Development and validation of risk prediction models for colorectal cancer in patients with symptoms

Wei Xu[1], Ines Mesa Eguiagaray[1], Theresa Kirkpatrick[1], Jennifer Devlin [23], Stephanie Brogan[4], Patricia Turner[4], Chloe Macdonald[5], Michelle Thornton[6], Xiaomeng Zhang[1], Yazhou He[1], Xue Li[1], Maria Timofeeva[23], Susan Farrington[23], Farhat Din[23], Malcolm Dunlop[23], Evropi Theodoratou[13*]

[1]Centre for Global Health, Usher Institute, University of Edinburgh; [2]Colon Cancer Genetics Group, Medical Research Council Human Genetics Unit, Medical Research Council Institute of Genetics & Molecular Medicine, University of Edinburgh; [3]Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh; [4]Clinical Research Team, Oncology Dept, Forth Valley Royal Hospital, Stirling Road, Larbert; [5]University Hospital Wishaw & University Hospital Monklands, NHS Lanarkshire; [6]Wishaw General Hospital, Wishaw.

***Corresponding Author**: *Evropi Theodoratou*, Centre for Global Health, Usher Institute, University of Edinburgh; e-mail, E.Theodoratou@ed.ac.uk

Introduction

Colorectal cancer (CRC) was the third most common cancer and the second leading cause of cancer-related death in the world (GLOBOCAN 2020 estimate).

CRC prediction models could provide the disease risk assessment to identify patients with higher cancer risk, and to assist clinical professionals in their decision-making about further clinical care such as risk-tailored cancer screening, testing, and treatments (Shipe et al., 2019).

In this study, we aimed to develop and validate risk prediction models that incorporated demographics, clinical features, and a weighted genetic risk score (wGRS) for individual prediction of CRC risk in patients with symptoms.

Methods

CRC prediction models were developed with internal validation in the Study of Colorectal Cancer in Scotland (SOCCS) and the Lothian Bowel Symptoms Study (LABSS) [CRC Cases: n = 1686 / Controls: n = 963] (Figure 1).

The two main strategies for the development of the final model are predictor selection and full model (Royston et al., 2009). Models A (baseline model + wGRS) and B (baseline model) were developed based on LASSO regression to select predictors. Models C (baseline model + wGRS) and D (baseline model) were built using all the variables.

Models' prediction performance (*calibration, discrimination*) were evaluated through Hosmer-Lemeshow (HL) test (calibration curves were plotted) and Harrell's C-statistics (ROC and PRC curves were plotted). The corrected C-statistics were calculated based on bootstrapping validation (1,000 bootstraps resamples). The continuous Net Reclassification Index (NRI) and Integrated Discrimination Index (IDI) were calculated after recalibration to compare models and assess the prediction increment. An online nomogram for the final model was built using Shiny.apps.

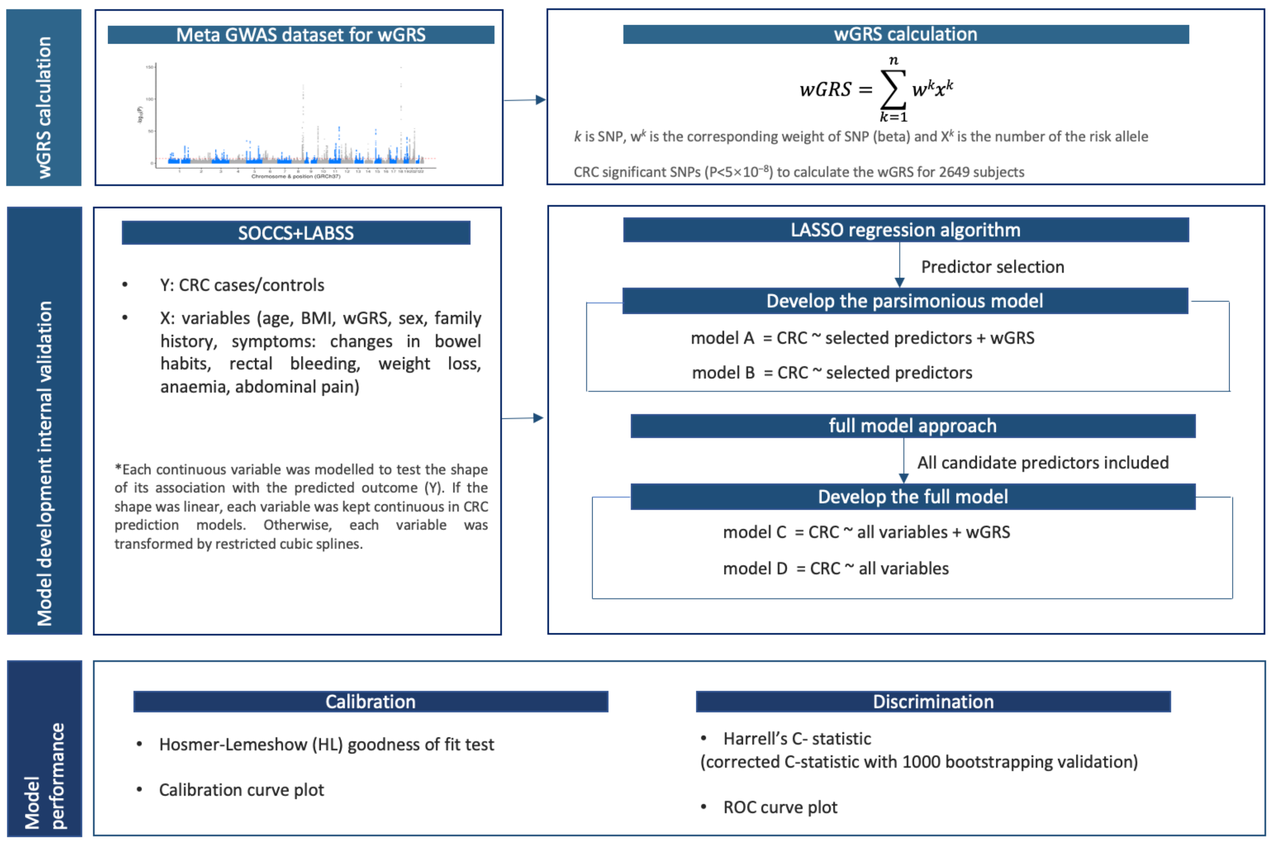

Figure 1: CRC risk prediction models construction and internal validation

Results

Models A (parsimonious LASSO model) and C (full model) that integrated $wGRS_{202}$ in combination with demographic and clinical predictors had better prediction performance, compared to baseline models B and D (Table 1; Figures 2-7). The findings suggested incremental predictive value had been introduced by the addition of $wGRS_{202}$ [Model A vs. B: NRI = 0.226 (0.149–0.335), IDI = 0.019 (0.013–0.024); Model C vs. D: NRI = 0.239 (0.154–0.340), IDI = 0.018 (0.013–0.023); P < 0.001].

Comparing the parsimonious model A and the full model C, there was no statistical difference in model predictive accuracy [C-statistic increment=0.001, P=0.479]. In this study, the increased time and cost to collect the larger number of predictors for the full model C outweighed the increased predictive accuracy. A good compromise between model parsimony and accuracy is important (Diaz-Ramirez et al., 2021). From a practical perspective, the parsimonious model A is easier to interpret, generalize, and use in practice.

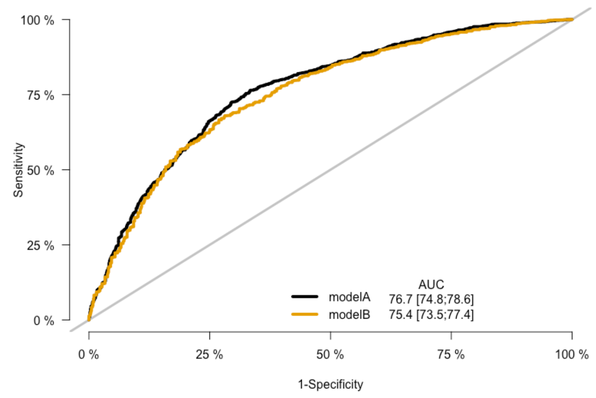

Figure 2: ROC curves-model A and model B comparison

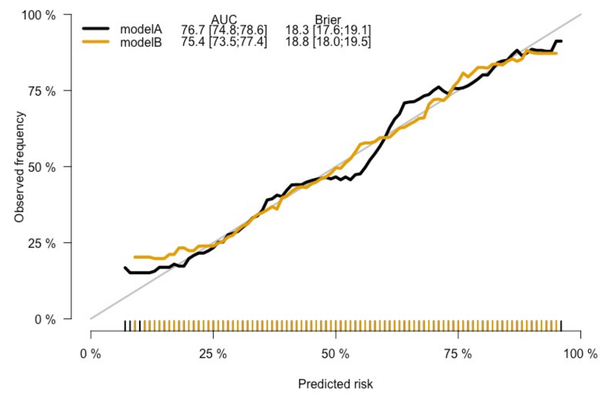

Figure 3: Calibration curves-model A and model B comparison

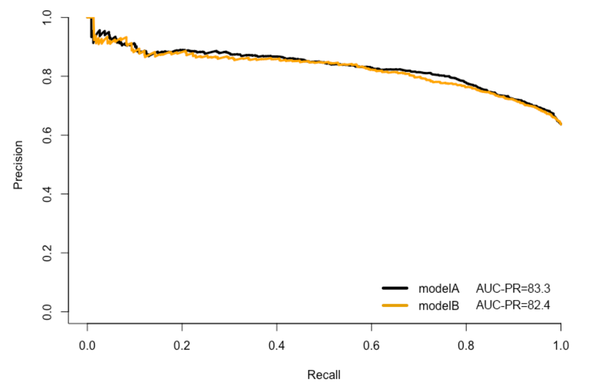

Figure 4: Precision recall curves-model A and model B comparison

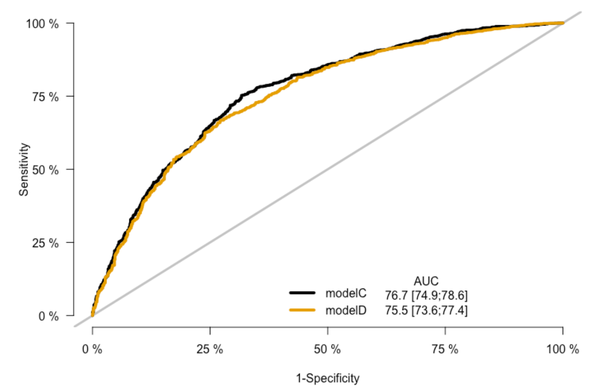

Figure 5: ROC curves-model C and model D comparison

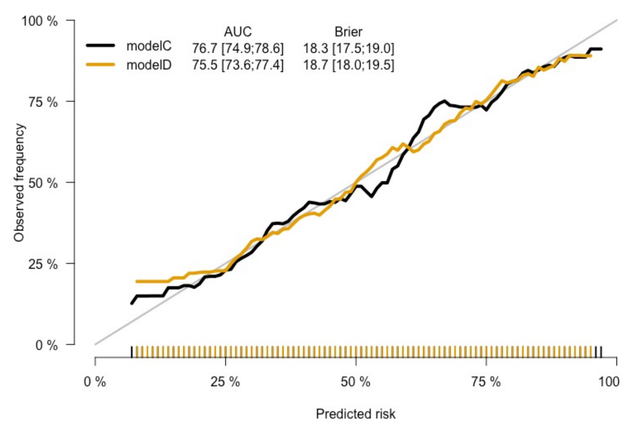

Figure 6: Calibration curves-model C and model D comparison

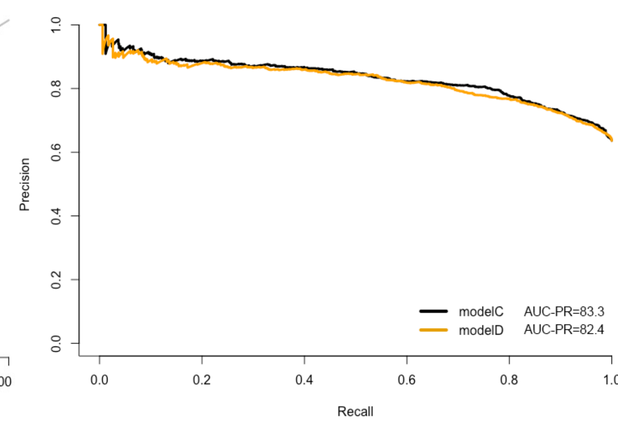

Figure 7: Precision recall curves-model C and model D comparison

An online CRC risk prediction nomogram/calculator A was built, which can be accessed through the link (https://crcpredictionmodel.shinyapps.io/dynnomapp/). The CRC risk for individuals can be calculated via inputting each patient's information.

Results

Each model's predictors, intercept, coefficients, discrimination, and calibration estimates are presented in Table 1.

Table 1. A summary of CRC risk prediction models comparison

Model	Method	Case	Control	λ	Predictors	Intercept	Coefficient	P-value	R^2	Brier	C-statistic	Corrected C-statistic	AUC-PR	HL P-value
Model A	LASSO	1686	963	0.0257	$wGRS_{202}$ Age Sex Change of bowel habit Abdominal pain	-1.3030	0.7612 0.0410 0.3611 -1.2411 -0.6784	5.31×10^{-13} 3.53×10^{-29} 7.19×10^{-5} 8.06×10^{-29} 7.65×10^{-12}	0.266	0.183	0.767	0.765 (1000 bootstrap)	0.833	0.024
Model B	LASSO	1686	963	0.0310	Age Sex Change of bowel habit Abdominal pain	-1.2124	0.0401 0.3690 -1.2411 -0.7020	1.06×10^{-28} 4.09×10^{-5} 1.34×10^{-39} 7.77×10^{-13}	0.244	0.188	0.754	0.753 (1000 bootstrap)	0.824	0.711
Model C	full model	1686	963	NA	$wGRS_{202}$ Age Sex BMI Family history Change of bowel habit Rectal bleeding Weight loss Anaemia Abdominal pain	-0.7679	0.7603 0.0410 0.3631 -0.0195 -0.0024 -1.2616 0.0402 -0.0112 -0.0531 -0.6786	6.91×10^{-13} 2.65×10^{-28} 7.05×10^{-65} 0.019 0.985 7.68×10^{-37} 0.686 0.928 0.679 1.55×10^{-11}	0.269	0.183	0.767	0.764 (1000 bootstrap)	0.833	0.018
Model D	full model	1686	963	NA	Age Sex BMI Family history Change of bowel habit Rectal bleeding Weight loss Anaemia Abdominal pain	-0.7170	0.0404 0.3714 -0.0191 -0.0349 -1.2667 0.0734 -0.0661 -0.6999 -0.6786	4.12×10^{-28} 3.94×10^{-5} 0.020 0.774 7.07×10^{-38} 0.455 0.966 0.602 2.03×10^{-12}	0.247	0.187	0.755	0.752 (1000 bootstrap)	0.824	0.428

AUC-PR: area under the precision recall curve; HL: Hosmer-Lemeshow

Discussion

CRC prediction models were developed with internal validation to predict individualized CRC risk in patients with symptoms.

Strengths:

- CRC prediction models were developed with internal validation to alleviate the models' overfitting and optimism.
- To our best knowledge, this is the first study that developed and internally validated prediction models using wGRS combined with demographic and clinical factors for CRC risk in patients with symptoms. The previously published 19 prediction models did not use genetic predictors (neither SNPs nor GRS).
- Our findings supported that Integration of genetic architecture into CRC classical prediction model could only marginally improve prediction performance.

Limitations:

- This risk prediction modelling study (N=2649) has a small sample size and may not be sufficiently representative of the population.
- Internal validation cannot address selection bias with recruitment, or measurement errors as validation is performed within the study population.
- The majority of CRC cases came from SOCCS, and all controls were from LABSS. The different variable collection methods in SOCCS (GP e-referrals) and LABSS (questionnaire) could bias the study's results.
- The developed CRC risk prediction models have not been externally validated.

References

SCAN ME