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

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

\textsuperscript{a} Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
\textsuperscript{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
\textsuperscript{c} Western General Hospital, Edinburgh, UK
\textsuperscript{d} Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
\textsuperscript{e} Edinburgh Cancer Centre NHS Lothian, Edinburgh, UK
\textsuperscript{f} Department of Gastrointestinal Surgery, West China Hospital, Sichuan University, Chengdu, 610041, P.R. 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 Valentinianomogram) 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

0960-7404/ © 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).
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. R²) 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...
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 = 25, 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

Table 1

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

*p-values for Chi-square test.
CRC, colorectal cancer; AUC, area under receiver's operating characteristic curve.
* Including D-statistic, sensitivity and specificity.
* 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

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 = 25, 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
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, 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 Table 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 4). 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...
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 ac-
cessibility. For the CRC community, therefore, these variables should
routinely be 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 guide-
lines 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 statis-
tically 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 in-
terpret 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 ana-
lysis. 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 Diag-
nosis (TRIPOD) [30], future model development studies should also
present more detailed prediction rules including the intercept base-
line 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 con-
ducted 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Outcome</th>
<th>No. Studies</th>
<th>Predictors</th>
<th>Pooled C-statistic (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basingstoke score</td>
<td>RFS</td>
<td>2</td>
<td>Lymph node status, tumor differentiation, CEA, no liver metastasis, diameter of primary tumor, extracapsular proliferation</td>
<td>0.74 (0.52-0.88) NA</td>
</tr>
<tr>
<td>Fong score</td>
<td>RFS</td>
<td>4</td>
<td>Positive resection margin, extrapleural lesion, lesion of regional lymph nodes for primary tumor, metastasis-free period, no metastases, the largest size of metastasis, CEA</td>
<td>0.62 (0.55-0.68) 0.47-0.74</td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>4</td>
<td></td>
<td>0.60 (0.45-0.74) 0.32-0.82</td>
</tr>
<tr>
<td></td>
<td>CSS</td>
<td>2</td>
<td></td>
<td>0.55 (0.43-0.67) NA</td>
</tr>
<tr>
<td>Watsukii score</td>
<td>RFS</td>
<td>3</td>
<td>No. metastases, bidirectional lesion, metastasis-free period, the largest size of tumor</td>
<td>0.60 (0.40-0.76) 0.02-0.99</td>
</tr>
<tr>
<td></td>
<td>CSS</td>
<td>2</td>
<td></td>
<td>0.56 (0.14-0.91) NA</td>
</tr>
<tr>
<td>MSKCC nomogram</td>
<td>OS</td>
<td>3</td>
<td>Age, sex, T stage, N stage, grade, no positive lymph node, no total lymph nodes examined</td>
<td>0.67 (0.47-0.82) NA</td>
</tr>
<tr>
<td>Nordiner score</td>
<td>OS</td>
<td>2</td>
<td>Age, tumor invasion into intestinal serosa, lesion of regional lymph nodes for primary tumor, metastasis-free period, no metastases, the largest size of metastasis, distance from resection edge to tumor</td>
<td>0.73 (0.00-1.00) NA</td>
</tr>
<tr>
<td></td>
<td>CSS</td>
<td>2</td>
<td></td>
<td>0.59 (0.15-0.92) NA</td>
</tr>
<tr>
<td></td>
<td>RFS</td>
<td>3</td>
<td></td>
<td>0.57 (0.53-0.60) 0.33-0.78</td>
</tr>
<tr>
<td>PSDSS score</td>
<td>OS</td>
<td>3</td>
<td>Clinical symptoms, primary tumor pathology, PCI</td>
<td>0.63 (0.56-0.69) 0.23-0.91</td>
</tr>
<tr>
<td>Valenti nomogram</td>
<td>LR</td>
<td>2</td>
<td>pT stage, C, Tstage, age, p, no stage, concomitant chemotherapy, adjuvant chemotherapy</td>
<td>0.70 (0.41-0.89) NA</td>
</tr>
<tr>
<td></td>
<td>DM</td>
<td>2</td>
<td>pT stage, p, stage, surgery type, adjuvant chemotherapy</td>
<td>0.74 (0.60-0.85) NA</td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>2</td>
<td>pTstage, C, Tstage, age, p, stage, surgery type, adjuvant chemotherapy, radiotherapy dose, sex</td>
<td>0.71 (0.59-0.80) NA</td>
</tr>
<tr>
<td>Kanemitsu nomogram</td>
<td>OS</td>
<td>2</td>
<td>Histology of primary tumor, no, pulmonary tumor, hilar or mediastinal lymph nodes, extracapsular disease, CEA</td>
<td>0.73 (0.02-1.00) NA</td>
</tr>
</tbody>
</table>

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.
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 as 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 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 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-world 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.
Competition 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


