], Xue Li[a], Farhat VN. Din[b,d], Ewan Brown[e], Maria Timofeeva[b,d], Ziqiang Wang[f], Susan M. Farrington[b,d], Harry Campbell[a], Malcolm G. Dunlop[b,d], Evropi Theodoratou[a,d,*]

[a] Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
[b] Colon Cancer Genetics Group, Medical Research Council Human Genetics Unit, Medical Research Council Institute of Genetics & Molecular Medicine, Western General Hospital, The University of Edinburgh, Edinburgh, UK
[c] Western General Hospital, Edinburgh, UK
[d] Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[e] Edinburgh Cancer Centre NHS Lothian, Edinburgh, UK
[f] Department of Gastrointestinal Surgery, West China Hospital, Sichuan University, Chengdu, 610041, PR China

ARTICLE INFO

Keywords:
Colorectal cancer
Survival
Surgery
Prediction model
Systematic review

ABSTRACT

Prediction models allow accurate estimate of individualized prognosis. Increasing numbers of models on survival of CRC patients with surgical resection are being published. However, their performance and potential clinical utility have been unclear. A systematic search in MEDLINE and Embase databases (until 9th April 2018) was performed. Original model development studies and external validation studies predicting any survival outcomes from CRC (follow-up ≥1 year after surgery) were included. We conducted random-effects meta-analyses in external validation studies to estimate the performance of each model. A total of 83 original prediction models and 52 separate external validation studies were identified. We identified five models (Basingstoke score, Fong score, Nordinger score, Peritoneal Surface Disease Severity Score and Valentini nomogram) that were validated in at least two external datasets with a median summarized C-statistic of 0.67 (range: 0.57–0.74). These models can potentially assist clinical decision-making. Besides developing new models, future research should also focus on validating existing prediction models and investigating their real-word impact and cost-effectiveness for CRC prognosis in clinical practice.

1. Introduction

Colorectal cancer (CRC) is responsible for 8.5% of deaths attributed to cancer worldwide [1]. The overall 5-year survival of CRC varies from 50% to 81% even within stage II CRC patients [2]. This within-stage variation can be explained to some extent by a wide range of other established prognostic factors such as carcinoembryonic antigen (CEA) [3]. Although surgery is the mainstay treatment modality, prognostic modelling integrating these factors may help optimize individualized clinical decision-making on targeting adjuvant treatment to those at most risk of relapsing and who may respond better to certain treatment modalities [4], so as to minimize the potential harms of overtreatment. Over the past decades, numerous statistical prediction models have been developed, incorporating various variables such as demographic [5], genetic [6] and clinic-pathological [5] factors. However, their performance, reliability and clinical validity have been unclear.

This systematic review aims to provide a comprehensive overview of current prognostication models for CRC patients undergoing surgical resection, to perform meta-analysis for models that have been validated in multiple datasets, as well as to evaluate the quality and performance of these model development and validation studies.

Abbreviations: CRC, colorectal cancer; CEA, carcinoembryonic antigen; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; OS, overall survival; DFS, disease-free survival; AUC, area under the receiver operating characteristic curve; CHARMS, CHecklist for critical Appraisal and data extraction or systematic Reviews of prediction Modelling Studies; EPV, events per variable; RFS, recurrence-free survival; CTC, circulating tumor cells; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; MSI, microsatellite instability; PCI, peritoneal cancer index

* Corresponding author. Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, United Kingdom.
E-mail address: E.Theodoratou@ed.ac.uk (E. Theodoratou).

https://doi.org/10.1016/j.suronc.2019.05.014
Received 12 March 2019; Received in revised form 7 May 2019; Accepted 18 May 2019

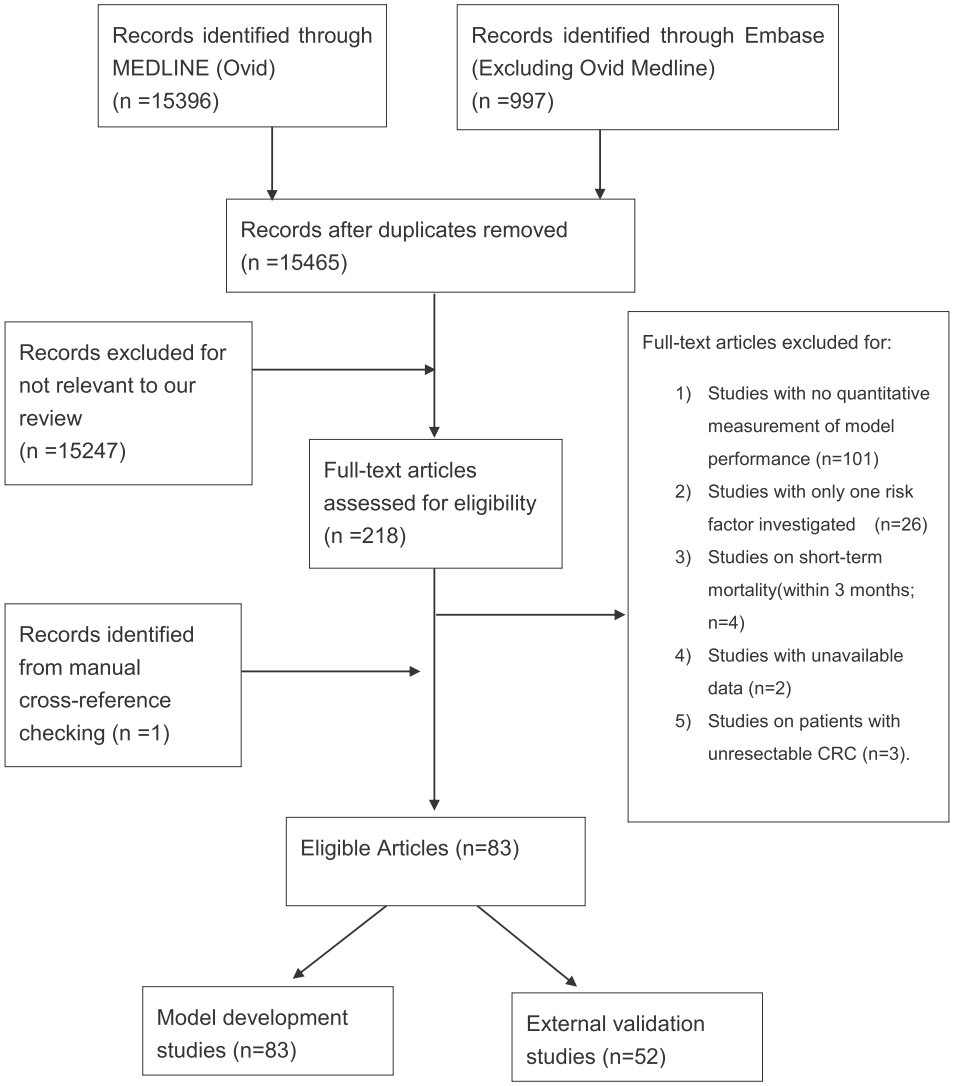

Fig. 1. Flow diagram of study selection.

2. Methods

2.1. Literature search and study selection

This study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [7]. A systematic search (limited to English and human studies) was performed in MEDLINE and Embase from inception to April 9th, 2018 to identify all relevant studies. Three sets of search terms, "Colorectal cancer", "Prognosis" and "Prediction model", were applied. The search strategy was formulated based on the search filter for identifying clinical prediction studies [8] and previous publications [9] (detailed search syntax presented in Supplementary Table S1). The reference list of each eligible article was also cross-checked.

We applied the following inclusion criteria: 1) studies developing or validating statistical model(s) based on time-to-event data to predict survival outcome (≥1 year) in CRC patients with surgical resection; 2) studies with at least two predictors; 3) studies that reported a quantitative measure of any aspect of model performance, such as metrics evaluating overall performance, discriminative ability and calibration. Conference abstracts, editorials and commentaries were excluded. Studies were also excluded if the prediction rule of the model was unavailable.

Two reviewers (YH and YO) screened the titles and abstracts independently. Potentially relevant articles were reviewed in full. Any disagreement was resolved by discussion, and a senior author (ET) was consulted if necessary.

2.2. Data extraction and critical appraisal

One reviewer (YH) extracted all relevant data (Table S2) following the guidelines of conducting systematic reviews of prediction model studies [10]. A second reviewer (ZW) verified the accuracy of the extracted data. Model performance metrics that evaluated discriminative ability (Harrell's C-statistic, also known as the area under the receiver operating characteristic curve (AUC)), calibration (e.g. calibration plot), and other metrics (e.g. R^2) were extracted. If a paper reported multiple models with different predictors or prediction rules, data were extracted separately for each model.

We appraised each model using the CHecklist for critical Appraisal and data extraction or systematic Reviews of prediction Modelling Studies (CHARMS) [11]. Based on this checklist, the risk of bias for each model was assessed following the criteria described in previous publications [12,13] which included six domains: 1) Participant selection; 2) Measurement and reporting of predictors; 3) Definition and measurement of the outcome; 4) Events per variable (EPV); 5) Attrition (loss to follow-up); 6) Data analysis. Details for the assessment rules are summarized in Supplementary Table S3. One reviewer (YH) appraised

Table 1
Summarized basic characteristics of included model development studies and external validation studies.

Variables	Model Development (*N* = 83)	External validation (*N* = 52)	*P*- values*
Participants (CRC patients)			
Cohort origin			
Europe	16 (19%)	23 (44%)	< 0.001
Asia	52 (63%)	19 (36%)	
America	15 (18%)	5 (10%)	
Other	0	5 (10%)	
CRC Stage			
I-III	45 (54%)	8 (15%)	< 0.001
IV	20 (24%)	44 (85%)	
Any	18 (22%)	0	
Tumor location			
Colon	15 (18%)	3 (6%)	0.005
Rectum	16 (19%)	3 (6%)	
Any	52 (63%)	46 (88%)	
Sample size			
< 500	28 (34%)	9 (17%)	0.04
> =500	55 (66%)	43 (83%)	
No. predictors			
< 5	30 (36%)	16 (31%)	0.28
5–10	50 (60%)	36 (69%)	
> 10	3 (4%)	0	
Outcome			
Overall survival	47 (57%)	24 (46%)	0.02
CRC-specific survival	13 (16%)	16 (31%)	
Disease-free survival	17 (20%)	11 (21%)	
Recurrence-free survival	7 (8%)	15 (29%)	
Other	10 (12%)	3 (6%)	
Model discrmination			
C statistic/AUC	76 (92%)	50 (96%)	0.35
Other[a]	4 (5%)	5 (10%)	
Model calibration			
Calibration plot	47 (57%)	7 (13%)	0.35
Hosmer-Lemeshow test	6 (7%)	0	
Internal validation			
Split sample	14 (17%)	NA	NA
Bootstrapping	13 (16%)	NA	
Cross validation	18 (22%)	NA	
Not reported	39 (47%)	NA	
Model presentation			
Nomogram	55 (66%)	NA	NA
Formula	21 (25%)	NA	
Other[b]	7 (8%)	NA	

*p-values for Chi-square test.
CRC, colorectal cancer; AUC, area under receiver's operating characteristic curve.
[a] Including D-statistic, sensitivity and specificity.
[b] Including score rule and decision tree.

all included studies. A second blinded reviewer (XL) evaluated a 25% random sample of all studies and cross-checked for any discrepancies.

2.3. Statistical analysis

Based on data availability, we performed meta-analyses of C-statistics across external validation studies that evaluated the same prediction model to estimate the overall discriminative performance for each model. The original dataset used to construct the model was not included in the meta-analysis to avoid inflated estimates [10]. We rescaled the C-statistic by applying a logit transformation [10]. The extracted 95% CI of a C-statistic was used to estimate its variance, and if this was not reported, the formula proposed by Debray et al. was used to approximate the 95% CI [10]. The C-statistic was considered statistically significant if the 95% CI excluded 0.5 [14]. Given the relatively small number of validation studies for each model and the inherent heterogeneity across external datasets with diverse populations and clinical settings, we adopted the restricted maximum likelihood (REML) estimation along with the Hartung-Knapp-Sidik-Jonkman (HKSJ) method under a random-effects model to estimate the pooled C-statistic and 95% CI [10, 15]. We also calculated the 95% prediction interval (PI) integrating the heterogeneity for the summarized C-statistic to indicate a possible range where a C-statistic of a future validation study may be located [16,17]. Due to unavailable data, we were unable to perform quantitative synthesis for other metrics evaluating model performance.

3. Results

3.1. Overview of eligible models

We obtained 15,465 unique records from the initial search. An additional validation study was identified from cross-checking the reference of eligible studies [18]. In total, 83 articles comprising 83 original model development studies and 52 separate external validation studies (Supplementary Table S4-S5) were included in this systematic review. The detailed study selection is summarized in Fig. 1.

Among the 83 model development studies, forty-five (54%) of these original models were based on early to locally advanced CRC (stage I-III) patients, and 24% (*N* = 20) focused on metastatic CRC. As for the predictors, these models included a median of 5 predictors (range 2–18). Age was the commonest predictor (*N* = 56, 67%). Other common predictors included CEA (*N* = 26, 31%), tumor grade or differentiation (*N* = 23, 28%), sex (*N* = 19, 23%), T stage (n = 16, 19%) and N stage (*N* = 16, 19%). Surgery type was adopted as a predictor in 13% (N = 11) of all models. The majority of the models (*N* = 73, 88%) were developed using Cox proportional hazards regression. Other methods included Weibull regression [19] and tree-based models [20]. The main outcome to be predicted was overall survival (OS) (*N* = 47, 57%), disease-free survival (DFS) (*N* = 17, 20%) and CRC specific survival (*N* = 13, 16%). The prediction time horizon varied from 1 year to 10 years, with 80% (*N* = 66) of the models reporting a 5-year prediction horizon. To adjust for potential overfitting, 44 (53%) models were internally validated using split-sample, bootstrapping or cross-validation. Twenty-eight (34%) models were validated in an external dataset by the same group of investigators. Only 11 (13%) models were externally validated by independent investigators. For model presentation, 55 of the 83 models (66%) were presented as nomograms, and the remainder as formulae, prediction rules, or web-based calculators. Detailed characteristics for each model development study are presented in Supplementary Table S4.

Among the 52 separate external validation studies (detailed characteristics in Table S5) and 22 (42%) of them validated original models identified in our systematic review. For the other 30 studies validating pre-existing models where the model performance was not evaluated in the initial model development reports, we evaluated their performance in these external validation studies. The study cohorts of external validation studies had significantly smaller sample size than model development studies (median 277 vs. 814, Mann-Whitney-Wilcoxon test: $P < 0.001$). The comparison of basic characteristics between model development and external validation studies are summarized in Table 1.

3.2. Critical appraisal

Risk of bias distribution of each domain for all included studies is

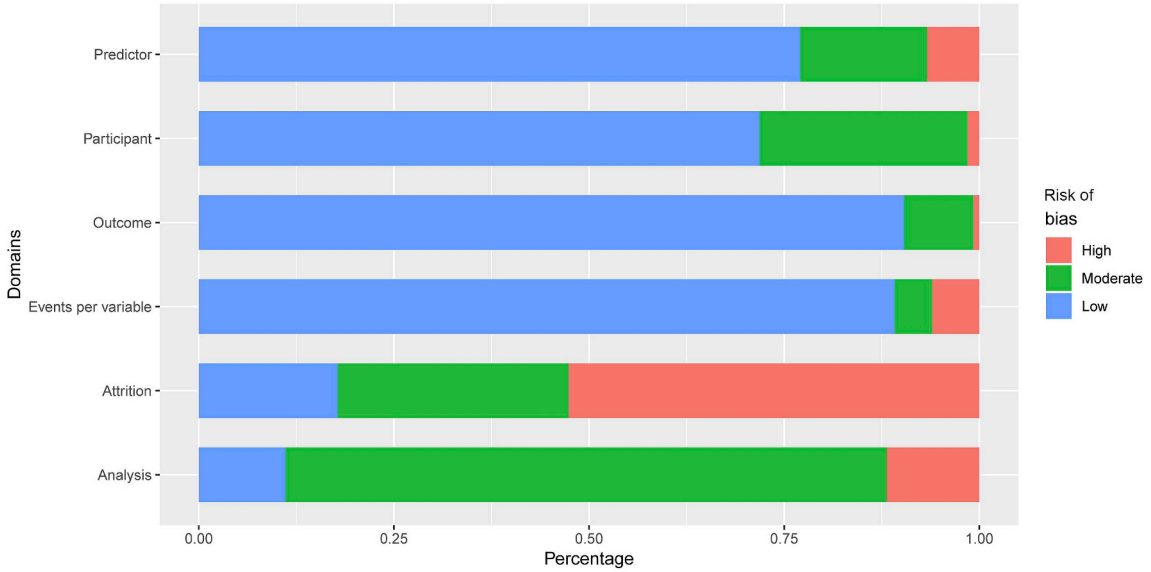

Fig. 2. Risk of bias assessment for six predefined domains for each included study. For participant selection, studies were rated as 'moderate' risk of bias if participants were possibly selected in a non-consecutive manner as this allowed for potential selection bias. We categorized studies to be high risk of bias if their selection criteria were inadequately described. With respect to the predictors, we assigned 'moderate' risk to studies where it was unclear whether the predictors were measured after the outcome was revealed, and 'high' risk to studies where the measurement of predictors was not clearly described. For the outcome domain, studies were assigned with 'moderate' risk when the measurement of CRC recurrence or progression was not clearly stated and 'high' risk if the whole follow-up procedure was not adequately described. For EPV, studies were scored as 'moderate' risk with an EPV between six and ten, and 'high' risk if their EPVs could not be calculated or were less than six. Studies were assigned with 'high' risk of attrition bias if insufficient information on loss to follow-up, and 'moderate' risk due to less than 20% of loss to follow-up. In relation to data analysis, studies were classified as 'moderate' risk given that either internal validation or missing data handling was not performed, and as 'high' risk if they neglected to report on either. The detailed classification rules are summarized in Table S3.

summarized in Fig. 2. Overall, only two models reported by one article were classified as low risk of bias for all domains [21]. The majority of the models were classified as 'low' risk for participant selection (N = 97, 72%), predictors (N = 104, 77%), outcome (N = 122, 90%), and EPV (N = 74, 89%). However, for dataset attrition, 71 studies (53%) were classified as 'high' risk, and with regard to data analysis, most studies (N = 104, 77%) were classified as 'moderate' risk of bias. The detailed scores of risk of bias for each domain are presented in Table S6 (model development studies) and Table S7 (external validation studies).

3.3. Model performance

The reported C-statistic for model development studies was significantly larger than external validation studies (median 0.73 vs. 0.66, Mann-Whitney-Wilcoxon test: $P < 0.001$).

We performed 15 meta-analysis for eight models (each single model can be applied to predict multiple survival outcomes) that had been externally validated at least twice: Basingstoke preoperative score, Fong score, Iwatsuki score, Memorial Sloan Katherine Cancer Center (MSKCC) nomogram, Nordinger score, Peritoneal Surface Disease Severity Score (PSDSS), Kanemitsu nomogram and Valentini nomogram. Their basic characteristics and estimate C-statistics from meta-analysis are presented in Fig. 3. We found significant discriminative ability for five models predicting six outcomes: the Basingstoke score (preoperative) predicting recurrence-free survival (RFS), the Fong score predicting RFS; the Nordinger score predicting RFS; the PSDSS score predicting OS; the Valentini nomogram predicting distant metastasis and OS. The pooled C-statistic of these six meta-analyses ranged from 0.57 to 0.74 (median 0.67). We were able to calculate the 95% PI for five meta-analyses (Fig. 3). The 95% PI of all the five models crossed 0.5, suggesting that a future validation study could possibly found a negative discriminative performance of that model.

The Fong score was the most commonly validated model. It utilized seven predictors (positive resection margin, extrahepatic lesion, lesion of regional lymph nodes for primary tumor, metastases-free period, number of metastases, the largest size of metastasis and CEA) to predict the RFS and OS of CRC patients with liver metastasis after curative resection. The meta-analysis found a significant C-statistic of 0.62 (95% CI: 0.55–0.68) for RFS prediction, but non-significant for OS 0.60 (C-statistic = 0.60 95% CI: 0.45–0.74). The strongest discriminative performance in relation to point estimates of C-statistics was observed for the Basingstoke preoperative score (C-statistic: 0.74, 95% CI: 0.52–0.88) for RFS and the Valentini nomogram (C-statistic: 0.74, 95% CI: 0.60–0.85) for distant metastasis.

For model calibration, 54 (40%) of all studies presented a calibration plot. Six studies employed the Hosmer-Lemeshow test to explore the overall goodness of model fit, and none of them reported a statistically significant departure of predicted outcomes from observed (Table S4). We were unable to quantitatively synthesize the model calibration because none of the studies reported the slope of the calibration plot or observed-to-expected events ratio.

4. Discussion

4.1. Interpretation and clinical application

To the best of our knowledge, this is the first systematic review and meta-analysis evaluating the performance of prediction models for survival outcomes of CRC patients with surgical resection. Prediction models can assist in estimating individualized prognosis, therefore guiding more precise treatment for CRC patients. In this study, we reviewed 83 original prediction models along with 52 external validation studies, and identified eight models that had been externally validated at least twice demonstrating significant discriminative performance.

With regard to predictors, most of the included models were based on common demographic and clinic-pathological factors. Genetic markers such as *RAS*, *BRAF* mutations and microsatellite instability (MSI) have already been recommended [22] to guide treatment for metastatic CRC. However, their predictive performance has barely been

Model	Participants	Outcome	No. Studies	Predictors	Pooled C-statistic (95%CI)	95% PI
Basingstoke score(preoperative)	CRC(liver metastasis+curative resection)	RFS	2	Lymph node status, tumor differentiation, CEA,no.liver metastasis,diameter of primary tumor,extrahepatic proliferation	**0.74(0.52-0.88)**	NA
Fong score	CRC(liver metastasis+curative resection)	RFS	4	Positive resection margin, extrahepatic lesion, lesion of regional lymph nodes for primary tumor,metastases-free period, no.metastases, the largest size of metastasis,CEA	**0.62(0.55-0.68)**	0.47-0.74
		OS	4		0.60(0.45-0.74)	0.32-0.82
		CSS	2		0.55(0.43-0.67)	NA
Iwatsuki score	CRC(liver metastasis+curative resection)	RFS	3	No. metastases,bilobar lesion,metastases-free period,the largest size of tumor	0.60(0.40-0.76)	0.02-0.99
		CSS	2		0.56(0.14-0.91)	NA
MSKCC nomogram	CRC(I-III+curative resection)	OS	3	Age,sex,T stage,N stage,grade,no.positive lymph node,no.total lymph nodes examined	0.67(0.47-0.82)	NA
Nordinger score	CRC(liver metastasis+curative resection)	OS	2	Age,tumor invasion into intestinal serosa,lesion of regional lymph nodes for primary tumor,metastases-free period, no.metastases,the largest size of metastasis,distance from resection edge to tumor	0.73(0.00-1.00)	NA
		CSS	2		0.59(0.15-0.92)	NA
		RFS	3		**0.57(0.53-0.60)**	0.33-0.78
PSDSS score	CRC(peritoneal metastasis)	OS	3	Clinical symptons, primary tumor pathology, PCI	**0.63(0.56-0.69)**	0.23-0.91
Valentini nomogram	RC(II-III)	LR	2	pT stage,cTstage,age,pN stage,concomitant chemotherapy,adjuvant chemotherapy	0.70(0.41-0.89)	NA
		DM	2	pT stage,pN stage,surgery type,adjuvant chemotherapy	**0.74(0.60-0.85)**	NA
		OS	2	pTstage,cTstage,age,pN stage,surgery type,adjuvant chemotherapy,radiotherapy dose,sex	**0.71(0.59-0.80)**	NA
Kanemitsu nomogram	CRC(lung metastasis+thoracotomy)	OS	2	Histology of primary tumor,no. pulmonary tumors,hilar or mediastinal lymph nodes,extrathoracic disease, CEA	0.73(0.02-1.00)	NA

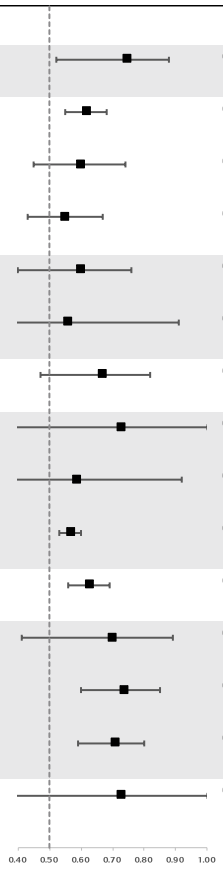

Fig. 3. Basic characteristics and summarized C-statistics of prediction models included in meta-analysis. Only external validation studies on the same prediction model were included for each meta-analysis. OS, overall survival, RFS, recurrence-free survival, CSS, colorectal cancer-specific survival, LR, local recurrence, DM, distant metastasis.PCI, peritoneal cancer index; CEA, carcinoembryonic antigen.

investigated in existing prediction models. Other strong prognostic factors for CRC such as chemo- or radiotherapy were only adopted in a small proportion of included models (13/83) due to limited data accessibility. For the CRC community, therefore, these variables should be routinely recorded in the future to develop stronger prediction models. Exploring the potential incremental predictive value of these prognostic predictors and other novel markers such as circulating tumor cells (CTC) [23] and immune-scores [24], is still of merit.

In relation to model performance, the Fong score is the most commonly studied model and it has been externally validated four times. The European Society for Medical Oncology (ESMO) consensus guidelines has discussed possible application of this score to guide adjuvant treatment for CRC with liver metastasis after hepatectomy [25], but no formal recommendations have been made. Our study identified statistically significant but modest discriminative ability for this score (C-statistic: 0.62 for RFS) as well as other models (range 0.55–0.74), which merits further improvement. Additionally, the relatively small number of external validations for each model and inherent heterogeneity across different clinical settings resulted in C-statistics with wide PIs crossing the null. The estimate discriminative performance of these models should therefore be interpreted with caution. Whilst most models adopted the C-statistic to evaluate the discriminative ability, its limitations have been widely discussed. For instance, it is hard to interpret the variation among C-statistics to compare the performance of different models derived from the same sample [26,27]. Novel metrics, such as the expected information for discrimination [28], may be adopted in future research. Our review also found that model calibration was poorly reported, which made it even more challenging to evaluate the model accuracy.

4.2. Risk of bias evaluation

The main sources for risk of bias for the current models stemmed from potential cohort attrition and methodological flaws in data analysis. The vast majority of included studies did not specify the presence and extent of loss to follow-up in the study cohort, which could bias the results and affect their validity [29]. With regard to data analysis, none of the external validation studies in our review reported how the missing data were dealt with, and only 22% of the model development studies employed missing data imputation. In addition, according to the CHARMS checklist and the proposed checklist of Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [30], future model development studies should also present more detailed prediction rules including the intercept or baseline survival to allow for individualized risk prediction rather than simply stratify CRC patients into risk groups. As for validation studies, our review identified a paucity of external validation studies that compared the validation dataset with the original model development dataset in terms of characteristics of participants and distribution of predictors. Model updating, if necessary, is also expected to be conducted and clearly presented in future validation studies. It should be noted that the CHARMS checklist is less sensitive to some sources of bias specific to survival analysis. For example, some predictors that can vary with time such as chemotherapy dosage, BMI and other

biomarkers are mostly assessed as a fixed baseline measurement, and other predictors such as second-line therapy are immeasurable at the baseline, resulting in possible time-dependent bias [31].

4.3. Model validation and impact studies

Model performance can be artificially inflated if the metrics are simply estimated based on the original sample that was used to develop the model [32]. This 'over-optimism' could be attenuated with internal validation. However, only half of the model development studies identified in our systematic review reported internal validation metrics. Fourteen (17%) of these models adopted split-sample approach despite this method being less favored due to its inefficiency [33]. Future studies should consider more sophisticated internal validation methods such cross-validation and bootstrapping [33]. External validation can, but is not limited to, quantify the potential overfitting of the original model and explore the generalizability of a model in diverse clinical settings [34]. It is ideally performed by independent investigators to avoid over-interpretation [34], but of note, only 13% of the new models in our review have been externally validated by independent investigators. Furthermore, all the external validation studies reported by independent investigators evaluated models constructed and published prior to 2011, and therefore, future work on validating newer CRC prognostic models is required.

It is also noteworthy that we failed to identify any impact studies, which are critical in defining the models' real-world impact by head-to-head comparisons [35]. Aside from that, cost-effectiveness should also be evaluated by health economic modelling, which is scarce in current CRC prognostic models [36]. Finally, few studies have explored how prediction models can be integrated into the clinical workflow [4], which will also have ramifications on their clinical utility.

4.4. Limitations

Our study has several limitations. Firstly, the majority of the included models were constructed and validated in developed countries. The performance of these models remains unclear, and therefore, needs to be validated and updated in other epidemiological settings. It is also imperative to develop and validate models in those less-studied areas especially where increasing CRC mortality rates have been observed (such as Eastern Europe and South America) [37]. Secondly, our literature search was restricted to English-language publications, inadvertently omitting models developed or validated in some other populations. Thirdly, the relatively small number of included validation studies (< 5) for each meta-analysis and between-study heterogeneity led to wide confidence intervals. Therefore, the results of each meta-analysis ought to be interpreted with caution, and need to be updated as more validation studies for these models become available. In addition, our meta-analysis was based on reported face value of model performance metrics such as C-statistics. Multiple adaptations that enable the calculation of the C-statistic from time-to-event data have been proposed [38,39]. However, most included models did not report this information, which made it challenging to harmonize the extracted statistics and could compromise the accuracy of the meta-analysis. Fourthly, this study aimed to comprehensively review the performance of existing prediction models for CRC prognosis. Potentially useful models that did not report a quantitative measure of model performance were excluded, although this has been mitigated to some extent by the inclusion and evaluation of any available external validation studies of these models. Lastly, studies without a clear prediction rule, such as models derived from genomic data using neural network, were also excluded. It is impractical for these exploratory models to be validated by independent investigators, and so they are beyond the scope of this systematic review.

5. Conclusion

Although there exist abundant prediction models on survival outcomes of CRC patients with surgical resection, only five of them (Basingstoke score, Fong score, Nordinger score, Peritoneal Surface Disease Severity Score and Valentini nomogram) have been externally validated in at least two datasets and demonstrate significant discriminative ability, which may potentially assist clinical decision-making. However, other aspects of these five models such as model calibration, their impact in real-word and cost-effectiveness should be further investigated before formal recommendation can be made for use in clinical practice. As for other models that have not been validated in independent datasets and are subject to risk of bias, current evidence is insufficient to evaluate their performance externally, which does not support for these models to be routinely applied. Future research should focus not only on constructing new models with novel predictors, but also on validating and investigating the impact of existing prediction models to improve prediction for CRC prognosis.

Authors' contributions

Literature search: YH, YO and XL.
Study selection: YH and YO.
Data extraction: YH, YO and ZW.
Data analysis: YH and XL.
Manuscript draft and revision: YH, YH, XL, FVD, SMF, ZW, MT, EB, HC, MGD, ET.
Approval to submission: ET.

Competing interests statement

The authors declare no competing interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. YH and XL are supported by the China Scholarship Council. ET is supported by a CRUK Career Development Fellowship (C31250/A22804). MGD is supported by a CRUK programme grant C348/A18927 and a MRC Human Genetics Unit Centre Grant (U127527202 and U127527198).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.suronc.2019.05.014.

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, Int. J. Cancer 136 (5) (2015) E359–E386.
[2] Colorectal Cancer Facts & Figures 2017-2019 (Accessed on 17th July,2018 : https://www.cancer.org/research/cancer-facts-statistics/colorectal-cancer-facts-figures.html).
[3] B.A. Spindler, J.R. Bergquist, C.A. Thiels, E.B. Habermann, S.R. Kelley, D.W. Larson, K.L. Mathis, Incorporation of CEA improves risk stratification in stage II colon cancer, J. Gastrointest. Surg. 21 (5) (2017) 770–777.
[4] A.J. Vickers, Prediction models in cancer care, CA A Cancer J. Clin. 61 (5) (2011) 315–326.
[5] T.L. Bowles, C.Y. Hu, N.Y. You, J.M. Skibber, M.A. Rodriguez-Bigas, G.J. Chang, An individualized conditional survival calculator for patients with rectal cancer, Dis. Colon Rectum 56 (5) (2013) 551–559.
[6] I.J. Goossens-Beumer, R.S. Derr, H.P. Buermans, J.J. Goeman, S. Bohringer, H. Morreau, U. Nitsche, K.P. Janssen, C.J. van de Velde, P.J. Kuppen, MicroRNA classifier and nomogram for metastasis prediction in colon cancer, Cancer Epidemiol. Biomark. Prev. 24 (1) (2015) 187–197.
[7] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, P. Group, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Ann. Intern. Med. 151 (4) (2009) 264–269 W264.
[8] B.J. Ingui, M.A. Rogers, Searching for clinical prediction rules in MEDLINE, J. Am.

Med. Inform. Assoc. 8 (4) (2001) 391–397.
[9] J. Brush, K. Boyd, F. Chappell, F. Crawford, M. Dozier, E. Fenwick, J. Glanville, H. McIntosh, A. Renehan, D. Weller, M. Dunlop, The value of FDG positron emission tomography/computerised tomography (PET/CT) in pre-operative staging of colorectal cancer: a systematic review and economic evaluation, Health Technol. Assess. 15 (35) (2011) 1-192, iii-iv.
[10] T.P. Debray, J.A. Damen, K.I. Snell, J. Ensor, L. Hooft, J.B. Reitsma, R.D. Riley, K.G. Moons, A guide to systematic review and meta-analysis of prediction model performance, BMJ 356 (2017) i6460.
[11] K.G. Moons, J.A. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D.G. Altman, J.B. Reitsma, G.S. Collins, Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist, PLoS Med. 11 (10) (2014) e1001744.
[12] H.A. Smit, M. Pinart, J.M. Anto, T. Keil, J. Bousquet, K.H. Carlsen, K.G. Moons, L. Hooft, K.C. Carlsen, Childhood asthma prediction models: a systematic review, Lancet Respir. Med. 3 (12) (2015) 973–984.
[13] M. Lamain–de Ruiter, A. Kwee, C.A. Naaktgeboren, A. Franx, K.G. Moons, M.P. Koster, Prediction models for the risk of gestational diabetes: a systematic review, Diagn. Progn. Res. 1 (1) (2017) 3.
[14] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed., John Wiley & Sons, New York, NY, 2000.
[15] J. IntHout, J.P. Ioannidis, G.F. Borm, The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method, BMC Med. Res. Methodol. 14 (2014) 25.
[16] J. IntHout, J.P. Ioannidis, M.M. Rovers, J.J. Goeman, Plea for routinely presenting prediction intervals in meta-analysis, BMJ Open 6 (7) (2016) e010247.
[17] J.P. Higgins, S.G. Thompson, D.J. Spiegelhalter, A re-evaluation of random-effects meta-analysis, J. R. Stat. Soc. Ser. A Stat. Soc. 172 (1) (2009) 137–159.
[18] Y. Takakura, M. Okajima, Y. Kanemitsu, S. Kuroda, H. Egi, T. Hinoi, H. Tashiro, H. Ohdan, External validation of two nomograms for predicting patient survival after hepatic resection for metastatic colorectal cancer, World J. Surg. 35 (10) (2011) 2275–2282.
[19] F. Peng, D. Hu, X. Lin, G. Chen, B. Liang, Y. Chen, C. Li, H. Zhang, Y. Xia, J. Lin, X. Zheng, W. Niu, An in-depth prognostic analysis of baseline blood lipids in predicting postoperative colorectal cancer mortality: the FIESTA study, Cancer Epidemiol. 52 (2018) 148–157.
[20] I. Arostegui, N. Gonzalez, N. Fernandez-de-Larrea, S. Lazaro-Aramburu, M. Bare, M. Redondo, C. Sarasqueta, S. Garcia-Gutierrez, J.M. Quintana, R.C.-C. Group, Combining statistical techniques to predict postsurgical risk of 1-year mortality for patients with colon cancer, Clin. Epidemiol. 10 (2018) 235–251.
[21] M. Rees, P.P. Tekkis, F.K. Welsh, T. O'Rourke, T.G. John, Evaluation of long-term survival after hepatic resection for metastatic colorectal cancer: a multifactorial model of 929 patients, Ann. Surg. 247 (1) (2008) 125–135.
[22] https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf Accessed on September 23rd 2018.
[23] N.N. Rahbari, M. Aigner, K. Thorlund, N. Mollberg, E. Motschall, K. Jensen, M.K. Diener, M.W. Buchler, M. Koch, J. Weitz, Meta-analysis shows that detection of circulating tumor cells indicates poor prognosis in patients with colorectal cancer, Gastroenterology 138 (5) (2010) 1714–1726.
[24] B. Mlecnik, M. Van den Eynde, G. Bindea, S.E. Church, A. Vasaturo, T. Fredriksen, L. Lafontaine, N. Haicheur, F. Marliot, D. Debetancourt, G. Pairet, A. Jouret-Mourin, J.F. Gigot, C. Hubert, E. Danse, C. Dragean, J. Carrasco, Y. Humblet, V. Valge-Archer, A. Berger, F. Pages, J.P. Machiels, J. Galon, Comprehensive intrametastatic immune quantification and major impact of immunoscore on survival, J. Natl. Cancer Inst. 110 (1) (2018) 01.
[25] E. Van Cutsem, A. Cervantes, R. Adam, A. Sobrero, J.H. Van Krieken, D. Aderka, E. Aranda Aguilar, A. Bardelli, A. Benson, G. Bodoky, F. Ciardiello, A. D'Hoore, E. Diaz-Rubio, J.Y. Douillard, M. Ducreux, A. Falcone, A. Grothey, T. Gruenberger, K. Haustermans, V. Heinemann, P. Hoff, C.H. Kohne, R. Labianca, P. Laurent-Puig, B. Ma, T. Maughan, K. Muro, N. Normanno, P. Osterlund, W.J. Oyen, D. Papamichael, G. Pentheroudakis, P. Pfeiffer, T.J. Price, C. Punt, J. Ricke, A. Roth, R. Salazar, W. Scheithauer, H.J. Schmoll, J. Tabernero, J. Taieb, S. Tejpar, H. Wasan, T. Yoshino, A. Zaanan, D. Arnold, ESMO consensus guidelines for the management of patients with metastatic colorectal cancer, Ann. Oncol. 27 (8) (2016) 1386–1422.
[26] M. Diouf, B. Chibaudel, T. Filleron, C. Tournigand, M. Hug de Larauze, M.L. Garcia-Larnicol, S. Dumont, C. Louvet, N. Perez-Staub, A. Hadengue, A. de Gramont, F. Bonnetain, Could baseline health-related quality of life (QoL) predict overall survival in metastatic colorectal cancer? The results of the GERCOR OPTIMOX 1 study, Health Qual. Life Outcomes 12 (2014) 69.
[27] K. Kawai, S. Ishihara, H. Yamaguchi, E. Sunami, J. Kitayama, H. Miyata, K. Sugihara, T. Watanabe, Nomograms for predicting the prognosis of stage IV colorectal cancer after curative resection: a multicenter retrospective study, Eur. J. Surg. Oncol. 41 (4) (2015) 457–465.
[28] P. McKeigue, Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic, Stat. Methods Med. Res. (2018) 962280218776989.
[29] J.R. Dettori, Loss to follow-up, Evid. Based Spine Care J. 2 (1) (2011) 7–10.
[30] K.G. Moons, D.G. Altman, J.B. Reitsma, J.P. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration, Ann. Intern. Med. 162 (1) (2015) W1–W73.
[31] C. van Walraven, D. Davis, A.J. Forster, G.A. Wells, Time-dependent bias was common in survival analyses published in leading clinical journals, J. Clin. Epidemiol. 57 (7) (2004) 672–682.
[32] F.E. Harrell Jr., K.L. Lee, D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Stat. Med. 15 (4) (1996) 361–387.
[33] E.W. Steyerberg, F.E. Harrell Jr., G.J. Borsboom, M.J. Eijkemans, Y. Vergouwe, J.D. Habbema, Internal validation of predictive models: efficiency of some procedures for logistic regression analysis, J. Clin. Epidemiol. 54 (8) (2001) 774–781.
[34] G.S. Collins, J.A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.M. Yu, K.G. Moons, D.G. Altman, External validation of multivariable prediction models: a systematic review of methodological conduct and reporting, BMC Med. Res. Methodol. 14 (2014) 40.
[35] K.G. Moons, D.G. Altman, Y. Vergouwe, P. Royston, Prognosis and prognostic research: application and impact of prognostic models in clinical practice, BMJ 338 (2009) b606.
[36] A. van Giessen, J. Peters, B. Wilcher, C. Hyde, C. Moons, A. de Wit, E. Koffijberg, Systematic review of health economic impact evaluations of risk prediction models: stop developing, start evaluating, Value Health 20 (4) (2017) 718–726.
[37] M. Arnold, M.S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global patterns and trends in colorectal cancer incidence and mortality, Gut 66 (4) (2017) 683–691.
[38] P.C. Austin, M.J. Pencinca, E.W. Steyerberg, Predictive accuracy of novel risk factors and markers: a simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model, Stat. Methods Med. Res. 26 (3) (2017) 1053–1077.
[39] P. Blanche, J.F. Dartigues, H. Jacqmin-Gadda, Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring, Biom. J. Biom. Z. 55 (5) (2013) 687–704.