Germline Genetic Variations and Survival Outcomes of Colorectal Cancer

Yazhou He

何亚舟

Thesis submitted for the degree of

Doctor of Philosophy

Usher Institute

The University of Edinburgh

2020

Declaration

I hereby declare that the research within this thesis solely describes my own work. I conducted all aspects of this research except where states otherwise by reference or acknowledgment. The work described has not been submitted for any other degree or professional qualification.

Signature：   Date: March 1st 2020

—To my dearest family

Related publications

The following publications are derived from the research contents presented in the thesis. All published papers were reproduced with permission from the publishers.

He Y, Ong Y, Li X, Din FV, Brown E, Timofeeva M, Wang Z, Farrington SM, Campbell H, Dunlop MG, Theodoratou E. Performance of prediction models on survival outcomes of colorectal cancer with surgical resection: A systematic review and meta-analysis. Surgical oncology. 2019 May 20. Epub ahead of print.

He Y, Theodoratou E, Li X, Din FV, Vaughan-Shaw P, Svinti V, Farrington SM, Campbell H, Dunlop MG, Timofeeva M. Effects of common genetic variants associated with colorectal cancer risk on survival outcomes after diagnosis: A large population-based cohort study. International journal of cancer. 2019 Nov 1;145(9):2427-32.

He Y, Timofeeva M, Li X, Din FV, Blackmur JP, Vaughan-Shaw P, Svinti V, Farrington SM, Campbell H, Dunlop MG, Theodoratou E. A comprehensive study of the effect on colorectal cancer survival of common germline genetic variation previously linked with cancer prognosis. Cancer Epidemiology and Prevention Biomarkers. 2019 Nov 1;28(11):1944-6.

Other publications during the PhD programme:

He Y, Li X, Gasevic D, Brunt E, McLachlan F, Millenson M, Timofeeva M, Ioannidis JP, Campbell H, Theodoratou E. Statins and multiple noncardiovascular outcomes: umbrella review of meta-analyses of observational studies and randomized controlled trials. Annals of internal medicine. 2018 Oct 16;169(8):543-53.

He Y, Timofeeva M, Farrington SM, Vaughan-Shaw P, Svinti V, Walker M, Zgaga L, Meng X, Li X, Spiliopoulou A, Jiang X et al. Exploring causality in the association between circulating 25-hydroxyvitamin D and colorectal cancer risk: a large Mendelian randomisation study. BMC medicine. 2018 Dec;16(1):1-11.

Li X, Meng X, **He Y**, Spiliopoulou A, Timofeeva M, Wei WQ, Gifford A, Yang T, Varley T, Tzoulaki I, Joshi P et al. Genetically determined serum urate levels and cardiovascular and other diseases in UK Biobank cohort: A phenome-wide mendelian randomization study. PLoS medicine. 2019 Oct;16(10).

Meng X, Li X, Timofeeva MN, **He Y**, Spiliopoulou A, Wei WQ, Gifford A, Wu H, Varley T, Joshi P, Denny JC et al. Phenome-wide Mendelian-randomization study of genetically determined vitamin D on multiple health outcomes using the UK Biobank study. International journal of epidemiology. 2019 Oct 1;48(5):1425-34.

Banwell VC, Phillips HA, Duff MJ, Speake D, McLean C, Williams LJ, **He Y**, Paterson HM. Five-year oncological outcomes after selective neoadjuvant radiotherapy for resectable rectal cancer. Acta Oncologica. 2019 Sep 2;58(9):1267-72.

Acknowledgements

First and foremost, my heartfelt gratitude goes to my principal supervisor **Professor Evropi Theodoratou** who set me up for this wonderful three-year journey at the crossroad of my career. I felt this was destined to happen when we first met and I introduced my name which refers to ‘Asia’ in Chinese, to which you responded with surprise that your name means ‘Europe’ in Greek. Thank you for opening up the window of the beautiful world of epidemiology for me. Special thanks for this generous offer for this PhD programme which I never thought of but turns out to be one of the most fascinating chapters of my life. I am grateful for all your guidance, support and encouragement along the entire journey without which I would by no means become who I am today.

I would like to express my sincerest thanks to my second supervisor **Professor Harry Campbell** who kindly invited me here for a short visit at the first place, which led to this amazing adventure in epidemiology. Your encouragement always cheered me up every time I felt disheartened. Thank you for your dedicated support throughout the course of this PhD.

My special thanks also go to **Professor Malcolm Dunlop** for being such an amazing advisor during my stay here. Thank you for welcoming and hosting me in the amazing Colon Cancer Genetics Group when I first arrived, and for sharing with me your critical thinking as well as that unique sense of humour as a surgeon. I also owe my deep gratitude to my second advisor **Dr Maria Timofeeva**. Your never-ending patience on every detail of my research projects has helped me through all the adversities. Your unrivalled work ethic has always been inspiring me to pursue next goals. I have greatly benefited from your meticulous guidance on analytical issues of my research.

I would also like to thank **Farhat Din, Susan Farrington, Victoria Svinti, Marion Walker, Stuart Reid, Peter Vaughan-Shaw, James Blackmur** and all other colleagues from the Colon Cancer Genetics Group for all your continued support for my research projects. I also received generous support from **Peter Joshi**, **Athina Spiliopoulou** and **Xia Shen** from the Usher Institute who provided constructive comments and suggestions based on their statistical expertise.

I would like to thank my friends here in Edinburgh. **Xue**, I cannot be more thankful for all your support and guidance on my research, and also for being the elder sister in our group and unselfishly sharing your ideas and experiences with us. My special thanks go to **Xiangrui, Xiaomeng, Wei** and **Tian** for being such great teammates to work with and also for being amazing friends to share happiness and sorrow. I also want to thank **Hongjiang, Peige, You, Xin, Ting and Zhicheng** who welcomed me into the Usher family and brought countless joy to our office. Particular thanks to my sisters and brothers at the Holyrood Residence: **Ziwei, Nianrong, Chuchu, Weijing, Ni, Ziwen, Li, Ziying,** and **Zhe.** If I had to do it over again, I would not hesitate for a single second to choose this residence just to meet you all. Thank you for continuously sharing your delicious dishes with me since what I cook is universally deemed as uneatable. More importantly, thank you all for bearing with my lame jokes.

I am grateful for **Professor Ziqiang Wang** and **Professor Yanhong Zhou—**my mentors during my medical training back in China. Your advices and encouragement have always been inspiring me to be a better person. Also, my thanks go to **Xiangbing, Mingtian, Tinghan, Qingbing, Xiao** and other colleagues from the West China Hospital, and **Xin, Chi, Jiarong, Shenyang, Haichuan, Yuyang, Junhua, Hong, Jingquan, Ming, Chengdi, Xiaosa, Weiwei, Tianli, Hao** and other friends from the West China School of Medicine. Although we have been thousands of miles away from each other, I always feel like I have never left.

An extra special thank you to my wife for your unconditional love and support, for being the most beautiful **SONG** in my life, and for being the beacon shining into any darkness of my heart.

亲爱的爸爸妈妈，感谢你们对我三十年如一日的爱，支持我每一个正确的或错误的决定。寸草之心，难报三春之辉。唯愿你们平安健康，福泽绵长。

Abstract

Background

Colorectal cancer (CRC) was the second commonest cancer and the third leading cause of cancer-related deaths worldwide in 2018. In the UK, the overall 5-year survival rate of CRC patients is approximately 60%. Colorectal cancer patients are staged based on the staging system recommended by the American Joint Committee on Cancer (AJCC). The 5-year survival rates vary from approximately 90% for stage I to 10% for stage IV CRC patients. Although the AJCC stage is the main indicator of patients' prognosis, there is still substantial variation in terms of the survival outcomes of CRC patients within each stage. This merits further examination of other prognostic factors to improve prediction of CRC survival. Previous evidence revealed that germline genetic background plays an important role in determining survival outcomes of CRC patients. However, the human germline genome consists of millions of genetic variants and no specific genetic loci have been robustly mapped in relation to prognosis of CRC patients to date. Firstly, this thesis seeks to systematically review existing literature and explore whether germline genetic variants have been adopted in published multivariable models in attempts to predict CRC survival. Secondly, multiple CRC patient cohorts were leveraged to investigate associations between germline genetic variants and survival outcomes of CRC patients after diagnosis.

Methods

A systematic literature search was conducted in MEDLINE and Embase databases to retrieve published multivariable prediction models that were developed to forecast survival outcomes of CRC. Risk of bias for included models was assessed using published evaluation tools and metrics evaluating model performance were extracted and quantitatively assessed using meta-analysis.

Multiple study cohorts were used in this thesis including the Study of Colorectal Cancer in Scotland (SOCCS), incident CRC cases from the UK Biobank cohort and datasets from three previously published clinical trials (QUASAR2, SCOT and VICTOR). Firstly, germline genetic variants associated with CRC survival that were reported by published genome-wide association studies (GWAS) were identified by searching the NHGRI-EBI GWAS catalogue. Associations between these variants and overall and CRC-specific survival were investigated as a replication study using

the SOCCS cohort. Then I explored the potential predictive value of these previously reported variants in the UK Biobank study by developing a genetic predictor combining these variants, and evaluated the predictive performance of the predictor along with other variables (age at diagnosis, sex, AJCC stage and tumour grade) using the SOCCS as an external validation cohort. The model performance was assessed in terms of the discriminative ability and model calibration. The next step was to conduct two candidate genetic association studies to test the potential effects of two groups of genetic variants—variants associated with CRC risk and variants associated with prognosis of other cancers—on survival outcomes of CRC patients from the SOCCS study. These two groups of variants were identified from two large GWAS meta-analyses and the GWAS catalogue. Stratified analyses were performed by sex, AJCC stage (stage II/III and IV) and tumour site (colon and rectum). Cox regression models were used to estimate effects—hazard ratios (HRs)--of genetic variants on survival outcomes with age at diagnosis, sex and AJCC stage as covariates. The false discovery rate (FDR) approach was used to correct for multiple testing. Genetic effects were tested under both the additive and recessive genetic models.

Finally, I performed a GWAS on both overall and CRC-specific survival by investigating a total of overall eight million autosomal genetic variants throughout the genome using the SOCCS study. The effect estimates for each variant were obtained using a Martingale-residual based approach. Discoveries of the GWAS were then replicated by performing meta-analysis combining effect estimates from the UK Biobank cohort and the three clinical trials. Stratified GWASs were also conducted in SOCCS for stage II/III and stage IV CRC patients separately. Enrichment analyses were employed to detect potential genomic signals enriched in possible genes and gene-sets that are involved in relevant biological pathways.

Results

The systematic literature review identified 83 original prediction models and 52 separate external validation studies. Five models (Basingstoke score, Fong score, Nordinger score, Peritoneal Surface Disease Severity Score and Valentini nomogram) were validated in at least two external datasets and showed positive discriminative ability in terms of model performance. No germline genetic variants had been used as prognostic predictors in published prediction models.

A total of 5,675 CRC patients from the SOCCS cohort, 2,474 incident CRC cases from the UK Biobank cohort and 4,771 CRC patients from the three clinical trials were included in the main analysis. By searching the GWAS catalogue, I identified 43 independent genetic variants ($r^2<0.2$) that were previously linked with CRC survival outcomes. After correcting for FDR, none of these 43 variants, under the additive genetic model, were significantly associated with either overall or CRC-specific survival of CRC patients from the SOCCS cohort. Only three variants (rs17026425, rs17057166 and rs6854845) at nominal significance (unadjusted $p<0.05$) showed concordant direction of effects with previously published GWASs, whereas one variant with uncorrected $p<0.05$ showed opposite direction of effect (rs11138220). The polygenic risk score (PRS) combining the 43 variants was not associated with CRC survival outcomes. No significant associations after adjusting for FDR were found in the stratified analysis. Although four variants (rs17280262, rs16867335, rs6854845 and rs17057166) showed potential effects when the recessive model of inheritance was used in SOCCS, I failed to replicate these effects using data from the UK Biobank cohort.

With respect to the predictive performance of the 43 variants in the UK Biobank cohort, the genetic predictor combining the 43 variants did not show statistically significant C statistics after internal validation, with the 95% confidence intervals (CIs) including the null (overall survival: C=0.510, 95%CI=0.498-0.521; CRC-specific survival: C=0.518, 95%CI=0.498-0.530). Similarly, non-significant C statistics were observed for the 43-variant predictor in the external validation analysis using the SOCCS cohort. Moreover, the prediction model composed of the 43 variants was poorly calibrated in both the UK Biobank and the SOCCS cohorts. The model performance remained nearly unchanged when combining the genetic predictor with other variables including age at diagnosis, sex, AJCC stage and tumour grade in the SOCCS cohort, suggesting no incremental predictive value had been introduced by the addition of genetic variants.

Regarding the other two groups of candidate genetic variants, a total of 128 independent variants ($r^2<0.2$) associated with CRC risk and 82 independent variants ($r^2<0.2$) associated with survival outcomes of other cancers were included. Overall, none of the variants were observed in statistically significant associations (after FDR correction) with CRC survival under the additive model using the SOCCS cohort. The CRC-risk PRS was not significantly associated with either overall or CRC-specific survival. Stratified analysis did not identify any significant associations after correcting for FDR. Three CRC-risk variants (rs10161980, rs9537521 and rs7495132) showed significant genetic effects (recessive model after FDR correction) on survival outcomes of CRC patients from the SOCCS, and a significant association between the TT genotype of the variant rs7495132 and CRC-specific survival was also observed in the UK Biobank cohort (HR=1.69, 95%CI=1.03-2.79, p=0.038).

In relation to the results of the GWAS, I identified one variant in chromosome 6 (rs143664541) that was significantly associated with both overall and CRC-specific survival (overall survival: HR=1.92, 95%CI=1.52-2.42, p=4.24x10^{-8}; CRC-specific survival: HR=2.17, 95%CI=1.69-2.78, p=1.14x10^{-9}). Another variant in chromosome 9 (rs75809467) was observed to be significantly associated with CRC-specific survival (HR=1.80, 95%CI=1.48-2.20, p=7.07x10^{-9}) of patients from the SOCCS study. However, meta-analysis combining the UK Biobank and the three clinical trials failed to replicate significant associations between the two GWAS-identified variants and overall survival of CRC patients. CRC-specific survival was not investigated in the replication analysis due to lack of available data. In stratified GWASs by AJCC stage, I identified a variant on chromosome 5 (rs323694) that was significantly associated with CRC-specific survival of stage II/III patients from the SOCCS cohort (HR=1.33, 95%CI=1.20-1.47, p=2.92x10^{-8}). Genome-wide gene based analysis revealed significant enrichment of genetic signals in the *CCDC135* gene in relation to CRC-specific survival (p=9.92x10^{-7}). For the gene-set based analysis, significant enrichment of signals was detected in genes involved in the biosynthetic process of galactolipids for overall survival (p=2.09x10^{-6}) and genes associated with up-regulating the differentiation of adipocytes for CRC-specific survival (p=2.52x10^{-7}).

Conclusions

Although the systematic literature review identified no germline genetic variants used as predictors for CRC survival in published prediction models. Five prediction models (Basingstoke score, Fong score, Nordinger score, Peritoneal Surface Disease Severity Score and Valentini nomogram) that include clinic-pathological predictors can potentially be applied to assist clinical decision-making.

This thesis also presents a comprehensive investigation of potential effects of germline genetic variants on survival outcomes of CRC patients. For genetic variants previously linked with CRC survival, the results of the thesis suggest poor

reproducibility of these variants given that none of these associations were successfully replicated in the SOCCS cohort. In addition, the combined effect of the 43 variants, represented by a PRS, on CRC survival is also negligible. There is also very limited predictive value of these variants as a group in predicting survival outcomes of CRC. Although small effects cannot be confidently excluded, major effects of these variants on CRC survival are unlikely.

For genetic variants associated with CRC risk, the lack of association between the CRC-risk PRS and survival outcomes of CRC indicates that the overall genetic susceptibility to CRC has no significant subsequent influence on survival outcomes. For each individual CRC-risk variant, their effects on CRC survival under the additive genetic model are unlikely to be clinically relevant. However, potential genetic effects under recessive model were detected for three CRC-risk variants (rs10161980, rs9537521 and rs7495132) in the SOCCS cohort, especially for the variant rs7495132 whose association with CRC-specific survival was successfully replicated in the UK Biobank cohort. These findings merit further investigation in future large-scaled studies. With respect to genetic variants associated with prognosis of other cancers, the results of this thesis do not support any significant effects of these variants on survival outcomes of CRC patients, indicating that there is a limited shared genetic basis across different types of cancers in terms of survival outcomes.

Although the GWAS-identified variant rs143664541 was not successfully replicated in meta-analysis of results from the UK Biobank and the three clinical trials, effects with concordant direction were observed across all the datasets on overall survival. Therefore, future large-scale investigation of this variant in association with CRC survival outcomes, especially for CRC-specific survival, are warranted. As to the other GWAS-identified variant rs75809467, further investigation in terms of its effect on CRC-specific survival is still needed, although no significant association was found between this variant and overall survival in the replication analysis. A potential variant rs323694 was identified from the GWAS of stage II/III patients. This variant, if replicated in the future, could be of clinical relevance in stratifying stage II/III CRC patients of varied prognostic profiles so as to assist informing tailored treatment strategies. The results of gene and gene-set based analysis provide preliminary evidence favouring future exploration of the biological roles of the *CCDC135* gene and pathways associated with the biosynthetic process of galactolipids and the differentiation of adipocytes in CRC progression.

Lay Summary

Colorectal cancer (CRC) is a group of malignancies located along the human large bowel. It is the second most common cancer and the third leading cause of cancer-related death worldwide. On average, approximately 60% of CRC patients live longer than five years after they are confirmed to have developed CRC. For each individual, however, survival time can vary substantially from less than a year to more than ten years. There has been a growing interest in investigating factors that may explain this patient-to-patient variation in terms of their survival time. Identification of such factors can help clinicians evaluate risk of death and predict how long a patient may survive. A better-informed decision can then be made by clinicians regarding how intensive the treatment strategy should be for the patient.

Researchers investigated families with parents and children both diagnosed with CRC. They found that the children tend to live a shorter life if their parents did not live long after their CRC diagnosis. This indicates that factors that can influence survival time of CRC patients could be inherited from their parents. Genetic information passed from parents to children is recorded in the human DNA sequence. The human DNA is constituted by billions of ‘small units’, known as nucleotides [which form genes], that are inherited from parents, the vast majority of which are consistent across all human beings. However, there are still millions of such units that are different among individuals, and these units are known as genetic variants. Genetic variants are present at varying frequencies in the population and they confer varied effects on a wide range of common human traits such as height, hair colour and susceptibility to different diseases. The thesis employs the survival outcomes of CRC patients as the trait of interest, and investigates associations between this trait and common genetic variants in the DNA sequence of CRC patients by using multiple large-scale human studies in the UK.

Firstly, the thesis comprehensively searched previously published literature to identify studies aiming to forecast survival outcomes of CRC patients. After reviewing 139 relevant publications in depth, no genetic variants had been used to help successfully predict survival outcomes of CRC. Then I searched for candidate variants that could potentially influence survival outcomes of CRC, including genetic variants that had previously been linked to CRC susceptibility or survival outcomes of other cancers. Associations between these variants and CRC survival were then examined. Three genetic variants, which had been linked to CRC susceptibility, also showed possible effects on survival outcomes of CRC. In addition to previously known variants, I also scanned the entire human DNA including millions of genetic variants in order to identify possible novel variants associated with CRC survival. This resulted in identification of two new variants with possible effects. Although the thesis provides suggestive evidence on associations between several genetic variants and CRC survival, further validation of these variants in other populations of CRC patients can still be beneficial before they are ready to be used as valid predictors in clinical practice. Moreover, investigations in the biological implications of these genetic variants will also be helpful to illuminate their possible roles in the progression of CRC, which will ultimately lead to improved clinical outcomes of CRC patients.

List of Figures

List of Tables

List of Abbreviations

AJCC	American Joint Committee on Cancer
ALL	acute lymphoblastic leukaemia
APC	adenomatous polyposis coli
ASCO	American Society of Clinical Oncology
BMI	body mass index
BTC	Betacellulin
CA19-9	carbohydrate antigen 19-9
CCDC135	Coiled-coil domain-containing protein 135
CCI	Charlson Comorbidity Index
CDF	cummulative distribution function
CEA	carcinoembryonic antigen
CHARMS	CHecklist for critical Appraisal and data extraction or systematic Reviews of prediction Modelling Studies
CIMP	CpG island methylator phenotype
CIN	chromosomal instability
CMS	consensus molecular subtypes
COX-2	Cyclooxygenase-2
CRC	colorectal cancer
CRP	C-reaction protein
CRTC4	CREB regulated transcription co-activator 3
CRUK	Cancer Research UK
CSS	colorectal cancer specific survival
CTC	circulating tumour cells
CTC	circulating tumor cells
ctDNA	circulating tumour DNA
DFS	disease-free survival
DPD	dihydropyrimidine dehydrogenase
dTMP	deoxythymidine monophosphate
dUMP	deoxyuridine monophosphate
EPV	events per variable
ESMO	European Society of Medical Oncology
EVI1	Ecotropic Viral Integration Site 1
FAP	familial adenomatous polyposis
FWER	familywise error rate
GDP	guanosine triphosphate
GP	genetic predictor
GTex	Genotype-Tissue Expression project
GWAS	genome-wide association study
HES	Hospital Episodes Statistics
HNPCC	hereditary nonpolyposis colorectal cancer
HR	hazard ratio
IARC	International Agency for Research on Cancer
ICD	International Classification Diseases
IQCM	IQ motif containing M
IQR	interquartile range

IVW	inverse-variance weighted estimator
LASSO	least absolute shrinkage and selection operator
LD	linkage disequilibrium
LncRNA	long non-coding RNA
LP	linear predictor
MAF	minor allele frequency
MAP	MUTYH-associated polyposis
MAPK	mitogen-activated protein kinase
MAR	missing at random
MCAR	missing completely at random
MDS1	Myelodysplastic Syndrome 1
MMR	mismatch repair
MNAR	missing not at random
MSI	microsatellite instability
OS	overall survival
PFS	progression-free survival
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRS	polygenic risk score
QQ	quantile-quantile
RCT	randomised controlled trial
RFS	recurrence(relapse)-free survival
ROC	receiver operating characteristic curve
SCNA	somatic copy number alterations
SEER	The Surveillance, Epidemiology, and End Results Programme
SF	shrinkage factor
SNP	single nucleotide polymorhpism
TB	tumour budding
TCGA	The Cancer Genome Atlas
TGF	transforming growth factor
TNF	tumour necrosis factor
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
TS	thymidylate synthase
URL	uniform resource locator
VEGF	vascular endothelial growth factor
WHO	World Health Organisation

Contents

Chapter 1 Background

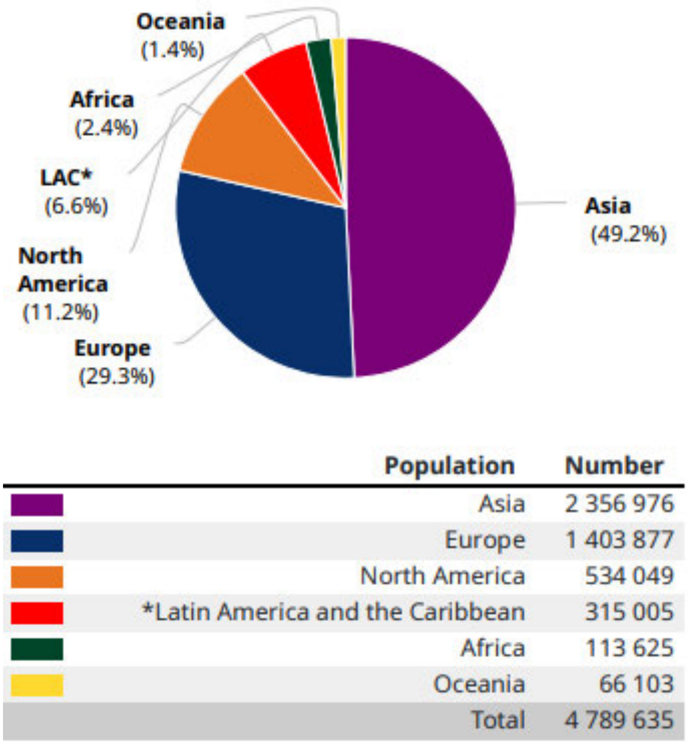
1.1 Introduction

Colorectal cancer (CRC), also known as bowel cancer, includes any malignancies located along the large intestine (from the cecum to the anorectal ring). It arises from the inner wall of the intestine. According to the anatomic site, CRC can be generally divided into colon and rectal cancer. In this chapter, background knowledge regarding CRC risk, diagnosis and prognosis will be introduced, so as to lay the foundation on which the research aims of this thesis will be proposed. Currently, CRC is the second commonest cancer worldwide (Bray *et al*, 2018), and this chapter will start by presenting the disease burden of CRC using the latest population-based statistics regarding both the prevalence and incidence rates of CRC. The metric of prevalence measures the number (or the proportion) of cases in a specific population at a given time point, whereas the incidence—the number of newly diagnosed CRC cases per population at risk—conveys probability of developing CRC during a given time period. Common risk and protective factors which can potentially affect CRC risk will then be introduced. In addition to epidemiological observations, biological evidence will also be reviewed to introduce the genetic and molecular pathogenesis of CRC. The second section features the diagnosis of CRC. In lieu of discussing the routine process and technologies employed in making CRC diagnosis, the focus of this section will be on diagnostic features that can inform treatment strategies and long-term prognosis, including tumour stage, histological type and grade. The last section will describe prognosis of CRC. In particular, population-based estimates on mortality and survival will be presented. Mortality is defined by the number of cases (or the proportion) in the general population who died of CRC within a specific time span. As opposed to mortality which is measured in the general population, survival rates are defined by the probability of being alive, or free of specific events such as postoperative recurrence, for CRC patients within a given time after their diagnosis. Common prognostic factors will then be presented with the main focus on genetic factors.

1.2 Colorectal cancer risk

1.2.1 Prevalence

According to the newly-released prevalence estimates from the International Agency for Research on Cancer (IARC) and the World Health Organisation (WHO), there are a total of 4.8 million people worldwide, as of 2018, living with colorectal cancer (CRC) within five years since the initial diagnosis (Cancer Fact Sheets 2018, IARC/WHO, URL1−1). Geographically, the highest prevalence is observed in Asia, accounting for 49.2% of the total number of cases. Europe ranks 2nd, accountable for 29.3% of the global number of CRC cases. A pie chart of geographic distribution of global CRC prevalence is shown in **Figure 1-1**.



	Population	Number
🟪	Asia	2 356 976
🟦	Europe	1 403 877
🟧	North America	534 049
🟥	*Latin America and the Caribbean	315 005
🟩	Africa	113 625
🟨	Oceania	66 103
	Total	4 789 635

Figure 1-1 Geographic distribution of 5-year global prevalence (2013-2018) of colorectal cancer. Reprinted from Cancer Fact Sheets, colorectum and anus (C-18-21), Copyright (2018) (URL1-1) with written permission from IARC/WHO.

In the UK, there was an estimate of 0.24 million individuals living with CRC in 2010 ever since their initial diagnoses. This number is projected to increase to 0.34 million by 2020 (approximately 1,021 survivors per 100,000 population)(Maddams *et al*, 2012). In Scotland, the latest estimates by Scottish Cancer Registry in 2017 reported an aggregated number of 24,174 survivors with their CRC diagnosed up to 20 years ago, resulting in a prevalence of 0.45% in the Scottish population (Scottish Cancer Registry, URL1-2).

1.2.2 Incidence

Based on the latest *GLOBOCAN* estimates in 2018, there was an estimate of 1.84 million newly diagnosed CRC cases worldwide in 2018 (Cancer Fact Sheets 2018, IARC/WHO, URL1-1). Moreover, the crude global CRC incidence is projected to reach more than 2.2 million newly-diagnosed CRC cases per year by 2030 (Arnold *et al*, 2017). Amongst all cancer types, CRC is the second most commonly diagnosed cancer, accounting for 10.2% of the 18.1 million newly-diagnosed cancer cases worldwide (Bray *et al*, 2018).

Given that the age structure of a certain population significantly affects the CRC incidence rate, the incidence rate is often standardised by taking the weighted mean of crude rates in each age groups to derive the age-standardised rate when comparing incidence rates across different populations. Presented in **Figure 1-2** is the age-standardised CRC incidence rate worldwide (Torre *et al*, 2015). As the figure indicates, higher incidence rates are widely observed in well-developed areas including Europe, North America and Oceania. Taking UK as an example, there were approximately 42,000 new CRC cases per year from 2014 to 2016, accounting for around 12% of all new cancer cases according to the data from Cancer Research UK (CRUK) (Bowel Cancer Statistics, CRUK, URL1-3). The age-standardised incidence rate in the UK was 69.3 per 100,000 population in 2016. In Scotland, there were 3,776 newly diagnosed CRC cases in 2017 with an age-standardised incidence rate of 73.7 per 100,000 population (Scottish Cancer Registry, URL1-2). In the USA, there was an average of 145,600 new CRC cases diagnosed per year from 2014 to 2016, with an age-standardised incidence rate of 38.6 per 100,000 population (Marley & Nan, 2016). Whilst absolute incidence rates remain highest in these well-developed countries, there has been a stable or declining trend of CRC incidence (Arnold *et al*, 2017) (URL1-1). The age-standardised CRC incidence rates in Scotland from 1993 to 2017 are plotted in **Figure 1-3** (Scottish Cancer Registry, URL1-2).

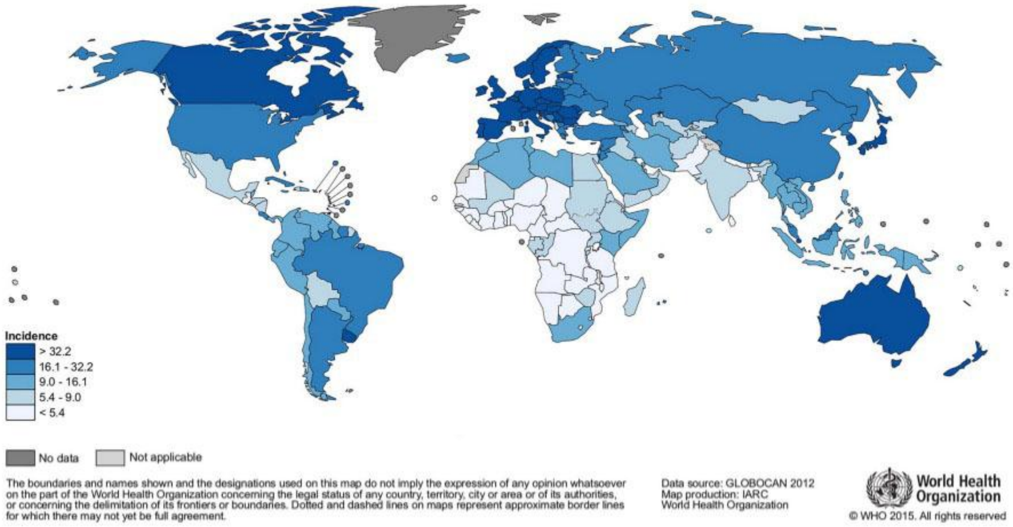

Incidence
> 32.2
16.1 - 32.2
9.0 - 16.1
5.4 - 9.0
< 5.4
No data
Not applicable
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.
Data source: GLOBOCAN 2012
Map production: IARC
World Health Organization
World Health Organization
© WHO 2015. All rights reserved

Figure 1-2 Worldwide colorectal cancer incidence rates (age adjusted according to the world standard population, per 100 000) in 2012. Adapted from (Arnold et al, 2017) with permission.

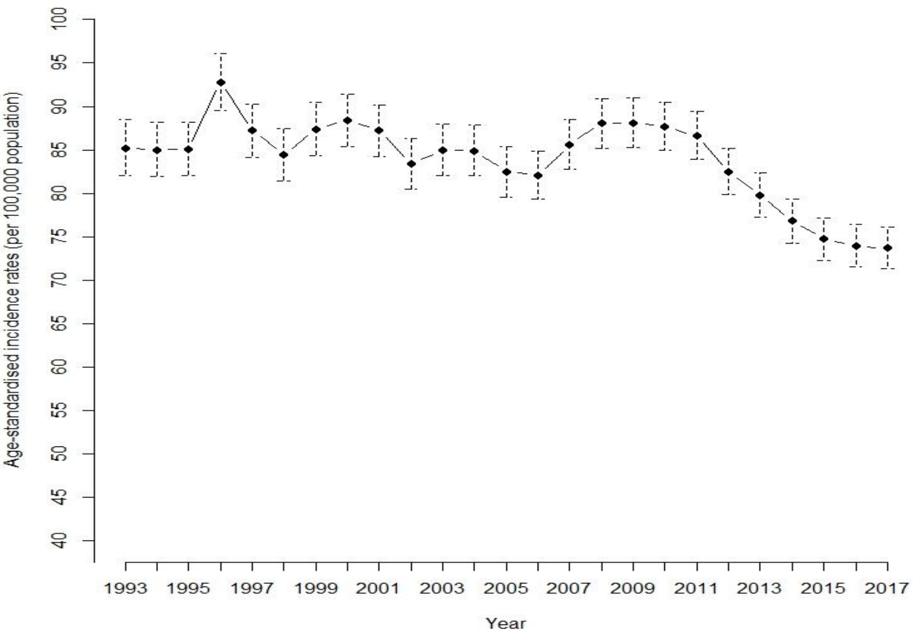


Figure 1-3 Colorectal cancer incidence trend in Scotland from 1993 to 2017.Created using data from the Scottish Cancer Registry (Scottish Cancer Registry, URL 1-2).

However, there are widening disparities in CRC incidence patterns, primarily between well-and less-developed countries. For less-developed areas where historically there had been lower CRC incidence rates, the number of newly diagnosed CRC cases has kept rising over recent decades. For instance, in China, the age-standardised incidence rate of CRC has increased from 14.3 to 25.3 per 100,000 population from 1990 to 2016 (Zhang *et al*, 2019). Similarly, in Thailand, this number is projected to increase by 42% from 2000 to 2025 (Virani *et al*, 2017). This increase could be attributed to the changing diet patterns, obesity and other lifestyle risk factors that can potentially increase CRC risk (Bray *et al*, 2018).

Stratified by sex, CRC is the third commonest cancer in men (10.9% in all men with CRC) following lung and prostate cancer. For women, it ranks fourth (9.5%) among all cancer types (Bray *et al*, 2018). The 2018 *GLOBOCAN* estimates indicate that men are subject to slightly higher CRC risk than women (23.6 vs. 16.3 per 100,000 population) (Bray *et al*, 2018). Although the magnitude varies, this sex difference in CRC incidence is consistent over the globe (Cancer Fact Sheets 2018, IARC/WHO, URL1-1). Colorectal cancer incidence rate increases with age. In the UK, approximately 44% of new CRC cases were diagnosed at the age of 75 or older in 2018 (Bowel Cancer Statistics, CRUK, URL1-3). **Figure 1-4** presents the distribution of age at diagnosis separated by sex in the UK from 2014 to 2016. The CRC incidence rates for both men and women peak at individuals between 85 to 89 years of age.

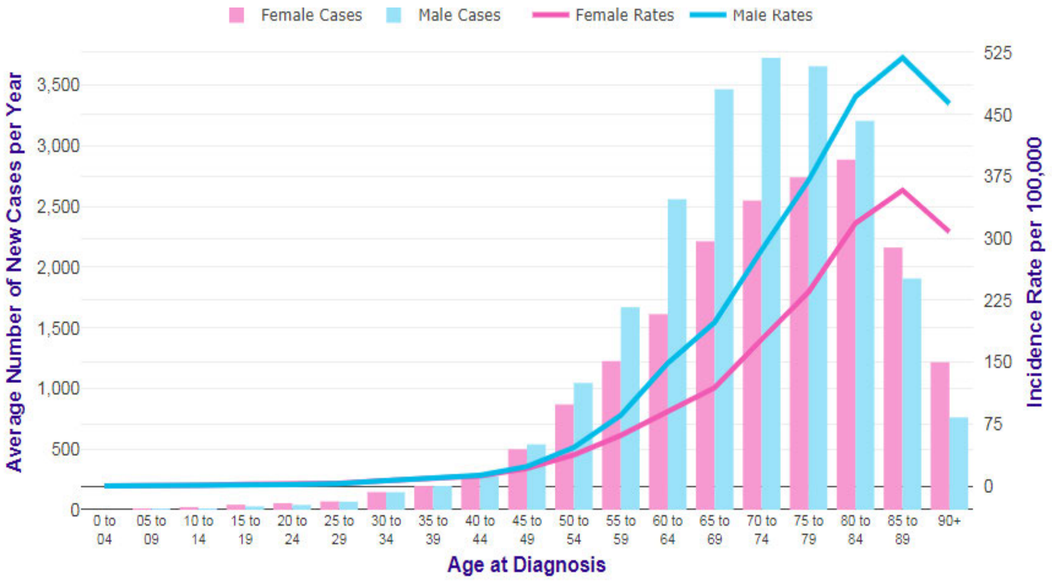


Figure 1-4 Distribution of age at colorectal cancer diagnosis in the UK from 2014 to 2016.Reproduced with permission from the graph created by Cancer Research UK (Bowel Cancer Statistics, CRUK, URL1-3).

Common risk and protective factors

Colorectal cancer can be categorised into familial and sporadic disease. The heritable components of CRC will be introduced in the next section. Sporadic CRC accounts for up to 60% of existing CRC cases, in which no family history has been reported at diagnosis. Aside from aforementioned demographic factors such as gender and age, epidemiological studies have identified a wide range of other factors associated with CRC risk. According to recommendations from the World Cancer Research Fund (WCRF) (Bowel Cancer, WCRF, URL1-4) and a systematic literature review by Johnson and colleagues (Johnson *et al*, 2013), commonly known risk and protective factors that are categorised as strong evidence are summarised as follows:

Risk factors:

—red and processed meat intake (Chao *et al*, 2005)

—cigarette smoking (Botteri *et al*, 2008)

—obesity (Edge *et al*, 2010)

—alcohol consumption (Fedirko *et al*, 2011)

—personal history of polyps and inflammatory bowel diseases (Munkholm, 2003)

Protective factors:

— physical activity (Wolin *et al*, 2009)

—dietary intake of fibre (Negri *et al*, 1998)

—Dairy products (Aune *et al*, 2012)

—Calcium supplements (Huncharek *et al*, 2009)

It is worth noting that the CRC screening strategy varies across the world and it per se can also influence the observed CRC incidence rate due to more cases detected. Colorectal cancer screening in the UK includes three tests: faecal immunochemical test, faecal occult blood test, and colonoscopy. Individuals over 60 years of age (50 for Scotland) are invited to participate the screening.

There are other factors identified by the systematic review that are associated with CRC risk. For example, individual studies reported that aspirin intake (Rothwell *et al*, 2010) and hormone replacement therapy in women (Johnson *et al*, 2009) were associated with lower risk of CRC. However, no significant associations were identified in meta-analyses of all published studies (Johnson *et al*, 2013). Therefore, more evidence is needed before any recommendations can be made.

These aforementioned modifiable risk and protective factors can have potentially far-reaching implications for CRC prevention provided that causal effects in these observed associations are established by randomised clinical trials (e.g. calcium intake) or other approaches like Mendelian randomisation studies (e.g. obesity) if clinical trials are not feasible.

1.2.4 Genetic and molecular pathogenesis

Genetic models of tumorigenesis

Colorectal cancer occurrence follows a well-understood transformation pattern—it starts from normal colorectal epithelium to benign adenomas and eventually progresses to invasive and metastatic CRC. This process is accompanied by stepwise accumulation of both germline and somatic genetic alterations.

As with other malignancies, CRC is in fact a group of heterogeneous conditions that encompasses a number of subtypes characterised by distinct genetic patterns. In particular, CRC can be classified into sporadic (50-60%), familial (30-40%) and hereditary (4-6%) CRC (American Society of Colon and Rectal Surgeons, ASCRS, URL 1-5). Sporadic cases, with no family history, are more likely to be diagnosed after the age of 50. Recently, an increased incidence of sporadic CRCs has been observed among young adults (<50 years of age) from high-income countries (Araghi *et al*, 2019). The formation of sporadic CRC features sequential acquisition of somatic mutations. In contrast, hereditary CRCs tend to occur at an earlier age with an identifiable high-penetrance germline predisposition. These hereditary cases can be further divided into groups of syndromes: one group manifests colonic polyposis, including familial adenomatous polyposis (FAP), MUTYH-associated polyposis (MAP) and other less commonly observed syndromes; the other group manifests without polyposis—namely hereditary nonpolyposis CRC (HNPCC, also known as Lynch syndrome).

A comparison between genetic patterns of sporadic and hereditary CRC from normal epithelium to invasive cancer is presented in **Figure 1-5**. For sporadic CRCs, multiple somatic mutations are sequentially accumulated to trigger formation of neoplastic lesion as well as to expedite progression to invasive cancer. Whereas in the case of hereditary CRCs, germline genetic alterations play an essential role in tumorigenesis. The distinction between polyposis and nonpolyposis hereditary CRC lies in the specific phase when germline genetic alterations take effect. For individuals with polyposis syndromes like the FAP, germline alterations mainly accelerate the formation of adenomas, but for nonpolyposis CRC, such as the Lynch syndrome, germline alterations primarily affect the progression rate from adenomas to invasive cancers (**Figure 1-5)**.

The third type is familial CRC. The genetic mechanisms of this type of CRC have been less clearly understood. These cases present inconsistent patterns compared with the aforementioned inherited syndromes. A family history is reported in these individuals who are at significantly higher risk of developing CRC. It is estimated that having a single first-degree relative diagnosed with CRC imposes a twice as high risk onto an individual compared with the general population (Tuohy *et al*, 2014).

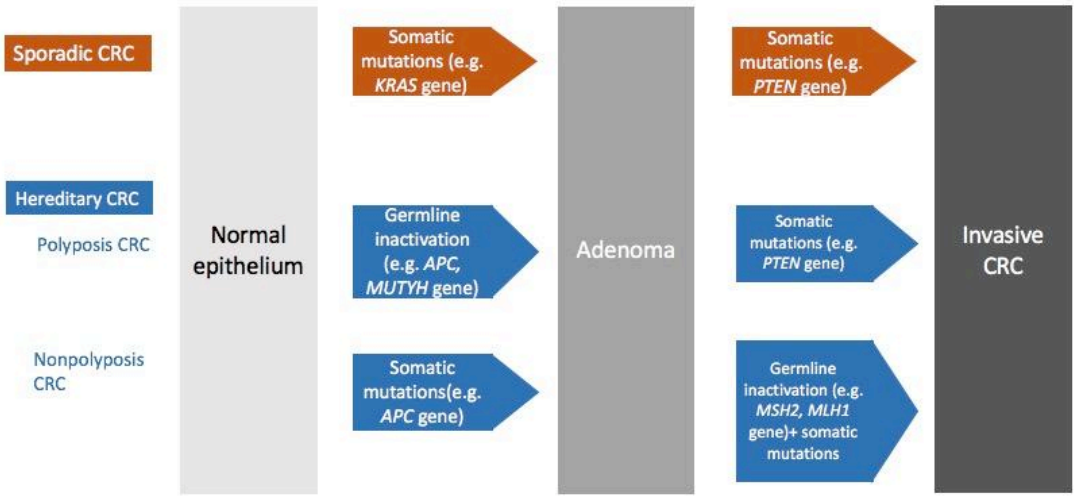

Sporadic CRC
Hereditary CRC
Polyposis CRC
Nonpolyposis CRC
Normal epithelium
Somatic mutations (e.g. KRAS gene)
Germline inactivation (e.g. APC, MUTYH gene)
Somatic mutations(e.g. APC gene)
Adenoma
Somatic mutations (e.g. PTEN gene)
Somatic mutations (e.g. PTEN gene)
Germline inactivation (e.g. MSH2, MLH1 gene)+ somatic mutations
Invasive CRC

Figure 1-5 Comparison of genetic models between sporadic and hereditary colorectal cancer. Recreated based on (Fearon, 2011).

Genetic mutations

Germline mutations

These mutations occur within germ cells and can be passed on to off-spring. Mutations in potential pathogenic genes of the CRC tumorigenesis can exert large effects on increased CRC risk (also known as high penetrance), although these mutations are usually rare (<1%) in the general population.

With regard to pathogenic genes responsible for CRC development, the *Adenomatous polyposis coli (APC)* gene is a tumour suppressor gene that has been extensively studied and plays a critical role in formation of colorectal adenomas. The APC protein, encoded by the *APC* gene, is able to assist in maintaining cell division, adhesion, and chromosome stability, thus can serve as a key suppressor that prevents uncontrolled cell proliferation (Fodde, 2002). Germline mutations in the *APC* gene can lead to premature truncation of the APC protein, rendering it unable to effectively control overgrowth of cells that eventually develop into adenomas (Fodde, 2002). An estimate of 90% individuals affected by FAP carry germline defects in the *APC* gene (Fearon, 2011). Besides FAP, inactivation of the *APC* gene is also involved in other hereditary CRC syndromes with polyposis such as Gardner syndrome and attenuated adenomatous polyposis coli.

The *MUTYH* gene is another common gene that harbours mutations causing another type of hereditary polyposis CRC syndrome, namely the MAP. The *MUTYH* gene encodes the MUTYH glycosylase which is engaged in the base excision repair pathway. In particular, this enzyme can excise adenine bases when nucleotide bases are incorrectly paired mostly due to oxidative DNA damage (Sampson *et al*, 2005). Homozygous germline mutations in the *MUTYH* gene can cause failure of DNA damage repair and subsequently provoke somatic mutations in relevant oncogenes—genes that can potentially cause cancer growth.

Pertaining to HNPCC, the genetic basis is constituted of germline mutations in a single copy of allele in one of the DNA mismatch repair (MMR) genes, for example the *MSH2* or *MLH1* gene, while the other allele is somatically silenced via mechanisms including loss of heterozygosity and promoter hypermethylation (Peltomaki, 2001). Loss of heterozygosity refers to a common genetic event in carcinogenesis where a heterozygous locus in the germline turns into homozygous in the tumour DNA. Promoter hypermethylation is an epigenetic change (inheritable alteration not involving the DNA sequence) in the promoter region featuring enrichment of CpG islands (A cytosine base followed immediately by a guanine base). Impaired DNA mismatch repair can subsequently lead to microsatellite instability (MSI)—changes in the number of repeats of short sequence in the tumour DNA— which is widely-accepted as one of the key activators of CRC tumorigenesis.

Somatic mutations

In addition to germline genetic defects, the role of acquisition of somatic mutations is also indispensable throughout CRC pathogenesis, particularly for the most common type of sporadic CRCs. Common somatic mutations in both oncogenes and tumour suppressor genes involved in the CRC tumorigenesis are listed in **Table 1-1**.

Table 1-1 Reported mutation frequencies of selected common somatic mutations observed in colorectal cancer tumour tissues

Gene	Mutation frequency*
Oncogenes	
KRAS	35-45%
PIK3CA	15-25%
BRAF	4-18%
EGFR	5-15%
CDK8	10-15%
CMYC	5-10%
NRAS	3-5%
Tumour suppressor genes	
APC	70-80%
p53	40-70%
FBXW7	14-20%
PTEN	10-20%
SMAD4	10-15%
SMAD2	5-10%

*Frequency estimates are based on references (Fearon, 2011) (Kudryavtseva *et al*, 2016; Li *et al*, 2015; Molinari & Frattini, 2013; Nguyen & Duong, 2018; Yeh *et al*, 2018).

As a well-established gatekeeper of tumorigenesis, the *APC* gene can also mutate somatically to initiate the formation of colorectal adenomas. An estimate of 80% of colorectal adenomas (hereditary and sporadic combined) are driven by defects in the

APC gene which initiate a whole chain of genetic alterations (Fearon, 2011). Approximately 80 somatic mutations are expected to be detected in a given CRC sample, among which an average of 15 are predicted to be potential driver mutations throughout CRC tumorigenesis (Wood *et al*, 2007).

A major group of oncogenes is the *RAS* family which consists of *KRAS, NRAS* and *HRAS* genes. Proteins encoded by these genes, namely K-RAS4A, K-RAS4B, H-RAS and N-RAS, are guanosine triphosphate (GTP) binding proteins that participate in transforming extracellular signals to intracellular regulatory factors, such as cell cycle proteins, so as to mediate cell differentiation, proliferation and apoptosis (Jinesh *et al*, 2018). The *RAS* family is among the most frequently observed genes with mutations associated with tumorigenesis of many human malignancies. With respect to CRC, approximately 40% of CRC cases carry *KRAS* mutations and less than 5% carry *NRAS* mutations (Fearon, 2011). Mutations in the *KRAS* gene occur in a relatively early stage of the CRC tumorigenesis, and can provoke the mitogen-activated protein kinase (MAPK) pathway, rendering uncontrolled cell growth (Jinesh *et al*, 2018).

Around 10% of CRCs are identified with mutations in another important oncogene known as the *BRAF* gene (Barras, 2015). The BRAF protein is a serine/threonine protein kinase that plays a vital role in key regulatory pathways related to carcinogenesis such as the Hippo signalling pathway which regulates cell proliferation and apoptosis. Hence, mutations in the *BRAF* gene exert profound effects in the CRC tumorigenesis (Barras, 2015).

In addition to these oncogenes above, there are also tumour suppressor genes harbouring somatic mutations engaged in CRC tumorigenesis. For instance, the *p53* gene is among the most well-known tumour suppressor genes and mutations in this gene are commonly observed in up to 70% of CRC cases (Fearon, 2011). The p53 tumour suppressor protein can bind to DNA sequences and regulate the expression of many vital genes that can maintain cell cycle, and apoptosis, particularly under circumstances with stimuli like DNA damage and hypoxia (Ozaki & Nakagawara, 2011). Mutations in the *p53* gene are frequently clustered in the p53 DNA binding domain, and they can impair normal p53 function by hindering it binding to relevant DNA sequences (Li *et al*, 2015). This effect is thought to be active in the transformation from adenoma to invasive CRC (Lopez *et al*, 2012).

Molecular pathways

On the basis of molecular events that have been partially discussed above, CRC tumorigenesis is generally classified into three molecular pathways: (a) the chromosomal instability (CIN) pathway, (b) the microsatellite instability (MSI) pathway and (c) the CpG island methylator phenotype (CIMP) pathway. Of note, these three pathways are not mutually exclusive, thus can be observed simultaneously in the same CRC sample.

The CIN pathway can be triggered both in sporadic and hereditary CRCs. Primarily, it is characterised by chromosomal abnormalities such as insertions, deletions or loss of heterozygosity. These abnormalities are frequently paired with inherited or acquired mutations that can potentially activate relevant pathways of CRC tumorigenesis (Pino & Chung, 2010). The second pathway that features MSI is typically identified in HNPCC. Germline mutations in MMR genes lead to dysfunction of MMR enzymes and, as a consequence, failures to repair mismatches are accumulated throughout the genome. Colorectal cancers that develop through this pathway can be typically identified with high level of MSI (noted as MSI-high tumours). With respect to the CIMP pathway, epigenetic alterations such as DNA methylations can suppress the level of the relevant oncogene and tumour suppressor gene expression, including but not limited to MMR genes. In particular, CRCs characterising the CIMP pathway are enriched with methylated CpG islands (Weisenberger *et al*, 2006). For instance, methylation of the CpG islands in the promoter region of the *MGMT* gene, which encodes the O6-methylguanine DNA methyltransferase (MGMT), can inhibit the corresponding gene expression, leading to aberrant genomic alterations (Inno *et al*, 2014). Common molecular events of the three pathways are summarised in **Table 1-2**

Table 1-2 Genetic characteristics of molecular pathways of colorectal cancer tumorigenesis

	CIN pathway	**MSI pathway**	**CIMP pathway**
Prevalence	60-70%	~15%	15-20%
Heredity	Hereditary/sporadic	Hereditary	Hereditary/sporadic
Genetic markers			
MSI status	MSS	MSI-high	MSI-high/low
CIN	+	-	-
***KRAS* mutation**	+	+/-	-
***BRAF* mutation**	-	-	+
***MLH1* status**	Normal	Mutated	Methylated
***MGMT* methylation**	-	-	+

CIN, chromosomal instability; MSI, microsatellite instability; MSS, microsatellite stability; CIMP, CpG island methylator phenotype. This table is created based on references (Al-Sohaily *et al*, 2012) and (Noffsinger, 2009).

Common germline genetic variations

Unlike hereditary CRC cases where major pathogenic germline mutations can often be ascertained, the mass of sporadic as well as familial CRC cases remains to be investigated in the sense of their genetic components related to higher CRC risk. There has been a growing awareness that the development of these types of CRC is attributed to complex interactive effects of the patients' genetic background and environmental risk factors.

Human genomes differ from each other at an estimated 4 to 5 million sites, of which more than 99.9% are single nucleotide polymorphisms (SNP) and short indels (Genomes Project *et al*, 2015). SNPs are substitutions of single nucleotides that occur at specific genetic loci with a relatively common frequency (>1%) in a given population.

Other variations include copy number variations and chromosomal translocations. These common genetic variations have been widely linked with a wide spectrum of phenotypes and can modify the susceptibility to diseases such as CRC, although the penetration is comparatively lower than that of rare germline mutations. Since early 2000s, the emergence of genome-wide association studies (GWAS) has successfully mapped numerous genetic risk loci to complex phenotypes including human malignancies. Additional details regarding study design and analysis of GWASs will be introduced at length in later chapters.

As to CRC, the two most recent (to date) meta-analyses of large GWASs identified over 100 independent risk loci throughout the human genome (Huyghe *et al*, 2019; Law *et al*, 2019). Among these hits, most of them are mapped to non-coding regions in the genome—DNA sequences that do not encode proteins. Although this indicates mostly regulatory effects of these genetic variants, some of them are enriched in or near genes involved in several known cancer related pathways such as the transforming growth factor beta (TGF-β) pathway and the Wnt-pathway, as well as other pathways responsible for immune response, cell apoptosis and differentiation (Law *et al*, 2019). Other variants may assist in revealing novel pathways related to CRC tumorigenesis that are not understood extensively. A list of basic characteristics of common genetic variations identified by GWASs that are associated with CRC risk are presented in Chapter 5 (**Table 5-19**). Notably, these CRC-risk variants were identified mostly in the European population, with a small number of variants that were uniquely identified in the Eastern Asian population. It is estimated that the common variants identified to date in European populations can explain approximately 11% of the 2.2 fold familial relative risk for CRC (Law *et al*, 2019).

Besides GWASs, a large number of hypothesis-driven studies of smaller scale have been conducted in order to explore associations between specific candidate genetic variants and CRC risk. Field synopsis is a systematic approach using meta-analysis and established criteria to summarise and appraise published candidate genetic association studies. The latest field synopsis by our group highlighted credible associations between 18 common genetic variants and CRC risk (Montazeri *et al*, 2019). These findings are highly concordant with previous GWASs.

It should be noted that these common genetic variants identified by association studies could either be or near the potential driver variant that needs to be determined by functional characterisation. Although evidence for the biological function of these genetic variants is still sparse, observed associations between these variants and CRC risk can be harnessed to improve risk prediction and stratification in the general population.

1.3 Diagnosis of colorectal cancer—staging, typing and grading

In clinical practice, a diagnosis of CRC is accompanied by information on the characteristics of the tumour, so as to assist clinicians in evaluating how serious the cancer is, predicting prognosis and determining potential treatment strategies. These characteristics mainly include the stage, type and grade of the tumour.

Staging

Staging is the process during which clinicians assess how much the tumour has grown and spread. In particular, the depth of tumour invasion into the bowel wall, the involvement of local lymph nodes and distant organs are evaluated. Staging assesses the extent of tumour progression for CRC patients and, therefore, it is currently considered as the principal prognostic indicator. Patients diagnosed with more advanced stage manifest poorer prognosis, thus are subject to more aggressive therapy. There have been two widely-used staging systems, namely the American Joint Committee on Cancer (AJCC) TNM system and the Dukes' system.

The Dukes' system—a specific classification system for CRC— was first proposed in 1932 by the British pathologist Cuthbert Dukes (1890-1977). It categorises CRC based on whether the tumour is local, regional (local lymph nodes involved) or metastatic. Although followed by several modified versions (Kyriakos, 1985), this system now serves mainly historical purposes and is rarely applied in today's practice.

The TNM staging system, however, is used more widely in recent clinical practice as it is also generally applicable to other cancers with the exception of leukaemia and tumours of the central nervous system. As suggested, this system evaluates three aspects of the bowel tumour: **T** (tumour) measures the extent of tumour growing into the bowel wall which is composed of an inner layer (mucosa), a middle layer (submucosa and muscle), and an outer layer (subserosa and serosa); **N** (nodes) represents any involvement of nearby lymph nodes; **M** (metastasis) shows whether the tumour has spread to any distant lymph nodes or organs such as the liver and lung. **Figure 1-6** depicts CRCs at different stages in relation to the basic anatomy.

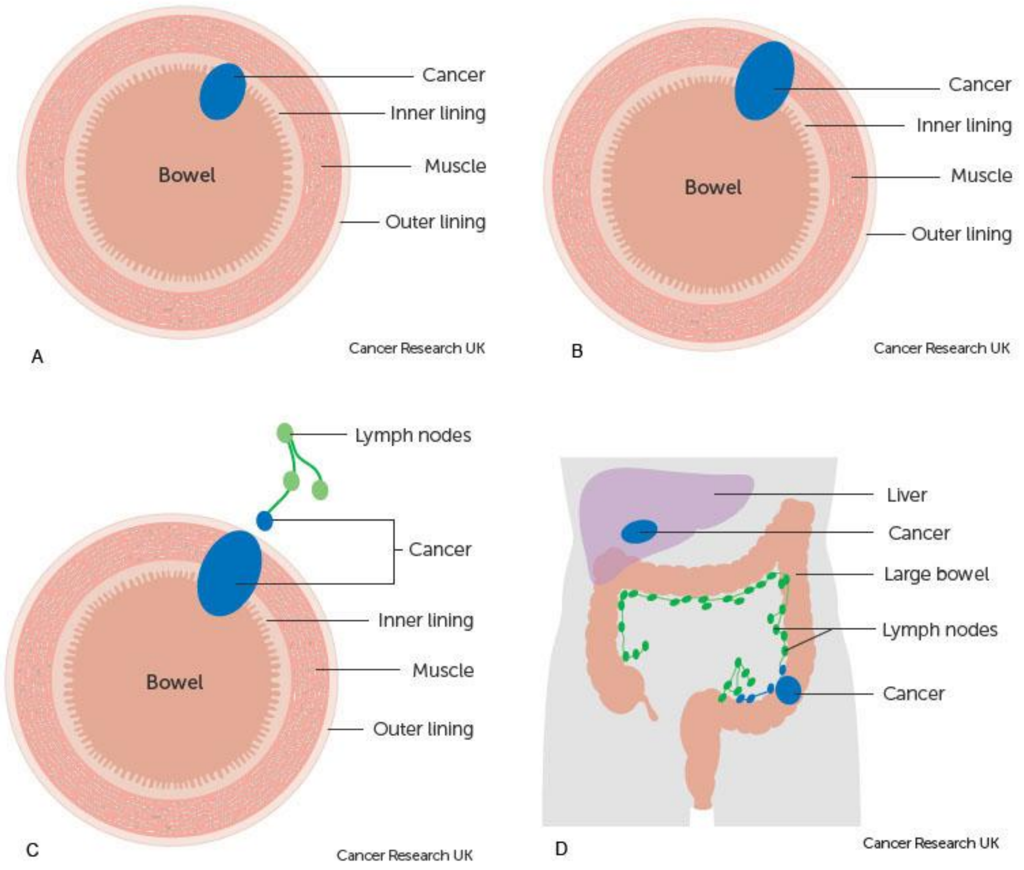


Figure 1-6 Anatomic diagram of colorectal cancer at different stages. A: stage I (Dukes' A); B: stage II (Dukes' B); C: stage III (Dukes' C); D: stage IV (Dukes' D). Adapted with permission based on graphs created by Cancer Research UK (Bowel Cancer Stages, Types and Grades, CRUK, URL1-6).

According to combinations of different measures in these three aspects, Roman numbers of 0 to IV are assigned to group CRC into five stages (the AJCC stage). Table 1.3 presents the basic classification rules of stage 0 to IV based on the 8^{h} edition TNM staging system of CRC (Colorectal Cancer Stages, American Cancer

Society, URL1-7). Each stage can be further divided into sub-stages according to detailed TNM measures (details can be found in URL1-7).

Table 1-3 Basic classification rules for the TNM and the Dukes' staging systems

AJCC stage	**Dukes' stage**	**Stage grouping**	**Description**
0	None	Tis N0 M0	Also known as carcinoma in situ (Tis). It has not grown beyond the inner layer (mucosa) of the bowel wall.
I	A	T1-2 N0 M0	The tumour has grown through the mucosa into the submucosa (T1), or into the muscularis propria (T2), but has not spread to nearby lymph nodes (N0) or to distant sites (M0).
II	B	T3-4 N0 M0	The tumour has grown into the outer layers (subserosa and serosa) but has not gone through them (T3). T4 refers to the tumour that has grown through the bowel wall with or without attachment to nearby tissues or organs. It has not spread to nearby lymph nodes (N0) or to distant sites (M0).
III	C	T1-4 N1-2 M0	The tumour has spread to 1 to 3 (N1) or more than 3 nearby lymph nodes. It has not spread to distant sites (M0).
IV	D	Any T Any N M1	The tumour has spread to distant parts of the peritoneum, distant set of lymph nodes or distant organs (M1).

By default, the staging system above refers to the pathological stage (noted as pTNM) which is ascertained by examining the tumour tissue removed from surgery. There is also clinical staging (noted as cTNM) which occurs prior to surgery, and is ascertained by evaluating results of physical examination, biopsies and imaging. Although not as accurate as pathological staging, clinical staging can inform necessary neoadjuvant therapy (chemotherapy or radiotherapy applied prior to surgery) and decision on the most appropriate surgical approach. Given the fact that neoadjuvant therapy can potentially downgrade the tumour stage, a prefix ‘y’ symbol is added to indicate staging after neoadjuvant treatment (noted as ypTNM).

Typing

The World Health Organisation first introduced a histological typing system in 1970 (Morson & Sobin, 1976). Up to 96% of the CRCs are classified as adenocarcinomas (URL1-6) which stem from the gland cells in the lining of the bowel wall. In addition to classical colorectal adenocarcinomas, there are two less frequent subtypes: mucinous and signet ring carcinomas. Not only do these subtypes morphologically differ under the microscope, they are related to distinct tumour biology, thus indicative of varied prognostic outcomes. There has been evidence reporting that patients with mucinous and signet ring carcinomas tend to be diagnosed with more advanced stage, and can be predisposed to worse survival (Nitsche *et al*, 2013). Other types of CRC such as squamous cell carcinoma, stromal tumours and carcinoid tumours are too rare to have been extensively investigated.

Grading

Analogous to the staging, a grading system has been introduced to provide an overall description on the amount of abnormality of tumour cells under the microscope in an attempt to quantify how fast the tumour is likely to grow and spread (also described as cancer differentiation). Pathologists assign grades from G1 to G4 to the CRC by examining the tumour tissue. Grade 1 (well-differentiated) represents a tumour that looks most like normal cells; grade 2 (moderately-differentiated) denotes a tumour with cells that fall between normal and abnormal; grade 3 (poorly-differentiated) refers to a tumour with abnormal cells; grade 4 (undifferentiated) tumours, are those tumours that look completely different from the tissue from which they originate. At times, G3 and G4 tumours are jointly presented as G3. Tumours of higher grades exhibit more invasiveness than ones of lower grades. Depicted in Figure 1.7 are microscopic views of CRCs from grade 1 to grade 3 (4).

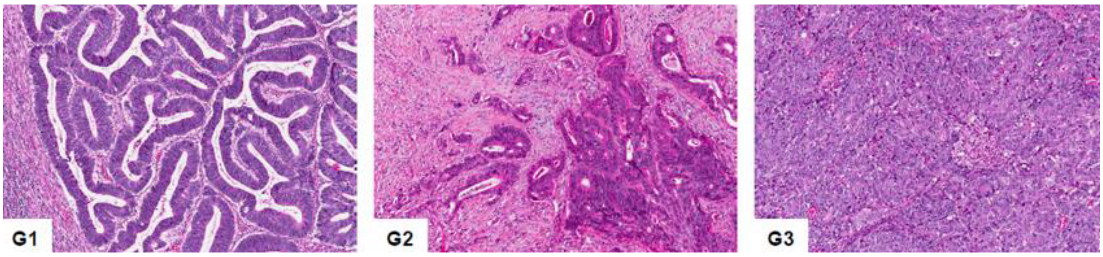
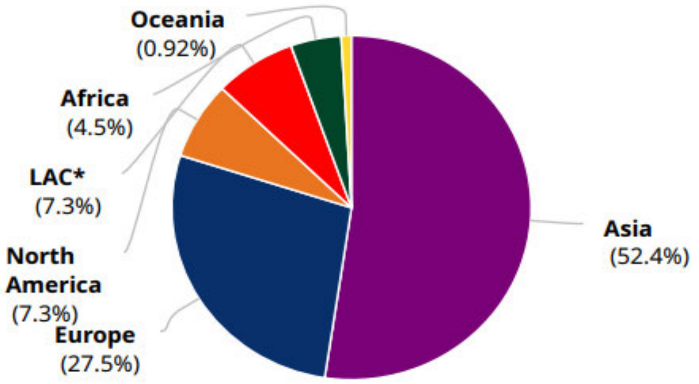


Figure 1-7 Grade 1 to 3 (4) colorectal cancer tissue under a microscope. Reproduced from the Atlas of Genetics and Cytogenetics in Oncology and Haematology with permission (Huret et al, 2013) (URL1-8).

1.4 Colorectal cancer prognosis

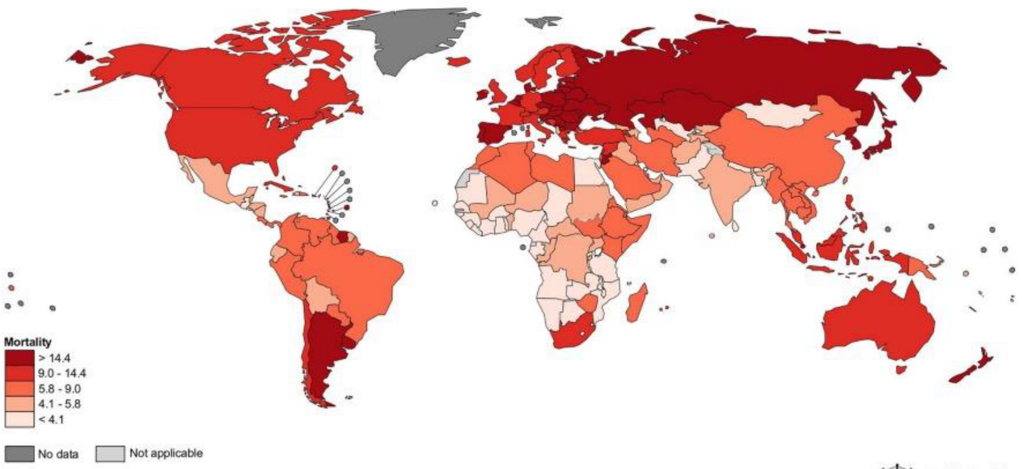
1.4.1 Mortality

In 2018, there was an aggregate number of 0.88 million deaths caused by CRC globally. By 2030, this number is estimated to increase to 1.1 million per year (Arnold *et al*, 2017). Similar to the global distribution of CRC prevalence, roughly 80% of CRC-related deaths occurred in Asia or Europe (ULR1-1). Presented here in **Figure 1-8** is the geographic distribution of CRC-related deaths in 2018. The latest statistics from the IARC indicate that CRC was the second leading cause of cancer-related death in 2018, with 9.2% of all cancer deaths attributed to CRC (Bray *et al*, 2018). With regard to sex-stratified mortality, CRC was the third highest cause of cancer related deaths in females and fourth in males (9.5% of female cases and 9% of male cases) (Bray *et al*, 2018). As shown in **Figure 1-9**, high CRC mortality (age-standardised) is mainly observed in specific areas such as east Europe and South America.



	Population	Number
🟪	Asia	461 422
🟦	Europe	242 483
🟧	North America	64 121
🟥	*Latin America and the Caribbean	64 666
🟩	Africa	40 034
🟨	Oceania	8 066
	Total	880 792

Figure 1-8 Geographic distribution of colorectal cancer related deaths in 2018. Reprinted from Cancer Fact Sheets, Colorectum and anus (C-18-21), Copyright (2018) (URL1-1) with written permission from IARC/WHO.



The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data source: GLOBOCAN 2012
Map production: IARC
World Health Organization

© WHO 2015. All rights reserved

Figure 1-9 Worldwide colorectal cancer mortality rates (age adjusted according to the world standard population, per 100 000) in 2012**.** Adapted from (Arnold *et al*, 2017) with permission.

In contrast to incidence where substantial variation exists between developing and developed countries, the difference of CRC mortality, however, lies primarily in the changing pattern with time where decreasing age-standardised CRC mortality rates have been observed mainly in many developed countries such as the USA, UK and Singapore (Arnold *et al*, 2017; Bray *et al*, 2018). In the UK, there were on average approximately 16,000 CRC deaths every year (44 per day) from 2014 to 2016. Over the past decade, however, the age-standardised CRC mortality rate in the UK has declined by 14% (Bowel Cancer Statistics, CRUK, URL1-3). In Scotland, the CRC mortality rate has been reduced from 47.4 per 100,000 person-years in 1992 to 31.1 in 2017 (detailed numbers for each year plotted in **Figure 1-10**) (Scottish Cancer Registry, URL1-2). Improved patient care along with advancing multi-disciplinary treatment strategies could be behind this. However, relatively stable CRC mortality rates have been observed in developing countries. For example, the CRC mortality rate in China was 11.65 per 100,000 population in 1990 and 11.34 in 2016 (Zhang *et al*, 2019). The stable trend of CRC mortality rates may point to the balance between improvement of treatment strategies and rapid growth of CRC incidence in these areas.

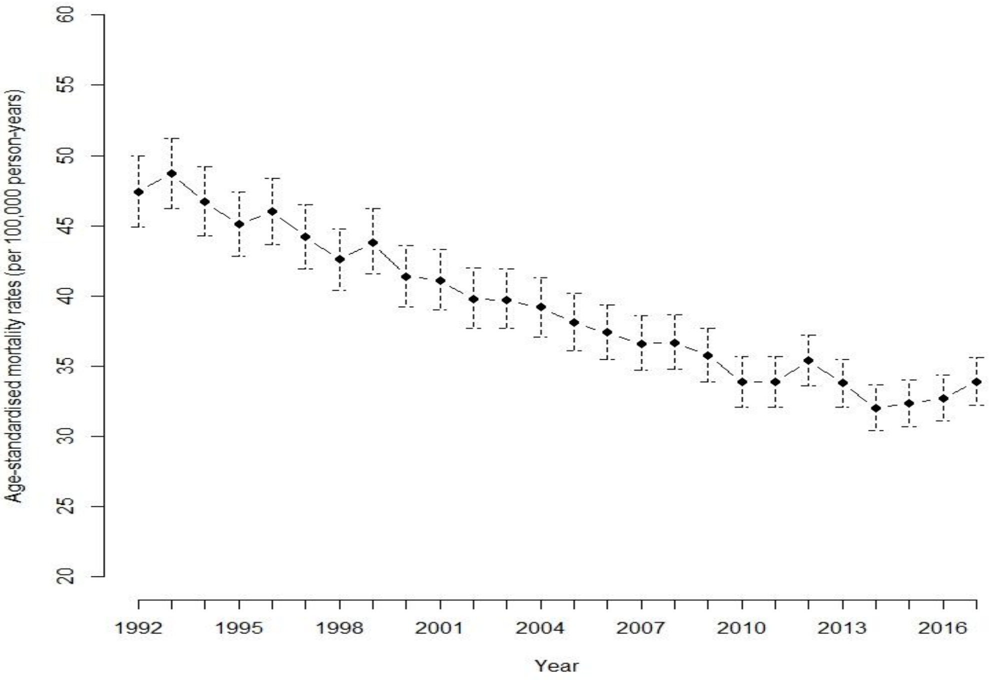


Figure 1-10 Colorectal cancer mortality trend in Scotland from 1992 to 2017. Created with data from Scottish Cancer Registry.

1.4.2 Survival

Given the fact that survival estimates reflect directly the clinical outcome of CRC patients after diagnosis, they are widely used both in evaluating the population disease burden and in the context of clinical research. This metric will therefore be adopted as the study outcome of the thesis.

Based on the cause of death, survival metrics can be divided into overall survival (OS) and CRC-specific survival (CSS)—the probability of surviving in the absence of other causes such as cardiovascular events or car accidents. Deaths of other causes were considered as censored when presenting the CSS. Both the OS and the CSS estimates absolute survival rates in a certain population.

However, in order to compare survival rates of CRC patients across different populations, net survival rates (also known relative survival rates) are widely adopted by taking the ratio of the proportion of survivors in the CRC patient cohort (absolute overall survival rate) to that in a reference cancer-free cohort with comparable common characteristics like age, sex, ethnicity and residence. Therefore, the net survival rate provides a useful measurement for comparison of survival estimates amongst different populations by peeling off the potential varying effects of death risk from other causes.

The CONCORD programme, a global surveillance programme of cancer survival including data from population-based cancer registries from 71 countries and territories, published separate age-standardised 5-year net survival estimates for colon and rectal cancer in 2018 (Allemani *et al*, 2018). Based on these estimates, survival rates of colon and rectal cancer vary widely across the word. From 2010 to 2014, the highest survival rates were observed in Australia (colon cancer: 70.7%, rectal cancer: 71.0%). However, in countries such as South Africa, these estimates were as low as 12.3% for colon cancer and 9.1% for rectal cancer (Allemani *et al*, 2018).

This substantial variation may reflect to a certain extent differences in CRC screening programmes, treatment strategies, as well as surveillance across these countries. For instance, loss to follow-up rates greater than 15% were more common in African countries than in Europe and North America, making the survival estimates less reliable (Allemani *et al*, 2018). Other than that, distinct distributions of CRC patient characteristics across these areas may also be an essential attribute to this variation of survival estimates.

Tumour stage is currently the most commonly known determinant of patients' survival. Stratified survival estimates of individuals of different stage at diagnosis are provided by some large cancer registries or surveillance programmes. The Surveillance, Epidemiology, and End Results Programme (SEER), operated by the National Cancer Institute of the USA, is the largest cancer surveillance programme in the world (NCI, 2018). According to the latest SEER estimates, the 5-year (2009-2015) net survival rate of CRC patients in the USA is 64.4%. In lieu of AJCC stage, the latest SEER statistics reported the stage-specific survival estimates by grouping CRC into localised (stage I, IIA and IIB cancer), regional (stage IIC and stage III) and distant (stage IV) tumours. As shown in **Figure 1-11**, the 5-year relative survival drops drastically as the tumour stage progresses (from 89.9% for localised to 14.2% for distant CRC) (NCI, 2018).

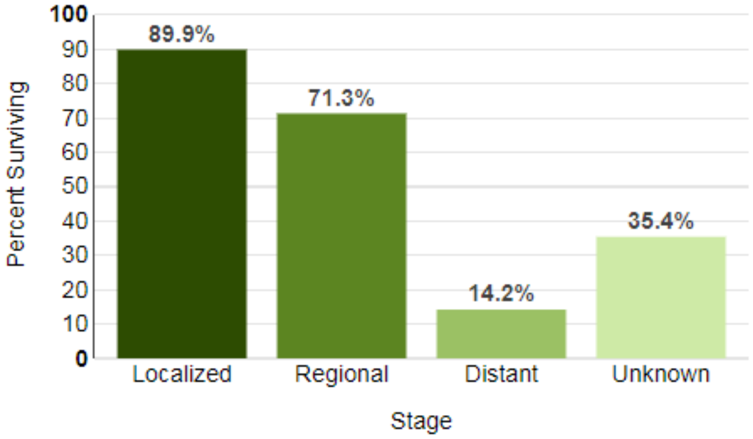


Figure 1-11 5-year relative survival estimates (2009-2015) by stage from the SEER programme of the USA. Reproduced from the NCI public data with permission.

In the UK, the 5-year net survival (2002-2006) of CRC patients (England and Wales) is approximately 60% based on the statistics from CRUK (URL1-6). In Scotland, the 5-year (2007-2011) net survival rate is 59.8% based on the latest estimates provided by the Scottish Cancer Registry (Scottish Cancer Registry, URL1-2). CRUK provides stage-specific relative survival estimates separately by gender. **Figure 1-12** demonstrates a clear trend of decreased survival rates with more advanced stage. Females show slightly favourable survival in each stage compared with males, although none of these differences are statistically significant (Bowel Cancer Statistics, CRUK, URL1-3).

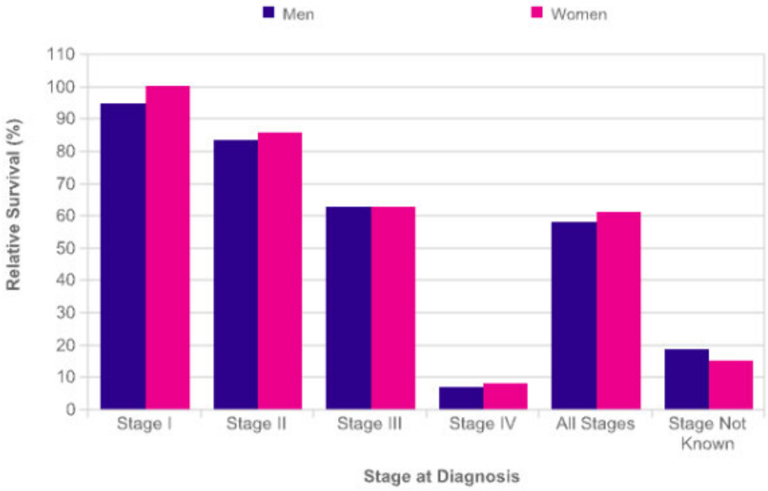
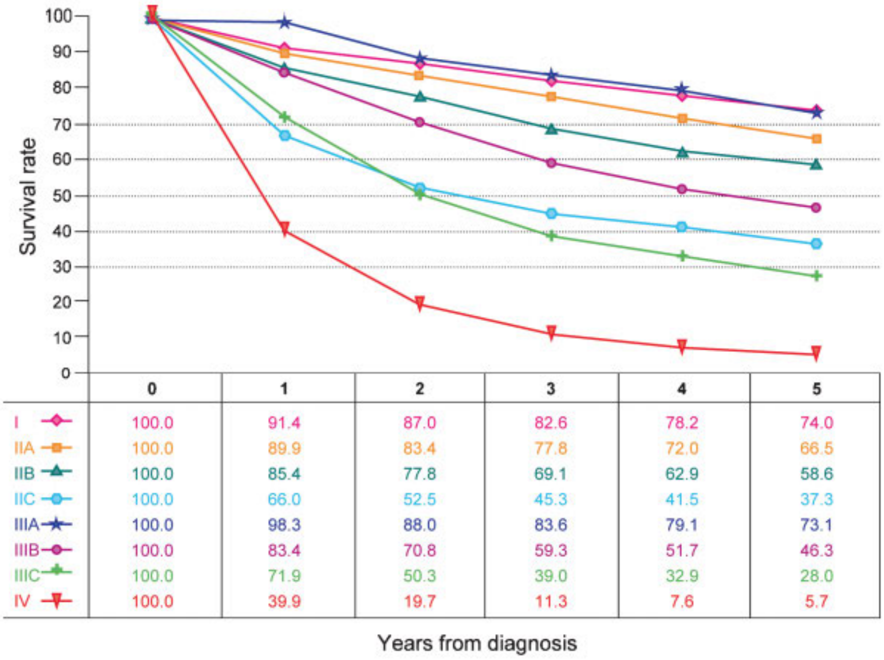


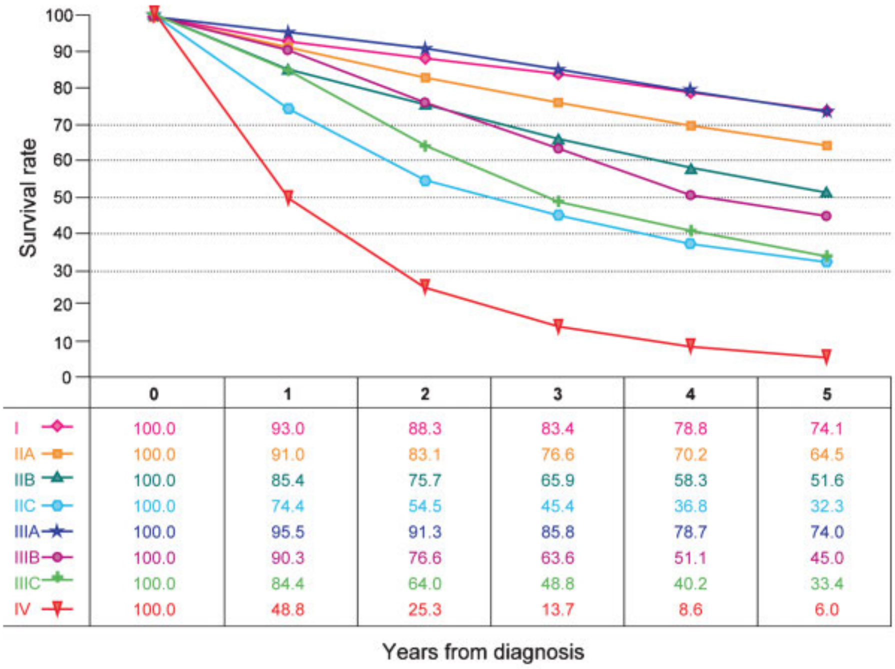
Figure 1-12 5-year relative survival rates (2002 to 2006) by stage for colorectal cancer patients in the UK. Reproduced with permission from the graph created by Cancer Research UK.

As described in section 1.3, the AJCC staging system is derived primarily on anatomic basis rather than empirical data regarding CRC prognosis. Therefore, the wide variation with respect to survival probabilities cannot be fully explained by cancer stage. There is still large heterogeneity in terms of individualised survival even within each specific stage, where the AJCC staging system (sub-stages such as IIA and IIB) may not be able to reliably further distinguish patients of varied prognosis. To take the SEER data as an example, the absolute survival estimates for each AJCC stage are plotted in **Figure 1-13** for colon and **Figure 1-14** for rectal cancer cases. As shown by the graphs, the 5-year overall survival rates for stage II and III CRC can vary from 28% to as high as 74%. Even within stage II, the differential of survival rates between IIC and IIA can be approximately 40%. Graphically, survival curves of stage II/III are heavily intertwined, indicating imperfect performance of the stratification by AJCC sub-stages.



	0	1	2	3	4	5
I ◇	100.0	91.4	87.0	82.6	78.2	74.0
IIA ■	100.0	89.9	83.4	77.8	72.0	66.5
IIB ▲	100.0	85.4	77.8	69.1	62.9	58.6
IIC ●	100.0	66.0	52.5	45.3	41.5	37.3
IIIA ★	100.0	98.3	88.0	83.6	79.1	73.1
IIIB ●	100.0	83.4	70.8	59.3	51.7	46.3
IIIC +	100.0	71.9	50.3	39.0	32.9	28.0
IV ▼	100.0	39.9	19.7	11.3	7.6	5.7

Figure 1-13 Absolute survival rates of 28,491 colon cancer patients stratified by stage from the SEER programme (1973-2005). The original source of this graph is the AJCC Cancer Staging Manual, seventh edition (2010)(Edge *et al*, 2010). Reproduced with written permission.



	0	1	2	3	4	5
I	100.0	93.0	88.3	83.4	78.8	74.1
IIA	100.0	91.0	83.1	76.6	70.2	64.5
IIB	100.0	85.4	75.7	65.9	58.3	51.6
IIC	100.0	74.4	54.5	45.4	36.8	32.3
IIIA	100.0	95.5	91.3	85.8	78.7	74.0
IIIB	100.0	90.3	76.6	63.6	51.1	45.0
IIIC	100.0	84.4	64.0	48.8	40.2	33.4
IV	100.0	48.8	25.3	13.7	8.6	6.0

Figure 1-14 Absolute survival rates of 9,860 rectal cancer patients stratified by stage from the SEER programme (1973-2005). The original source of this graph is the AJCC Cancer Staging Manual, seventh edition (2010)(Edge *et al*, 2010). Reproduced with written permission.

Beyond these aforementioned metrics in which deaths are used as the primary event of outcome, other alternative metrics, including disease-free survival (DFS), recurrence (relapse)-free survival (RFS), and progression-free survival (PFS), are also commonly adopted especially in the context of clinical trials. Whilst the OS or CSS is often deemed as the ‘gold standard’ endpoint of survival outcomes given that they are immune from measurement bias, the median survival time for CRC patients (all stages combined) can be as long as roughly 40 months (Laohavinij *et al*, 2010), hence increasing potential risk for cohort attrition as well as cost for long-time follow-up. These alternative endpoints can be accurately measured in clinical trials with more intense surveillance in a shorter observation time-window. Since events like CRC recurrence or progression usually occur prior to death, trials with such alternative endpoints as the primary outcome often require a smaller sample size to reach a pre-specified statistical power. In addition, there has been evidence supporting these endpoints as good surrogates with satisfied approximation to the OS (Oba *et al*, 2013). Definitions and common usage of some frequently used endpoints are summarised in the following **Table 1-4**.

Table 1-4 Summary of common endpoints of survival outcomes in clinical research.

Endpoint	Definition	Common clinical settings
Overall survival	The length of time from initial diagnosis (or primary treatment) to death from any cause.	Population statistics or clinical trials
CRC-specific survival	The length of time from initial diagnosis (or primary treatment) to death from CRC related causes. Deaths from other causes are often considered as censored.	Population statistics or clinical trials
Disease-free survival (Recurrence/Relapse-free survival)	The length of time from primary treatment (usually curative surgery) to disease recurrence.	Trials for adjuvant (postoperative) therapy
Progression-free survival	The length of time from primary treatment (non-curative) to disease progression.	Trials for metastatic (non-resectable) CRC

Of note, all the metrics discussed above are probability estimates measuring the proportion of survivors (free from pre-specified events) within a given time period. The other approach to evaluate survival is to estimate the time length from initial diagnosis of CRC (or primary treatment) to the endpoint of interest. In this case, an order statistic—the median of all the observed survival time, is often used as an estimate. The median survival time provides an intuitive estimate on how long an average individual can potentially survive.

1.4.3 Prognostic factors

As discussed in the previous section 1.3, diagnostic features of CRC, such as TNM measures, tumour site and grade, are important indicators for CRC prognosis.

However, even these factors combined cannot accurately predict survival outcomes of CRC patients. The survival estimates from **Figure 1-13** and **Figure 1-14** reveal substantial residual variation of prognosis for CRC-affected individuals, which merits incorporating more prognostic factors. Moreover, such factors, if modifiable, can lead to novel interventions to prolong survival time. For unmodifiable predictors like genetic variations, accurate prediction can inform decisions about more enhanced surveillance and more intensive treatment.

In concordance with the topic of this thesis, genetic factors, with a focus on germline genetic factors, will be introduced in this section. Thus far, most of the widely known clinical guidelines as well as organisation, except for the Canadian Cancer Society (Canadian Cancer Society, URL1-9), have not officially listed and recommended prognostic factors of CRC patients. Therefore, factors associated with CRC prognosis will be selected and introduced in this section mainly based on the recommendations from the Canadian Cancer Society.

Germline genetic variations

Empirical evidence has demonstrated an effect of inheritable genetic background on survival outcomes of CRC patients. A Swedish population-based study first revealed a familial concordance of cancer-specific survival by investigating over 4,800 pairs of parents and children diagnosed with the same type of primary cancer (Lindstrom *et al*, 2007). For CRC, in particular, children with a parent who died within 10 years of CRC diagnosis tended to have significantly poorer CRC-specific survival compared with those whose parent lived longer than 10 years since diagnosis (Lindstrom *et al*, 2007). Notably, this concordance was not identified among parents and children diagnosed with different types of primary cancers, indicating that this effect is most likely attributed to shared genetic basis rather than common environmental factors or similar behaviours. Current evidence suggests that the impact of germline variations on survival outcomes of CRC patients could possibly be mediated by their effects on tumour progression and response to treatment.

Germline genetics and cancer progression—As indicated previously, cancer progression and metastasis are the most lethal aspects of tumour behaviour that largely determines the survival outcomes of CRC patients. In the late 1990s, animal breeding studies developed the metastatic tumour mouse model by transferring a specific oncogene—the polyoma middle-T antigen transgene (Lifsted *et al*, 1998), and found that mice of different genetic origins exhibit varying tendencies of the primary mammary tumour to disseminate to the lung (Lifsted *et al*, 1998). This observed variation in metastatic potential is most likely attributed to the germline genome, considering the same initial oncogene was introduced in all the mice under study. In the following study, Hunter and colleagues successfully mapped this varied metastatic potential of primary mammary tumours in the mouse model to the genetic locus of Mtes1 on the chromosome 19 of the mice (Hunter *et al*, 2001). Another source of evidence comes from gene expression data. Ramaswamy et al. identified a distinct gene expression signature of 17 genes that is significantly associated with the metastatic potential of multiple types of human solid tumour including CRC (Ramaswamy *et al*, 2003). The pattern of the differential expression of these genes between tumours of high and low metastatic tendency was subsequently replicated in the mouse model by Hunter et al. with a fixed driver oncogene (Hunter *et al*, 2003), which further underscored the role of germline genetics in cancer metastasis. Despite shared genetic basis for metastasis of different types of cancers, currently there is a dearth of such evidence from animal models specifically developed to study CRC metastasis.

Impact on treatment response—Drug metabolism entails complex networks inside the human body engaging a very large number of biological molecules including transporters, enzymes and receptors. Genetic variations can modulate key steps of drug metabolism, and as a consequence cause varied responses to drugs. In terms of CRC, major drugs involved in chemotherapy include cytotoxins such as fluoropyrimidines (5-fluorouracil, capecitabine, S1 and tegafur), irinotecan and oxaliplatin. Targeted agents such as cetuximab and bevacizumab are also commonly used. Taking 5-fluorouracil as an example, it is a fluoropyrimidine analogue that can inhibit thymidylate synthase (TS)—an enzyme encoded by the *TYMS* gene catalysing the conversion of deoxyuridine monophosphate (dUMP) to deoxythymidine monophosphate (dTMP). Inhibiting TS supresses the formation of thymidine which is an indispensable part of DNA replication. Therefore, 5-fluorouracil can efficiently restrain proliferation of cancer cells. A germline variation in the number of tandem repeats (repetitions of single or multiple nucleotides adjacent to each other) in the promoter region of the *TYMS* gene has been identified to be associated with the *TYMS* gene expression, leading to varied responses to the 5-fluorouracil (Iacopetta *et al*, 2001). A patient's response to the treatment is measured by the extent to which the tumour has shrunk after chemotherapy. The inhibitory effect of 5-fluorouacil also applies to normal cells, which incurs toxic effects, such as myelosuppression and hand-foot syndrome, on CRC patients. The 5-fluorouracil is catabolised by an enzyme—the dihydropyrimidine dehydrogenase (DPD)—encoded by the *DYPD* gene. Evidence suggests that the DPD function is regulated by germline variations within the *DYPD* gene; individuals carrying the *DYPD* risk variant show deficient DPD function, and therefore are prone to higher risk of toxic effects (Amstutz *et al*, 2011). Other genes, such as the *MTHFR* and *ABCB1* gene, harbouring variations linked with treatment response of CRC patients to various drugs have been summarised by previous systematic reviews (Ab Mutalib *et al*, 2017). In addition to chemotherapy, emerging evidence also supports an important role of germline genetics in regulating the patients' sensitivity to radiotherapy (Kerns *et al*, 2014). Albeit proven efficacy of these drugs in improving clinical outcomes, robust associations between germline variations affecting the treatment response and CRC survival have not been established. This could be due to restricted data availability regarding treatment in large cohorts.

As previously mentioned, hereditary CRC cases such as HNPCC exhibit strong genetic predisposition driven by high-penetrance germline mutations. These cases only account for 4-6% of all CRC cases and the mutation profiles vary remarkably among patients, rendering each single pathogenic mutation too rare to be investigated individually in prognostic studies. Hence, previous efforts tended to aggregate these cases together and compared them to sporadic cases to explore overall effects of germline alterations as a whole on survival outcomes. For example, HNPCC patients had once been linked with better survival than sporadic cases (Sankila *et al*, 1996) before subsequent evidence found no significant difference in survival after adjusting for a variety of clinic-pathological factors that were largely discrepant between the two types of CRCs (Bertario *et al*, 1999). Results of a large cohort from our group also found lack of association between germline mutations in DNA mismatch repair genes (*MLH1, MSH2 and MSH6*) and survival outcomes (Barnetson *et al*, 2006). Current evidence suggests limited value of adding these rare germline mutations in predicting survival outcomes of CRC patients.

With respect to common germline genetic variants with lower penetrance, thus far there has been no published meta-analysis of GWASs on survival outcomes of CRC. According to the NHGRI-EBI GWAS Catalogue (URL1-10)—an online portal with all published GWASs indexed, three individual GWASs have been conducted investing CRC survival outcomes including the OS, CSS (Phipps *et al*, 2016; Xu *et al*, 2015) and time to distant metastasis (Penney *et al*, 2019); two of them were based on CRC patients from the same cohort (Penney *et al*, 2019; Xu *et al*, 2015). The largest GWAS by Phipps et al. identified no genetic variants significantly associated with either the OS or CSS among 3,494 CRC patients (Phipps *et al*, 2016). Although the other smaller cohort with 431 CRC patients reported several potential signals associated with time to metastasis, these findings were prone to false positivity or overestimated associations due to the limited sample size and lack of replication. These common germline variants reported by previous GWASs will be presented and discussed at length in Chapter 5. There have also been a number of published candidate genetic association studies focusing on specific sets of genetic variants. However, these variants were mostly investigated by single small studies. Therefore, hitherto accumulated evidence is insufficient to conduct a field synopsis and meta-analysis to summarise and further appraise potential prognostic roles of these variants.

In summary, the overall effect of germline genetic background on survival outcomes of CRC patients has been well supported by previous evidence. However, this effect is yet to be further dissected in order to identify specific loci that can potentially be employed as predictors to better inform patients' prognosis.

Somatic alterations and molecular subtypes

There has been extensive discussion of the prognostic role of somatic events acquired during CRC development. The Canadian Cancer Society recommends three somatic alterations that can potentially be used to predict survival: MSI, *KRAS* and *BRAF* mutations (Canadian Cancer Society, URL1-9).

As mentioned in section 1.2.4, deficiency of mismatch repair (MMR) genes can result in high level of MSI (instability of more than 30% of microsatellite loci) which is one of the genetic signatures enriched mainly in localised CRCs instead of metastatic ones. There has been evidence from meta-analyses supporting an association between

MSI-high tumours and improved survival compared with MSI-low or MSS tumours (Guastadisegni *et al*, 2010). However, this association is less prominent in metastatic CRCs possibly due to the low frequency of MSI-high tumours and the presence of a rich set of other somatic mutations observed in metastatic CRCs (Venderbosch *et al*, 2014).

A large meta-analysis found that *KRAS* mutation carriers manifest worse survival outcomes in comparison to wild type CRC patients (Petrelli *et al*, 2015). Given the fact that *KRAS* mutations are actually a group of mutations that occur at different positions of the gene, prognostic effects vary among these mutations. For example, Andreyev et al. looked into the prognostic effects of a spectrum of *KRAS* mutations and found that only one mutation in codon 12 was independently associated with CRC survival (Andreyev *et al*, 2001).

The *BRAF* mutations are mostly observed in codon 600 (also known as mutation V600E). Evidence from meta-analysis revealed a significant detrimental effect of the V600E mutation on overall survival of CRC patients (Ardekani *et al*, 2012). This effect, however, disappears in MSI-high CRCs, indicating possible interactions between this mutation and the MMR deficiency (Taieb *et al*, 2017).

There has been growing interest in devising molecular classification systems to better inform prognosis in addition to the current TNM staging system. These systems adopt somatic events including mutation profiles, molecular pathways and other characteristics discussed in section 1.1.2. Listed in **Table 1-5** is currently the most widely accepted classification system, known as the consensus molecular subtypes (CMS) (Guinney *et al*, 2015). The CMS harmonised six previous systems using clustering algorithms based on the relatedness of genetic and molecular characteristics.

Table 1-5 Summary of consensus molecular subtypes of colorectal cancer

Classification	**Genetic features**	**Median survival months* (95% CI)**
CMS1(Immune type)	High MSI and CIMP, high BRAF mutations, immune infilatration and activation	11.7(10.9-18.0)
CMS2(Canonical type)	High SCNA, Wnt and MYC activation	42.0(39.3-54.4)
CMS3(Metabolic type)	Mixed MSI status, low SCNA and CMIP, high KRAS mutations, metabolic deregulation	26.0(20.9-36.0)
CMS4(Mesenchymal type)	High SCNA, stromal infiltration, TGF-beta activation, angiogenesis	30.8(24.4-43.5)

*Median survival estimates were based on (Lenz *et al*, 2019). CMS, consensus molecular subtypes; SCNA, somatic copy number alterations; TGF, transforming growth factor. MSI, microsatellite instability; CMIP, CpG island methylator phenotype; CI, confidence interval.

As shown in the table above, CMS1 tumours are characterised by high burden of MSI, CIMP and *BRAF* mutations and highly expressed genes involved in immune function (Guinney *et al*, 2015). CMS2 CRCs can be distinguished by activation of the Wnt and MYC pathway, whereas CMS3 CRCs mainly entail impaired metabolic pathways (Guinney *et al*, 2015). As for the CMS4 subtype, markers of lymphocytes as well as monocytes are expressed in these CRCs, which points to the epithelial-to-mesenchymal transition of the tumour cells (De Sousa *et al*, 2013). The prognostic significance of the CMS was validated by an external cohort of 581 CRC patients (Lenz *et al*, 2019). The study found that the CMS as a risk factor was significantly associated with the OS and PFS of CRC patients. However, as indicated by the median survival estimates in **Table 1-5**, the predictive performance of the CMS is still suboptimal particularly for CMS3 and CMS4 patients who manifested similar survival outcomes (Lenz *et al*, 2019).

Other genetic factors

With biotechnology advancing rapidly, more and more novel genetic markers have been detected and linked with prognosis of CRC. A review by Compton published in the UpToDate® systematically summarised factors associated with CRC prognosis (Compton, 2019) (URL1-11). According to Compton's review, other genetic factors that have not been introduced in section 1.4.3 mainly include: gene expression profiles of oncogenes and tumour suppressor genes (Munro *et al*, 2005) (Ellis *et al*, 2000), epigenetic changes such as methylation levels (Jiang *et al*, 2014), microRNA levels (Gao *et al*, 2018), and circulating tumour cells and tumour DNA (Garlan *et al*, 2017) (Chou *et al*, 2018). A full list of these markers can be found in the review (Compton, 2019). Although a wide range of genetic factors have been identified, these markers were mostly reported by small studies with inconsistent findings or by studies without replication (Compton, 2019). Therefore, their exact effects on CRC prognosis still remain to be investigated.

Non-genetic factors

Pathological factors

Based on the recommendations of Canadian Cancer Society (URL1-9), main pathological factors associated with CRC prognosis include: tumour stage, histological type and grade, lymphovascular invasion, and surgical margin.

As described in section 1.3, the TNM stage that encompasses information on how far the tumour has advanced is instrumental in predicting survival outcomes of CRC patients. The numeric stage (I to IV), however, causes inevitable information loss from the original T, N and M measurements. In addition to accounting for the T, N and M stages separately, other modified measurements such as the lymph node ratio (the number of positive lymph nodes divided by the number of nodes examined) (Rausei *et al*, 2013), as well as other advanced data technologies such as machine learning (Hueman *et al*, 2019) have also been devised in an attempt to optimise prediction on CRC survival. However, the trade-off between stratification performance and attainability for clinicians ought to be balanced.

The histological type of CRC, for example the presence of signet ring cells, and tumour grade of differentiation harbour indications on biological behaviour related to tumour aggressiveness and therefore can serve as common prognostic factors of CRC. These effects have been well supported by large registry data (NCI, 2018).

Lymphovascular invasion, defined by identification of tumour cells within veins or lymphatic vessels, can also inform poor prognosis (Yuan *et al*, 2017). Of note, lymphovascular invasion is listed in recommendations from both the American Society of Clinical Oncology (ASCO) and the European Society of Medical Oncology (ESMO) as one of the risk factors for postoperative recurrence of stage II patients (Benson *et al*, 2004; Schmoll *et al*, 2012), which justifies more intensive therapy for these patients. Additional recommended pathological factors that increase the risk of postoperative recurrence include pathological T4 stage (pT4), grade 3 or 4 tumour and perineural invasion (Benson *et al*, 2004; Schmoll *et al*, 2012). Similar to lymphovascular invasion, perineural invasion has been associated with poor prognosis of CRC patients as identified by a previous meta-analysis (Knijn *et al*, 2016).

Surgery with curative intent is instrumental in the management of CRC patients. A surgical margin refers to the unaffected normal tissue surrounding the removed tumour. Inferior survival outcomes of CRC have been associated with narrow margins, which indicates that the resected tumour is close to the surgical edge (Bernstein *et al*, 2009).

Clinical factors

Bowel obstruction with or without perforation is generally reported to be associated with adverse survival outcomes of CRC (Chen & Sheen-Chen, 2000). Current guidelines also consider obstruction and perforation as risk factors that may warrant adjuvant chemotherapy for stage II CRC patients (Schmoll *et al*, 2012). It is worth noting that CRC cases with bowel obstruction and perforation tend to exhibit more invasive histopathological features that can potentially confound the observed association between these clinical manifestations and survival outcomes (Ghazi *et al*, 2013).

The other prognostic factor listed by the Canadian Cancer Society is the circulating level of carcinoembryonic antigen (CEA). It is commonly used in clinical practice to monitor potential risk of postoperative recurrence of CRC. This association is independent from tumour stage (Thirunavukarasu *et al*, 2011). However, there is still a heated dispute over the best cut-off value for CEA level.

Lifestyle factors

Lifestyle after CRC diagnosis may also influence patients' survival outcomes, although these factors were not officially recommended by either the Canadian Cancer Society (URL1-9) or the systematic review by Compton (Compton, 2019). A meta-analysis of prospective cohort studies found that increased physical activity after CRC diagnosis was associated with improved survival outcomes (Schmid & Leitzmann, 2014). With respect to dietary factors after diagnosis, Meer et al. systematically reviewed published literature, and found no consistent findings regarding associations between dietary factors, such as overall dietary patterns, meat intake and alcohol intake, and survival outcomes of CRC patients (van Meer *et al*, 2013). Future investigations are still needed to provide more solid evidence regarding the effects of lifestyle factors on CRC survival.

Treatment

It is well established that the treatment CRC patients receive can greatly influence their survival outcomes. Prognostic effects of certain treatment are estimated mostly via prospective studies, particularly randomised controlled trials (RCT), and thus are generally deemed credible evidence. In reality, however, the treatment variable itself is often not included in multivariable models, because of the challenges in obtaining detailed information in a large cohort regarding treatment regimens - such as number of completed cycles and presence of discontinuation of chemo- or radiotherapy. The flow of CRC management varies across different regions. In view of the scope of this thesis, only over-arching principles of patient management will be introduced here, according to the guidelines from the National Institute for Health and Care Excellence (NICE) (URL1-12).

In general, curative surgery is the centrepiece for all localised and regional CRCs. For stage I and part of stage II patients without any risk factors mentioned previously, no additional chemotherapy is needed following the surgery. Adjuvant chemotherapy is usually given to stage II (with one or more aforementioned risk factors) and stage III patients after surgery in order to reduce recurrence risk. Therapeutic agents for chemotherapy mainly include 5-fluorouracil/capecitabine alone or combined with oxaliplatin as the first-line regimen. Irinotecan is usually used as the second-line agent when the first-line does not work or incurs severe side effects.

Regarding metastatic CRCs, a multidisciplinary team (MDT) effort is recommended. A curative surgery may be considered if both the primary and metastatic tumours are deemed potentially resectable. Chemotherapy, radiotherapy or targeted therapy (monoclonal antibody) may be applied to alleviate symptoms and shrink tumours. Notably, for rectal cancer that has grown into nearby tissues, preoperative (also known as neoadjuvant) short course radiotherapy or chemoradiotherapy may be considered to increase the chance of a tumour-free surgical margin.

Other non-genetic factors

Based on the systematic review by Compton (Compton, 2019), a wide range of factors have been reported to be associated with CRC prognosis, yet have not been officially recommended. These factors mainly include:

●tumour budding (TB, defined as a single or cluster of tumour cells at the invasive margin of the cancer) (Rogers *et al*, 2016)

●host immune function (e.g. density of tumour-infiltrating lymphocytes in the tumour tissue)(Pages *et al*, 2005)

●intra-tumour microvessel density (Des Guetz *et al*, 2006)

●tumour location(Petrelli *et al*, 2017)

Further validation is needed before these factors can be officially recommended to guide CRC management.

1.5 Summary

Colorectal cancer is the second most common and third most lethal cancer around the world. As of 2018, approximately 4.8 million people were living with CRC. In terms of incidence, an estimate of 1.84 million newly-diagnosed CRC cases was observed worldwide in 2018, and there will be an estimate of 2.2 million newly-diagnosed CRC patients per year by 2030. Both inheritable genetic components and environmental factors can contribute to CRC susceptibility. Somatic genetic alterations in oncogenes (e.g. *KRAS*) and tumour suppressor genes (e.g. *p53*) are sequentially acquired during the development of CRC. Based on the incidence pattern, CRC can be categorised into hereditary (4-6%), familial (30-40%) and sporadic (50-60%) cases. Hereditary CRC s are mostly driven by rare germline mutations with high penetrance, such as mutations in the *APC* and *MUTYH* genes, whereas common germline variations with low penetrance have been found to be associated with increased risk of familial and sporadic CRC cases. In addition to genetic markers, environmental factors such as red or processed meat intake, cigarette smoking and alcohol consumption have also been linked to increased CRC susceptibility.

There was a total of 0.88 million CRC-related deaths worldwide in 2018, and this number is projected to be 1.1 million per year by 2030. The 5-year overall survival rate of CRC patients is approximately 60%. Colorectal cancer patients can be grouped into different stages (0 to IV) based on the extent of tumour invasion and dissemination at diagnosis. Patients diagnosed with more advanced stages show significantly worse survival outcomes. Although tumour stage serves as a main indicator of CRC prognosis, a substantial amount of variation in terms of survival outcomes has been observed for patients diagnosed with the same stage. For example, the 5-year overall survival rates for stage II CRC patients can vary from 36% to 67%. Other known factors that may influence patients' survival mainly include pathological features such as tumour grade and lymphovascular invasion, clinical manifestations such as bowel obstruction and treatment strategies for patients. For genetic markers, somatic mutations involved in CRC carcinogenesis (e.g. *KRAS* and *BRAF* mutations) could have subsequent effects on survival outcomes of CRC. Previous evidence found familiar concordance of CRC patients in terms of their survival outcomes, indicating possible prognostic effects of germline genetic variations. However, there is lack of solid evidence supporting associations between specific germline genetic loci and survival outcomes of CRC.

Chapter 2 Study aims and objectives

As introduced in Chapter 1, there has been both epidemiological and biological evidence supporting the overall effect of germline genetic background on survival outcomes of CRC. However, the genetic architecture of CRC survival still remains poorly understood, due to the paucity of hitherto identified genetic risk loci robustly associated with CRC survival. Although previous GWASs with small sample sizes reported possible associations between a few genetic variants and CRC survival, these associations have not been replicated and are subject to risk of false positive findings. Moreover, it is also unknown whether incorporating these germline genetic variants into established prediction models of other prognostic factors can further improve predicting survival outcomes of CRC. Given this large knowledge gap, the overarching aim of this thesis is to examine associations between germline common variations and survival outcomes of CRC patients after diagnosis, to identify genetic variants that may be predictive, and to develop prediction models on CRC survival by integrating genetic predictors. The overall study design of the thesis is presented in **Figure 2-1**.

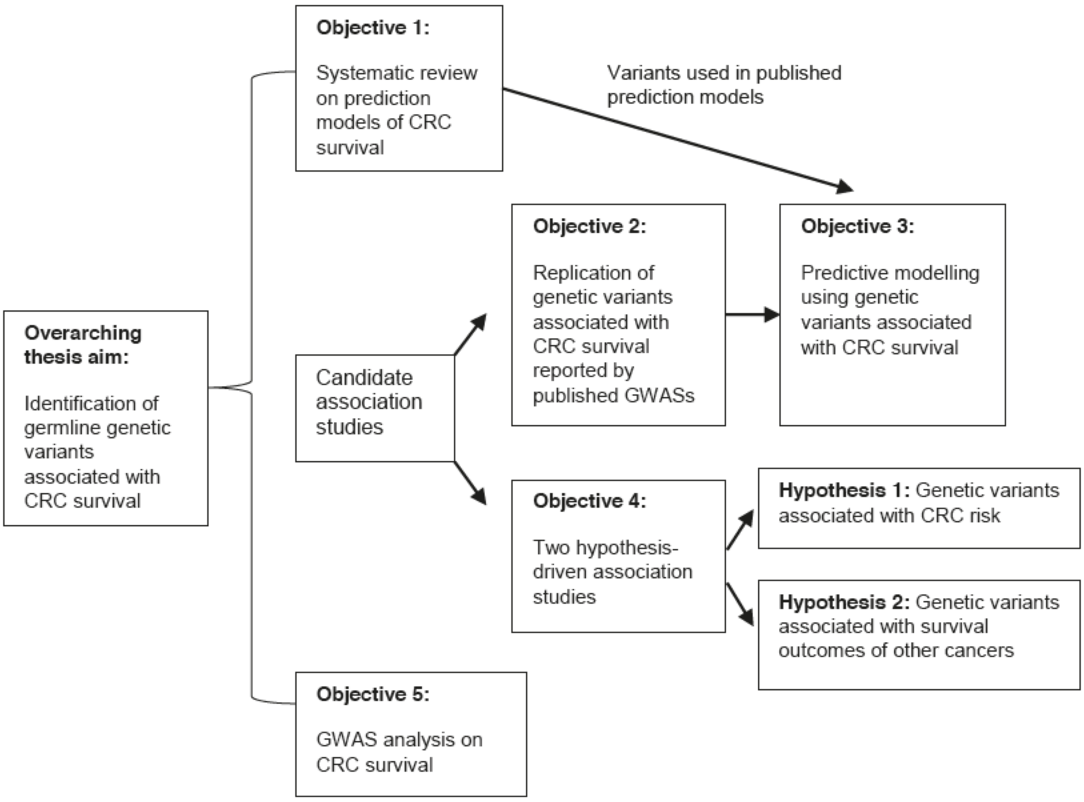

Overarching thesis aim:
Identification of germline genetic variants associated with CRC survival
Objective 1:
Systematic review on prediction models of CRC survival
Variants used in published prediction models
Candidate association studies
Objective 2:
Replication of genetic variants associated with CRC survival reported by published GWASs
Objective 3:
Predictive modelling using genetic variants associated with CRC survival
Objective 4:
Two hypothesis-driven association studies
Hypothesis 1: Genetic variants associated with CRC risk
Hypothesis 2: Genetic variants associated with survival outcomes of other cancers
Objective 5:
GWAS analysis on CRC survival

Figure 2-1 Overall study design of the thesis

To be specific, this thesis contains the following objectives:

Objective 1: To conduct a systematic literature review on published prediction models regarding CRC prognosis (presented in Chapter 3)

As introduced in Chapter 1, a wide spectrum of genetic and non-genetic risk factors has been reported to be associated with CRC survival. However, it remains unclear which of these factors have been adopted in multi-variable models to predict survival outcomes of CRC and to what extent these factors can add extra predictive value on the basis of established predictors such as tumour stage. Therefore, the first objective of the thesis is to systematically search the published literature, and identify all prediction models developed to forecast survival outcomes of CRC patients. I will summarise both genetic and non-genetic prognostic factors that have been adopted in these prediction models. Also, the performance of these models will be evaluated using established tools and methodologies. Had any germline genetic variants been adopted in published prediction models, they would be further validated in the subsequent analyses of the thesis.

Objective 2: To validate previously reported associations between germline genetic variants and CRC survival (presented in Chapters 4-6)

Overall, there has been a paucity of solid evidence supporting associations between specific genetic risk loci and CRC survival. Currently, study designs for investigating population-based genetic associations include: candidate genetic association studies and genome-wide association studies (GWAS). Ioannidis et al. reported that ~99% of disease risk loci identified in candidate genetic association studies were unable to be validated due to flawed study design and various sources of biases (Ioannidis *et al*, 2011); thus critical appraisal on the credibility of these claimed associations is crucial before efforts of replicating them in large cohorts. Given the lack of a field synopsis that could systematically summarise and appraise previous candidate genetic association studies, the main focus of this objective will be on validating genetic variants reported by GWASs that investigated survival outcomes of CRC. Findings of GWASs are also subject to risk of false positive findings due to millions of markers being tested in one study. Although procedures like multiple-testing correction can to some extent reduce this risk, it is still essential to validate GWAS-identified signals in large well-conducted cohorts. As discussed in Chapter 1, only three GWASs have been published and only a few variants, identified in a small cohort of 431 CRC patients, survived the stringent criteria of genome-wide significance ($p<5x10^{-8}$). This further underpins the need of replication. For the second objective of the thesis, germline genetic variants that have been linked to CRC survival by GWASs will be retrieved by searching the GWAS catalogue (URL1-10). Associations between these variants and overall and CRC-specific survival will be validated in a Scottish cohort (details about the cohort will be presented in Chapter 4).

Objective 3: To develop a multi-variable prediction model combining previous GWAS-identified genetic variants and other non-genetic prognostic factors (presented in Chapters 4-6)

As opposed to the second objective where the focus is on the individual association between each variant and CRC survival, I aim to look into the predictive value of these previously identified variants as a group in predicting survival outcomes of CRC, combined with common non-genetic predictors such as AJCC stage. Efforts of identifying prognostic factors serve the ultimate goal of informing clinical decision-making towards improved clinical outcomes for patients at various levels of risk. This relies on accurate prediction on prognosis for each patient. Over the past decade, there has been an increasing interest in genomic prediction with the genetic underpinnings of disease outcomes progressively unravelled by large GWASs. In the context of CRC, incorporation of GWAS-identified genetic variants has led to improved prediction on CRC risk compared with models with phenotypic variables only, although the incremental margin of performance is moderate (McGeoch *et al*, 2019). However, similar predictions on prognostic outcomes have not been reported thus far. In this phase, genetic variants identified by previous GWASs on CRC survival will be used to develop a genetic predictor in CRC cases from the UK Biobank cohort (details in Chapter 4). Then the performance of this genetic predictor together with other non-genetic factors will be tested in the Scottish cohort.

Objective 4: To investigate associations between CRC survival outcomes and candidate genetic variants using a hypothesis-driven approach (presented in Chapters 4-6)

This objective is to explore the impact of two groups of genetic variants of interest on survival outcomes of CRC. Genetic variants used in this section will be identified from previously published GWAS meta-analyses and the GWAS catalogue (URL1-10) based on two prior hypotheses. Associations of these variants with overall and CRC-specific survival will be tested in the Scottish cohort. Any statistically significant associations will be further validated in CRC cases from the UK Biobank cohort.

Hypothesis 1: Genetic variants associated with CRC risk can subsequently affect tumour progression and metastasis, and therefore may be associated with CRC survival. As introduced in the background, genetic events of CRC tumorigenesis, such as the *KRAS* and *BRAF* mutations, have also been shown to be associated with CRC prognosis, indicating possible continuing effects of these pathogenic genetic alterations on CRC progression and metastasis. Although survival difference between CRC patients carrying inherited predisposition, for example HNPCC, and sporadic cases is yet inconclusive, there has been interest in exploring the impact of genetically determined CRC susceptibility on subsequent survival outcomes. Recent large

GWAS meta-analysis expanded the spectrum of known common germline variants associated with CRC risk, allowing for further investigation in the prognostic significance of these CRC-risk variants.

Hypothesis 2: In relation to prognosis, there is possibly shared genetic basis across different cancer types. Hence, genetic variants associated with survival outcomes of other cancers may also influence CRC survival. Early evidence of family-based large observational studies found that family history of a specific cancer also increased risk for other cancers, indicating common genetic basis of risk for multiple cancers (Amundadottir *et al*, 2004). This finding was further strengthened by later combined analysis of GWAS data on risk for multiple cancers. Jiang and colleagues quantified genetic correlation among six cancers (head and neck, breast, lung, ovarian and colorectal cancer), and revealed shared genetic components contributing to risk for these cancers (Jiang *et al*, 2019). Although genetic correlations among prognostic outcomes of multiple cancers have not been explored due to the dearth of accumulated data, there has been other evidence suggesting these shared genetic components. For instance, several regulators in key pathways, such as the Notch signalling pathway, can modulate the invasion-metastasis cascade of multiple cancers by governing critical processes like epithelial-mesenchymal transition and tumour angiogenesis(Hu *et al*, 2012). Beyond CRC, there have been GWASs investigating survival outcomes of a few types of other cancers; genetic variants identified in these studies will be tested as the second part of this objective.

Objective 5: To discover potential novel genetic variants associated with CRC survival by performing a genome-wide association study (presented in Chapters 4-6)

In general, the genetic architecture of CRC survival remains unclear. Under circumstances with limited prior knowledge, GWAS is a powerful approach for discoveries on disease related genetic risk loci, given large well-characterised cohorts of CRC patients being aggregated. In contrast to candidate gene approach, in a GWAS, the whole human genome is scanned, and thus novel genetic variants located in specific genomic regions involved in the prognosis of CRC can be found. The final objective of the thesis is to conduct a GWAS on CRC survival using the Scottish cohort. Any discoveries will be validated by a pooled analysis of CRC cases from the

UK Biobank cohort and three previously published CRC clinical trials (additional details in Chapter 4).

Chapter 3 Systematic literature review

3.1 Introduction

In Chapter 1, I presented a summary of background epidemiological evidence regarding both genetic and non-genetic factors that can potentially affect survival outcomes of CRC. These factors are also known as 'predictors' when used to estimate the probability of a future event of interest, for example death within five years after diagnosis for CRC patients. This estimation process leverages information from a spectrum of predictors, and is widely recognised as risk prediction which can be realised by developing statistical prediction models. In clinical practice, prediction models can assist clinicians in their decision-making based on the specific risk profile of candidate predictors for a given patient. Subsequent to Chapter 1 where prognostic factors are summarised, this chapter aims to further look at the real-world clinical utility of CRC prognostic factors in the setting of risk prediction by systematically reviewing published prediction models on CRC survival and conducting evidence synthesis to quantitatively evaluate the predictive performance of these models.

This chapter presents a published paper in *Surgical Oncology* entitled '***Performance of prediction models on survival outcomes of colorectal cancer with surgical resection: A systematic review and meta-analysis***' (He *et al*, 2019a). As introduced in Chapter 1, clinical and pathological factors, such as AJCC stage and tumour grade, are most commonly accepted as prognostic predictors for CRC, and these factors rely on assessment of tumour specimens retrieved from surgical resections. Therefore, the systematic review focuses on published studies including CRC patients who underwent surgical resections. Based on current clinical guidelines (see Chapter 1 page 38), this resulted in inclusion of all prediction models for stage I to III CRC patients and part of models for stage IV patients whose primary and metastatic tumours are considered resectable. Following the structure of the publication, this chapter is composed of introduction, methods, results, discussion and conclusion. As the main investigator of this published work, I conducted the literature search, study selection, data extraction and statistical analysis. A parallel review was conducted independently to screen for eligible studies by a medical student (*Ong Y.* from the Western General Hospital, Edinburgh). Another investigator—*Wang Z.* (West

China Hospital, Chengdu) helped check the extracted data, and *Li X.* (Usher Institute, Edinburgh) independently appraised the quality of a randomly selected subset of included studies. I drafted the manuscript and revised it based on the comments from the editor and peer reviewers of the journal. Other authors (*Theodoratou E, Farrington S, Campbell H, Dunlop M, Timofeeva M, Din F, and Brown E*) critically reviewed and edited the manuscript.

3.2 Systematic review and meta-analysis on prediction models of survival outcomes of colorectal cancer

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

Authors: Yazhou He[1,2], Yuhan Ong[3], Xue Li[1], Farhat VN. Din[2,4], Ewan Brown[5], Maria Timofeeva[2,4], Ziqiang Wang[6], Susan M. Farrington[2,4], Harry Campbell[1], Malcolm G. Dunlop[2,4], Evropi Theodoratou[1,4]*

1. Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
2. 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
3. Western General Hospital, Edinburgh, UK
4. Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
5. Edinburgh Cancer Centre NHS Lothian, Edinburgh, UK.
6. Department of Gastrointestinal Surgery, West China Hospital, Sichuan University, Chengdu 610041, P.R. China

***Corresponding author:** *Dr. Evropi Theodoratou*, Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, United Kingdom; Tel: (+44) 131-650-6194, Fax: (+44) 131-650-6194；

E-mail: E.Theodoratou@ed.ac.uk

3.2.1 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. There is a pressing need for more external validation studies so as to evaluate the performance of other abundant published prediction models that have not been adequately validated. Future research should also focus on investigating the real-word impact and cost-effectiveness of existing prediction models for CRC prognosis in clinical practice.

3.2.2 Introduction

Colorectal cancer (CRC) is responsible for 8.5% of deaths attributed to cancer worldwide(Ferlay *et al*, 2015). The overall 5-year survival of CRC varies from 50% to 81% even within stage II CRC patients. This within-stage variation can be explained to some extent by a wide range of other established prognostic factors such as carcinoembryonic antigen (CEA)(Spindler *et al*, 2017). 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(Vickers, 2011), 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(Bowles *et al*, 2013), genetic(Goossens-Beumer *et al*, 2015a) and clinic-pathological(Bowles *et al*, 2013) 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.

3.2.3 Methods

Literature search and study selection

This study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement(Moher *et al*, 2009). 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(Ingui & Rogers, 2001) and previous publications(Brush *et* *al*, 2011) (detailed search syntax presented in **Table 3-1**). The reference list of each eligible article was also cross-checked.

Table 3-1 Search strategy for the systematic review

Database: MEDLINE (limit to English language and human studies; not review or editorial or letter or comment)
exp Colorectal Neoplasms/ or ((rectal or rectum or colonic or colon or colorectal) adj2 (cancer* or carcinoma* or neoplas* or tumor* or tumour* or malignan* or adenocarcinoma*)).mp.
AND
exp Prognosis/ or prognos*.mp. or Survival Analysis/ or Survival/ or surviv*.mp. or mortality.mp. or exp Mortality/ or metastas*.mp. or Neoplasm Metastasis/ or recurren*.mp. or Neoplasm Recurrence, Local/
AND
exp Models, Statistical/ or predict*. ab,ti. or validat* .ab,ti. or Validation Studies/
Database: Embase (limit to English language and humans; not review or editorial or letter or comment, excluding MEDLINE journals)
exp colorectal cancer/ or rectum tumor/ or colon tumor/ or ((rectal or rectum or colonic or colon or colorectal) adj2 (cancer* or carcinoma* or neoplas* or tumor* or tumour* or malignan* or adenocarcinoma*)).mp.
AND
prognos*.mp. or cancer prognosis/or exp survival or surviv*.mp. or metastasis/ or metastas* .mp. or cancer recurrence/ or recurren*.mp. or mortality/ or cancer mortality/
AND
exp prediction/or predict*.ab,ti. or exp statistical model/or validat*.ab,ti.

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.

Data extraction and critical appraisal

One reviewer (YH) extracted all relevant data following the guidelines of conducting systematic reviews of prediction model studies(Debray *et al*, 2017). 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)(Moons *et al*, 2014). Based on this checklist, the risk of bias for each model was assessed following the criteria described in previous publications (Lamain–de Ruiter *et al*, 2017; Smit *et al*, 2015) 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 summarised in **Appendix Table S1**. One reviewer (YH) appraised all included studies. A second blinded reviewer (XL) evaluated a 25% random sample of all studies and cross-checked for any discrepancies.

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 (Debray *et al*, 2017). We rescaled the C statistic by applying a logit transformation (Debray *et al*, 2017). 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(Debray *et al*, 2017). The C statistic was considered statistically significant if the 95% CI excluded 0.5(Hosmer & Lemeshow, 2000). 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 {IntHout, 2014 #850}. We also calculated the 95% prediction interval (PI) integrating the heterogeneity for the summarised C statistic to indicate a possible range where a C statistic of a future validation study may be located (Higgins *et al*, 2009; IntHout *et al*, 2016). Due unavailable data, we were unable to perform quantitative synthesis for other metrics evaluating model performance.

3.2.4 Results

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(Takakura *et al*, 2011). In total, 83 articles comprising 83 original model development studies and 52 separate external validation studies (**Appendix Table S2-S3**) were included in this systematic review. The detailed study selection is summarised in **Figure 3-1**.

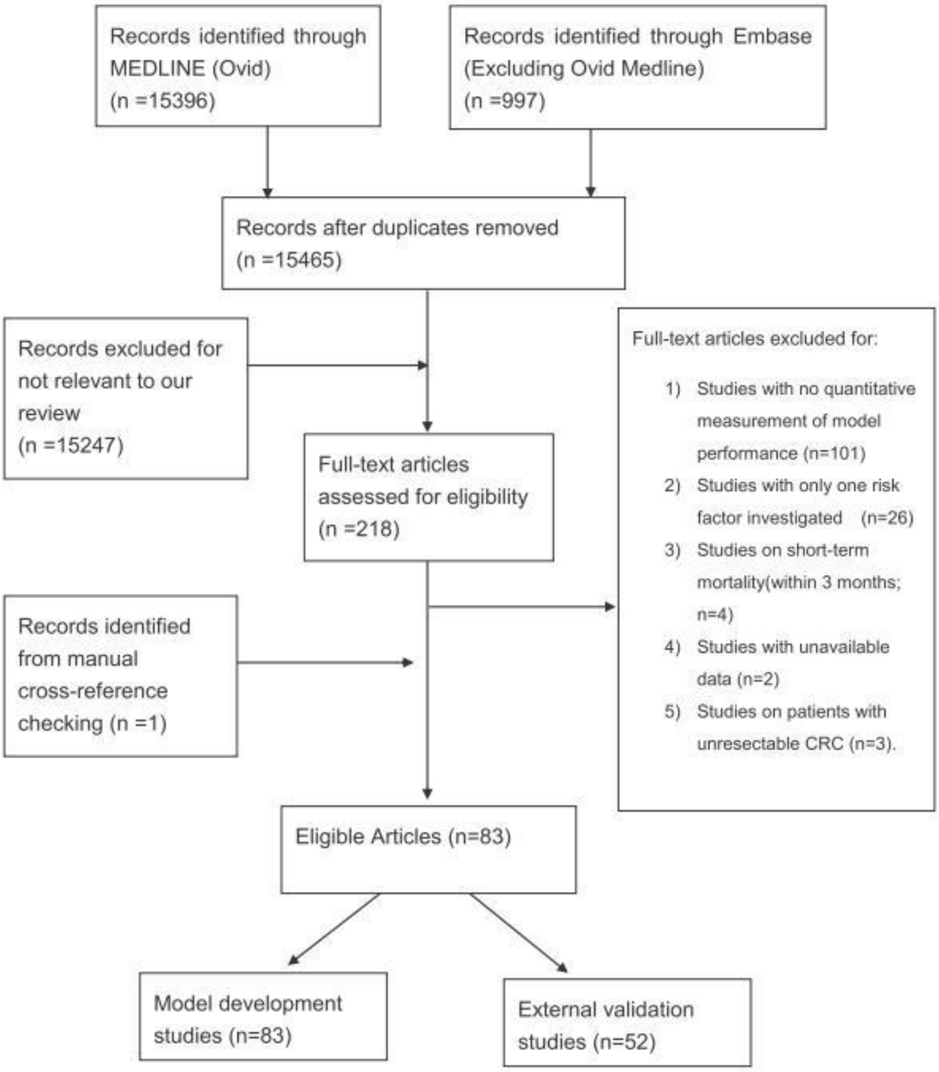

Records identified through MEDLINE (Ovid) (n =15396)
Records identified through Embase (Excluding Ovid Medline) (n =997)
Records after duplicates removed (n =15465)
Records excluded for not relevant to our review (n =15247)
Full-text articles assessed for eligibility (n =218)
Full-text articles excluded for:
1) Studies with no quantitative measurement of model performance (n=101)
2) Studies with only one risk factor investigated (n=26)
3) Studies on short-term mortality(within 3 months; n=4)
4) Studies with unavailable data (n=2)
5) Studies on patients with unresectable CRC (n=3).
Records identified from manual cross-reference checking (n =1)
Eligible Articles (n=83)
Model development studies (n=83)
External validation studies (n=52)

Figure 3-1 Flow diagram of study selection

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 to 18). Age was the commonest predictor (*N*=56, 67%). Other common predictors included CEA (*N*=26, 31%), tumour 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(Peng *et al*, 2018) and tree-based models(Arostegui *et al*, 2018). 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 **Appendix Table S2**.

Among the 52 separate external validation studies (detailed characteristics in **Appendix Table S2**), 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 summarised in **Table 3-2**.

Table 3-2 Summarised basic characteristics of included model development studies and external validations

Variables	Model Development(*N*=83)	External validation(*N*=52)
Participants (CRC patients)		
Cohort origin		
Europe	16(19%)	23(44%)
Asia	52(63%)	19(36%)
America	15(18%)	5(10%)

Variables	**Model Development(*N*=83)**	**External validation(*N*=52)**
Other	0	5(10%)
CRC Stage		
I-III	45(54%)	8(15%)
IV	20(24%)	44(85%)
Any	18(22%)	0
Tumour location		
Colon	15(18%)	3(6%)
Rectum	16(19%)	3(6%)
Any	52(63%)	46(88%)
Sample size		
<500	28(34%)	9(17%)
>=500	55(66%)	43(83%)
No. predictors		
<5	30(36%)	16(31%)
5-10	50(60%)	36(69%)
>10	3(4%)	0
Outcome		
Overall survival	47(57%)	24(46%)

Variables	**Model Development(*N*=83)**	**External validation(*N*=52)**
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%)
Other†	4(5%)	5(10%)
Model calibration		
Calibration plot	47(57%)	7(13%)
Hosmer-Lemeshow test	6(7%)	0
Internal validation		
Split sample	14(17%)	NA
Bootstrapping	13(16%)	NA
Cross validation	18(22%)	NA
Not reported	39(47%)	NA
Model presentation		
Nomogram	55(66%)	NA
Formula	21(25%)	NA
Other**	7(8%)	NA

*including D-statistic, sensitivity and specificity.
**including score rule and decision tree.
CRC, colorectal cancer; AUC, area under receiver's operating characteristic curve.

Critical appraisal

Risk of bias distribution of each domain for all included studies is summarised in **Figure 3-2**. Overall, only two models reported by one article were classified as low risk of bias for all domains (Rees *et al*, 2008a). 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 **Appendix Table S4** (model development studies) and **Appendix Table S5** (external validation studies).

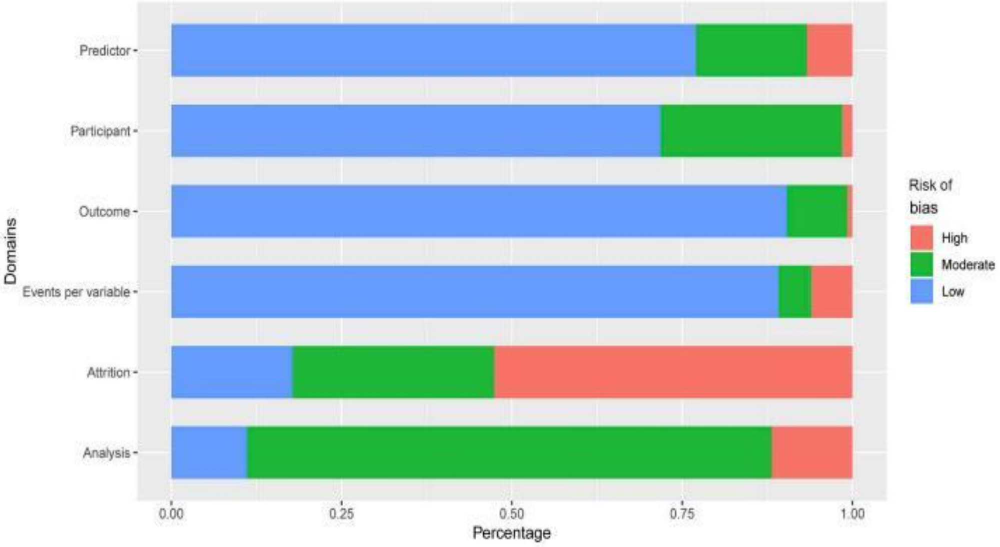


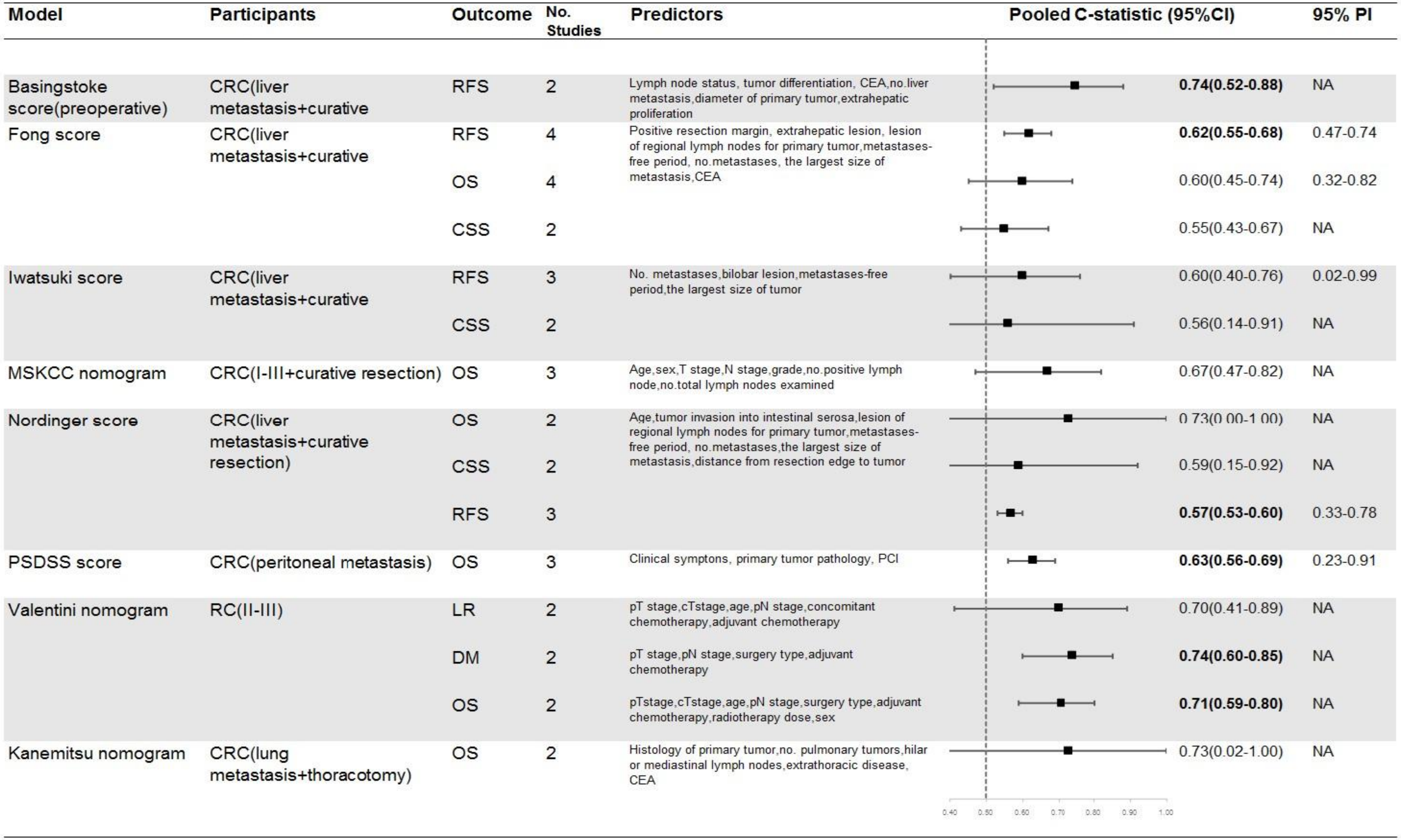
Figure 3-2 Risk of bias assessment for six predefined domains for each included study. For participant selection, studies were rated as ‘moderate’ risk of bias if participants were possibly selected in a non-consecutive manner as this allowed for potential selection bias. We categorized studies to be 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 **Appendix Table S1**.

Model performance

Of all studies, 126 (93%) reported a C statistic to assess the discriminative ability of the model. 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 including 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 Centre (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 **Figure 3-3**. 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 (**Figure 3-3**). 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.



Model	Participants	Outcome	No. Studies	Predictors	Pooled C-statistic (95%CI)	95% PI
Basingstoke score(preoperative)	CRC(liver metastasis+curative	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	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	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

Figure 3-3 Summarised C statistics of prediction models included in meta-analysis. Adapted from the original publication with permission

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 tumour, 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) and the Valentini nomogram (C statistic: 0.74, 95% CI: 0.60-0.85).

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.

3.2.5 Discussion

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 individualised 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 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 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) (Rahbari *et al*, 2010) and immune-scores(Mlecnik *et al*, 2018), 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(Van Cutsem *et al*, 2016), 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 to 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 (Diouf *et al*, 2014; Kawai *et al*, 2015). Novel metrics, such as the expected information for discrimination(McKeigue, 2018), 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.

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(Dettori, 2011). 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) (Moons *et al*, 2015), 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(van Walraven *et al*, 2004).

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(Harrell *et al*, 1996). 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 (32%) of these models adopted split-sample approach despite this method being less favored due to its inefficiency (Steyerberg *et al*, 2001). Future studies should consider more sophisticated internal validation methods such cross-validation and bootstrapping(Steyerberg *et al*, 2001). 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 (Collins *et al*, 2014). It is ideally performed by independent investigators to avoid over-interpretation(Collins *et al*, 2014), 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(Moons *et al*, 2009). Aside from that, cost-effectiveness should also be evaluated by health economic modelling, which is scarce in current CRC prognostic models(van Giessen *et al*, 2017). Finally, few studies have explored how prediction models can be integrated into the clinical workflow(Vickers, 2011), which will also have ramifications on their clinical utility.

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)(Arnold *et al*, 2017). 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 (Austin *et al*, 2017; Blanche *et al*, 2013). 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.

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. Most of these scores 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.

3.3 Summary

This chapter presents a published systematic literature review regarding prediction models on survival outcomes of CRC patients. I reviewed 83 original model development studies and 52 external validation studies. This review found that the majority of published prediction models had not been validated by external datasets, and their performance could potentially be overestimated in the original report. As for models that had been validated in at least two datasets, five models were identified by meta-analysis with significant discriminative ability to predict survival outcomes of CRC. No cost-effectiveness analyses or model impact studies were identified in this review, indicating that future efforts are still warranted before these prediction models can be applied in routine practice to guide more accurate patient management. In this review, I also summarised the predictors employed in published prediction models. However, none of the included prediction models adopted germline genetic variations as predictors to develop their models, indicating the paucity of solid evidence on associations between any germline variants and survival outcomes of CRC. The following chapters of the thesis will leverage multiple patient cohorts and explore potential effects of germline genetic variations on CRC survival.

Chapter 4 Materials and methods

4.1 Introduction

This chapter mainly describes materials and methods used in this thesis. It consists of two sections. In the first section, datasets including the Study of Colorectal Cancer in Scotland (SOCCS), CRC cases from UK Biobank and three previously published clinical trials will be introduced. The second section describes at length the study design and the main steps of data analysis for genetic association studies, prediction modelling and the GWAS.

4.2 Data sources

4.2.1 The Study of Colorectal Cancer in Scotland

The Study of Colorectal Cancer in Scotland (SOCCS) is a population-based case control study which has been actively recruiting CRC patients and matched healthy controls (age, sex and health board) from all areas of Scotland since February 1999. It should be noted that controls were only used for other research purposes such as investigating risk factors for CRC susceptibility. For analysis in this thesis, only CRC cases from the SOCCS were used as a case cohort to explore survival outcomes of patients. The main aim of the SOCCS study is to investigate genetic and environmental factors contributing to CRC risk and survival outcomes. This study was funded by CRUK, Medical Research Council (MRC) and Chief Scientist Office of the Scottish Executive (CSO).

Research Ethics approval

The SOCCS study was approved by the MultiCentre Research Ethics committee for Scotland (MREC; approval number MREC/ 01/0/0), 18 Local Research Ethics committees, 18 Caldicott guardians and 16 NHS Trust management committees (Theodoratou *et al*, 2008). Informed consents were signed by participants regarding their DNA samples and other relevant clinical and lifestyle data being stored and used by the research team based in the University of Edinburgh as well as collaborators from other research groups. Each participant was assigned a unique identification number and all data were entered into an anonymised Access database (except the genotype data which are also anonymised and stored separately).

Inclusion and exclusion criteria for CRC cases

The SOCCS study consists of two phases. Participants recruited from February 1999 to 2006 formed the first phase of the study. For this phase, all incidental CRC cases throughout Scotland were included in the study if they were:

1) histologically diagnosed with CRC

2) were 16 to 84 years of age

3) permanently resident in Scotland

The diagnosis of CRC was confirmed histologically by referring to patients' pathological reports.

Patients were excluded if they were:

1) recurrent colorectal cancer cases;

2) unable to provide informed consent, for example if they were too ill or had mental health disorders.

The second phase of the SOCCS study consists of incident CRC patients recruited from 2007 onwards at the Western General Hospital, Edinburgh. In this phase, UK residents of 16 years or older with a diagnosis of CRC at any time were included. Other inclusion and exclusion criteria remained the same.

Genotype data

Blood samples were collected at recruitment and leucocyte DNA was extracted following standard protocols. DNA samples were genotyped using the Illumina®

HumanHap300, HumanHap240S and OmniExpressExome BeadChip 8v1 arrays. Detailed specifications of these genotyping arrays can be found at the Illumina® Support Centre (URL4-1). As of December 2018, a total of 6,366 CRC cases and 14,692 controls were genotyped in the SOCCS study.

Quality control

Dr Maria Timofeeva, a statistical geneticist in the Colon Cancer Genetics Group, performed the quality control and imputation for the genotype data of the SOCCS cohort. Detailed technical information can be found in previous publications (He *et al*, 2018b; Law *et al*, 2019). In particular, the quality control was conducted in concordance with the protocol proposed by Anderson and colleagues (Anderson *et al*, 2010). Participants were excluded based on the following criteria, and detailed numbers of individuals excluded in each step are presented in Chapter 5.

1) High missing rate of genotyping (>5%). This is an indicator of possible low quality of the DNA sample, which can potentially impair the genotyping accuracy.

2) Abnormal heterozygosity (>3 standard deviations from the mean). Extreme heterozygosity rates indicate possible contamination of the DNA sample or inbreeding.

3) Discordant sex classification.

4) Individuals recruited twice or who had first-degree relatedness to other included participants.

5) Evidence of non-white European ancestry. This was evaluated by principal component analysis (PCA) in conjunction with samples of European ancestry from the 1000 Genome Project (URL4-2). PCA is a method that evaluates the resemblance among samples under study by clustering them based on a range of their features. For example, samples in the SOCCS study were clustered according to a few hundred thousand of genetic variants arrayed. This method incorporates common genetic variants (minor allele frequency>1%) throughout the germline genome into a restricted number of independent principal components (PCs) that are numbered by a descending order of importance--that is, PC1 accounts for the largest amount of genetic difference across all included individuals. Principal components 1 and 2 are plotted in **Figure 4-1** and **Figure 4-2** to show the clusters of individuals in the SOCCS and the 1000 Genome Project.

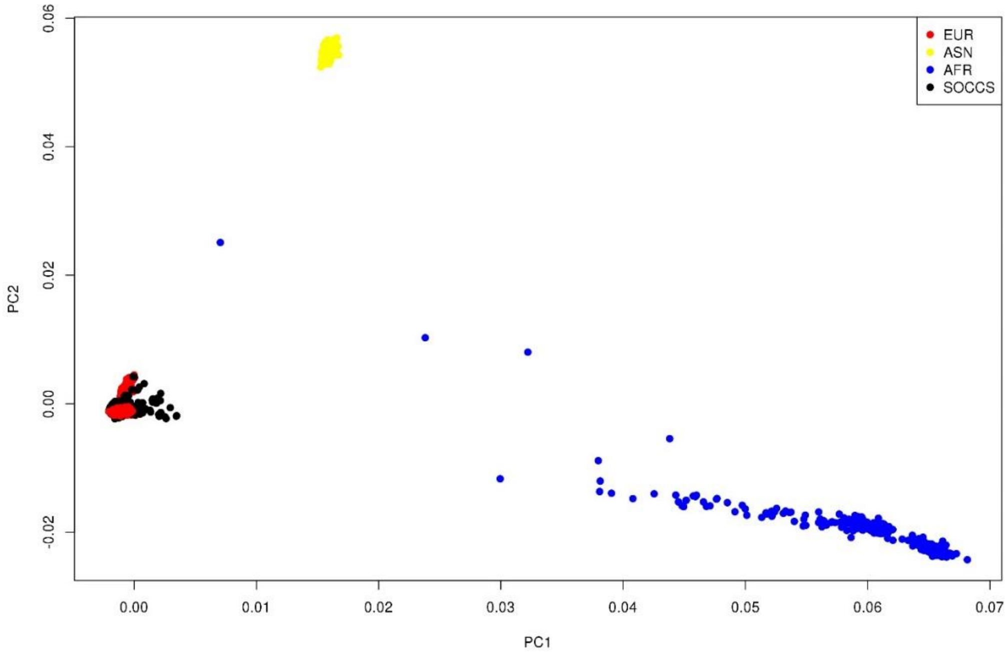


Figure 4-1 Principal components clustering of individuals in the SOCCS study with European (EUR), Asian (ASN) and African (AFR) populations from the 1000 Genome Project. Created by Dr. Maria Timofeeva (Institute of Genetics and Molecular Medicine, Edinburgh) and used with permission.

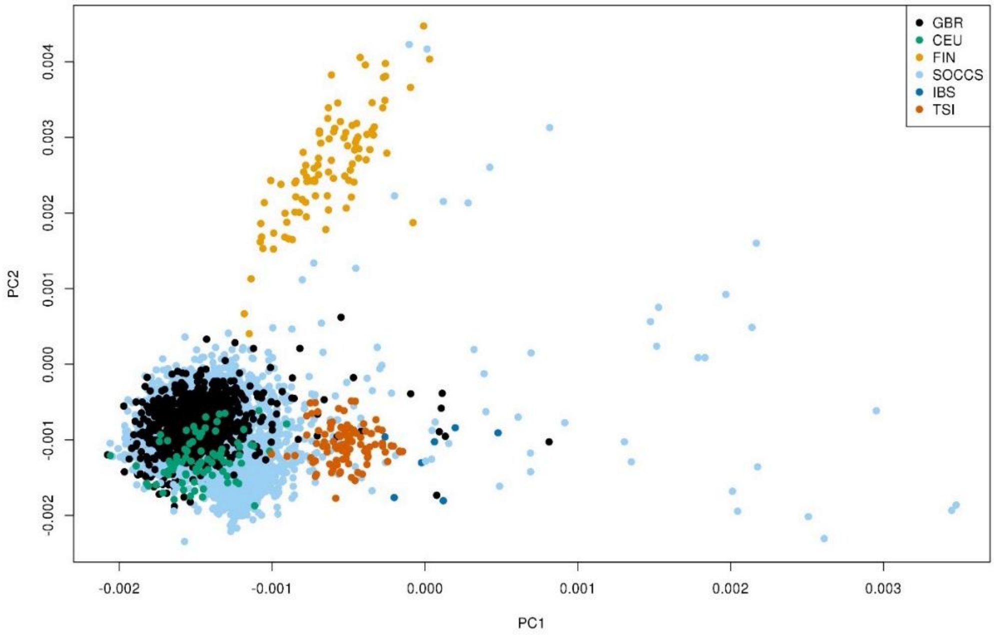


Figure 4-2 Principal components clustering of individuals in the SOCCS study with European populations of the Northern and Western European Ancestry (CEU), Toscani in Italia (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR) and Iberian in Spain (IBS) from the 1000 Genome Project. Created by Dr. Maria Timofeeva (Institute of Genetics and Molecular Medicine, Edinburgh) and used with permission.

As shown in these two figures, all SOCCS samples clustered tightly with the five European populations sequenced as part of the 1000 Genome project (**Figure 4-2)**—Northern and Western European Ancestry (CEU), Toscani in Italia (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR), and Iberian Population in Spain (IBS), yet no clustering with Asian and African populations was detected (**Figure 4-1**). The closely clustered samples from the SOCCS cohort indicate that population structure is unified and therefore relatively unlikely to affect observed associations between genetic markers and phenotypes.

Imputation

Genotyping arrays are normally designed to detect a limited number of genetic variants (usually 250,000 to 500,000 variants) instead of the entire set of common genetic variants throughout the human genome. Especially for large-scale genetic association studies with thousands of samples included, it is often too costly to sequence the whole genome for each individual. Nonetheless, genetic variants genotyped by arrays, also known as tag variants, are still informative given the ubiquitous linkage disequilibrium (LD) throughout the genome. Linkage disequilibrium reflects the fact that alleles at nearby genetic loci tend to be inherited together, leading to non-random associations among these alleles. In an attempt to leverage these associations among correlated genetic loci, a range of imputation methods have been proposed to infer undetected genotypes that can be used to map genetic risk loci for specific disease outcomes. Imputation increases the density of genetic variants and hence adds extra statistical power to discover potential genetic loci. The imputation process for the SOCCS study is briefly introduced here.

After an extensive procedure of quality control, the genotype data of the SOCCS study were phased using the SHAPEIT (v2.r837) software (Delaneau *et al*, 2011). Phasing, also known as haplotype estimation, refers to the statistical method used to infer the haplotypes—combinations of genotypes that are inherited together— based on arrayed genotypes. For instance, the genotypes of two loci are detected as ‘AB’ and ‘AB’ (A and B refer to different alleles); then the haplotypes in theory can be ‘AB/AB’ or ‘AA/BB’. A variety of probabilities models have been developed to estimate the most likely haplotypes—that is, to reconstruct the chromosome origin for each allele. The SHAPEIT software adopts a linear complexity method to estimate haplotypes and additional details regarding the estimation process are presented elsewhere

(Delaneau *et al*, 2011). There has been evidence showing that inferring haplotypes prior to imputing genotypes can improve accuracy and efficiency of imputation, and therefore has been recommended by the operating manual of the IMPUTE2 software (URL4-3).

The inferred haplotypes were then passed to the IMPUTE2 software for the imputation of untyped genotypes (URL4-3). The IMPUTE2 module operates based on the imputation method proposed by Howie et al (Howie *et al*, 2009). **Figure 4-3** depicts the conceptual framework of imputation. The basic idea is to compare the genotyped sample with a reference panel from the same ancestry that has been more densely genotyped or sequenced, so as to infer unknown genotypes of the sample leveraging the genomic LD structure.

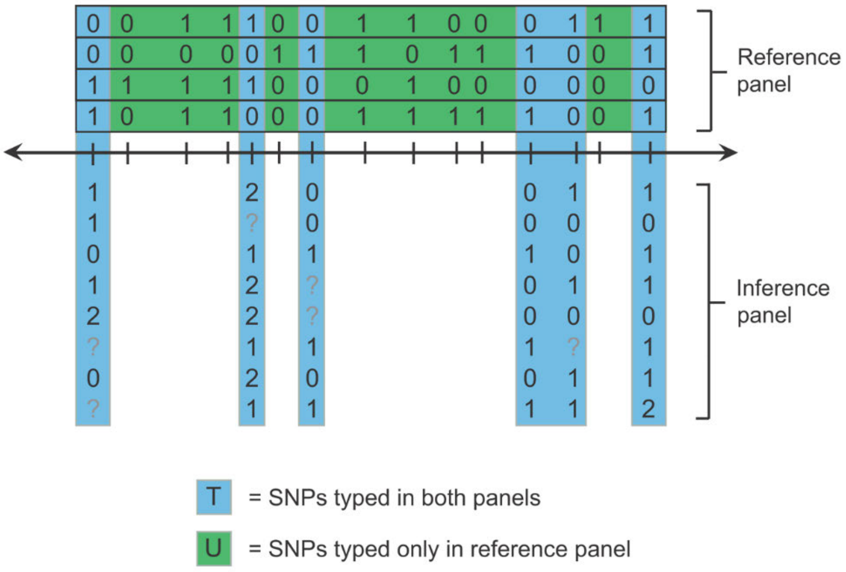

Reference panel
Inference panel
T = SNPs typed in both panels
U = SNPs typed only in reference panel

Figure 4-3 Conceptual framework for genotype imputation. Horizontal boxes denote haplotypes of the reference panel. Reproduced from (Howie *et al*, 2009) with permission. The original source is covered by a Creative Commons Attribution License.

The reference panels we used for imputation were the UK10k release (ALSPAC and TWINSUK studies, April 2014 release), and the 1000 Genome V3 (December 2013 release). All variants were coded and mapped to their chromosome positions based on the Genome Reference Consortium Human Build 37 (GRCh37)—a digital database of DNA sequences which was derived from a number of volunteers recruited from the USA and released in February 2009 (URL4-4). We excluded variants that have discrepancies in strand (5’ end to 3’ end or 3’ end to 5’ end) and chromosome position information across the two reference panels (<1% of variants). The two reference panels were merged in IMPUTE2, and the imputation was conducted in 5Mbp chunks. Monomorphic variants [only one allele present in the dataset, rare variants with minor allele counts <20 and poorly imputed variants] were excluded from the analysis. Poor quality of imputation for a specific variant was defined as an information score less than 0.80. The information score estimates the ratio between the effective sample size accounting for imputation uncertainty and the real sample size (Marchini & Howie, 2010). For example, an information score of 0.5 indicates that the imputed genotype data are equal to half of the sample size of accurately genotyped data in terms of estimating a pre-defined genetic effect. A total of 8,328,632 autosomal genetic variants were included in the analysis.

Phenotype data

Covariates

As introduced in Chapter 1 (page 36), prognostic factors that are officially listed by the Canadian Cancer Society include: AJCC stage, surgical margins, lympho-vascular invasion, CEA levels, bowel obstruction or perforation, tumour grade, histological type, microsatellite instability (MSI0, *KRAS* gene mutation, *BRAF* gene mutation). Among these factors, the AJCC and tumour grade were available in our SOCCS datasets and therefore were extracted with other basic demographic variables. These variables were used as covariates and are summarised in the following **Table 4-1**.

Table 4-1 Summary of non-genetic variables extracted from the database of the SOCCS study

Variables	Descriptions
Age at diagnosis	Collected from clinical records
Sex	Biological sex ascertained from genetic data
AJCC stage	Derived from the TNM stage based on the 6th edition AJCC staging manual
Tumour grade	Collected from pathological reports

Survival outcomes

Overall and CRC-specific survival were both employed as outcomes for analysis. The death status and date for each CRC patient in the SOCCS study were retrieved by linking the database to the Scottish Cancer Registry on January 1st 2018. Notably, it takes up to six months for the Scottish Cancer Registry to enter all updates of death records received from the National Death Registry. Hence, deaths occurred between July 2017 to January 2018 could possibly be missed in our dataset. In order to minimise this bias, we set the endpoint of follow-up at July 1st 2017—that is, deaths between July 1st 2017 and January 1st 2018 were treated as censored. The survival time for each patient was defined as the time span from the date of definitive treatment (starting date of surgery or chemo/radiotherapy for patients without surgery) and the date of death or July 1st 2017 whichever happened first. I assigned the cause of death based on information provided by death certificates. Rules for assigning the cause of death for each case are summarised below:

1) All deaths with the primary diagnosis of CRC reported in death certificates were noted as CRC-related deaths (N=1,212).

2) If the death certificate mentioned the presence of ‘metastasis’ or ‘carcinomatosis’, then the cause of death was presumed to be CRC (N=92).

3) If the case was reported with unknown primary tumour, then the death was assigned as CRC unless the certificate clearly stated otherwise (N=37).

4) If the death certificate stated that either the primary or metastatic tumour site had been unresectable, then the cause of death was noted as CRC (N=8).

5) If the certificate stated that the cause of death was due to a visceral or intra-abdominal complication which could be directly related to CRC or the treatment of the disease, the cause of death was attributed to CRC (N=9).

6) If no clear diagnosis was reported by the death certificate, then the death was deemed as non-CRC related.

A randomly selected sample of 200 death records were evaluated independently by a colorectal surgeon, Dr Peter Vaughan-Shaw, and the concordance of results were checked. Only two cases (1%) were identified with discordant results upon which agreement was reached after discussion.

4.2.2 UK Biobank

UK Biobank is a large, population-based prospective cohort study designed to investigate genetic and environmental determinants of a wide spectrum of human complex traits and disorders. From 2006 to 2010, the study recruited over 500,000 participants aged 40-69 years throughout the UK. More details regarding participant recruitment for the UK Biobank can be found in the online study protocol (URL4-5).

Research Ethics approval

The UK Biobank study was approved by North West Multicentre Research Ethics Committee (Reference 11/NW/0382). Each participant signed an electronic consent form at one of the 22 assessment centres across the UK. The dataset of CRC cases used in this thesis was based on a study proposal approved by the UK Biobank (Project No. 7441: 'Investigation of genetic, environmental and gene x environmental interaction in colorectal cancer risk and survival', Principal investigator: Professor Evropi Theodoratou). The Project 7441 adopted a case-control design including all the prevalent and incident CRC cases along with healthy controls (cases vs controls:

1:4) matched by age, gender, date of blood sampling, ethnicity and region of residence. Similar to the SOCCS study, only CRC cases from the Project 7441 were used as a case cohort to explore survival outcomes of patients.

Identification of CRC cases

UK Biobank retrieves diagnoses of medical conditions by linking participants to the Hospital Episode Statistics (HES) and the cancer registry. Diagnoses are coded using both the 9th and 10th version of the International Classification of Diseases (ICD). Classification codes corresponding to CRC are summarised in the **Table 4-2**. The project (7441) includes individuals from the UK Biobank cohort with diagnoses of ICD9 codes 153.0-154.1 and ICD10 codes C18.0-C20. It is worth mentioning that UK Biobank also documented self-reported conditions at recruitment. Self-reported CRC cases were excluded from the analysis unless they were confirmed by ICD9 or ICD10 codes.

Table 4-2 ICD9 and ICD10 codes for colorectal cancer

Sites	ICD9	ICD10
Colon		
Hepatic flexure	153.0	C18.3
Transverse colon	153.1	C18.4
Descending colon	153.2	C18.6
Sigmoid colon	153.3	C18.7
Cecum	153.4	C18.0
Appendix	153.5	C18.1
Ascending colon	153.6	C18.2
Splenic flexure	153.7	C18.5

Sites	ICD9	ICD10
Overlapping lesion of colon*	153.8	C18.8
Colon, unspecified sites	153.9	C18.9
Rectum		
Rectosigmoid junction	154.0	C19
Rectum (includes rectal ampulla)	154.1	C20

*Subcategory 153.8 and C18.8 denotes malignancies that overlap two or more continuous sites or the point of origin cannot be determined.

Genotype data

Peripheral blood samples for each participant in UK Biobank were taken at recruitment and the DNA was extracted following a standard protocol described elsewhere (URL4-6). Genotyping was conducted using the Affymetrix UK BiLEVE Axiom array for the initial 50,000 participants; a customised array—the Affymetrix UK Biobank Axiom®—was used to genotype the remaining 450,000 participants. These two arrays tagged over 95% of overlapped genetic variants. Additional details regarding the genotyping arrays can be found online (ULR 4.6).

Quality control

Two rounds of marker-based quality control were performed by the Affymetrix laboratory and the Wellcome Trust Centre for Human Genetics (WTCHG) respectively. In brief, the Affymetrix laboratory applied a cluster-based method which contained a range of metrics to identify variants that were genotyped with poor accuracy. More technical details can be found elsewhere (URL4-7). Based on the filtered variants from the initial round, the WTCHG designed a panel of six statistical tests to further detect poorly genotyped variants by examining the consistency across various experimental factors such as array difference and batch effect. Descriptions on the panel of statistical tests can be found in the previous publication (Bycroft *et al*, 2018). An aggregate of 805,426 genetic variants passed these two rounds of marker-based quality control.

The UK Biobank cohort consists of participants of diverse ancestral origins across the UK. The WTCHG conducted PCA to dissect the population structure of the whole cohort, although self-reported ethnic background was also documented in the study. The first four principal components were plotted in the **Figure 4-4** below with self-reported ethnic background marked by different colours and signs.

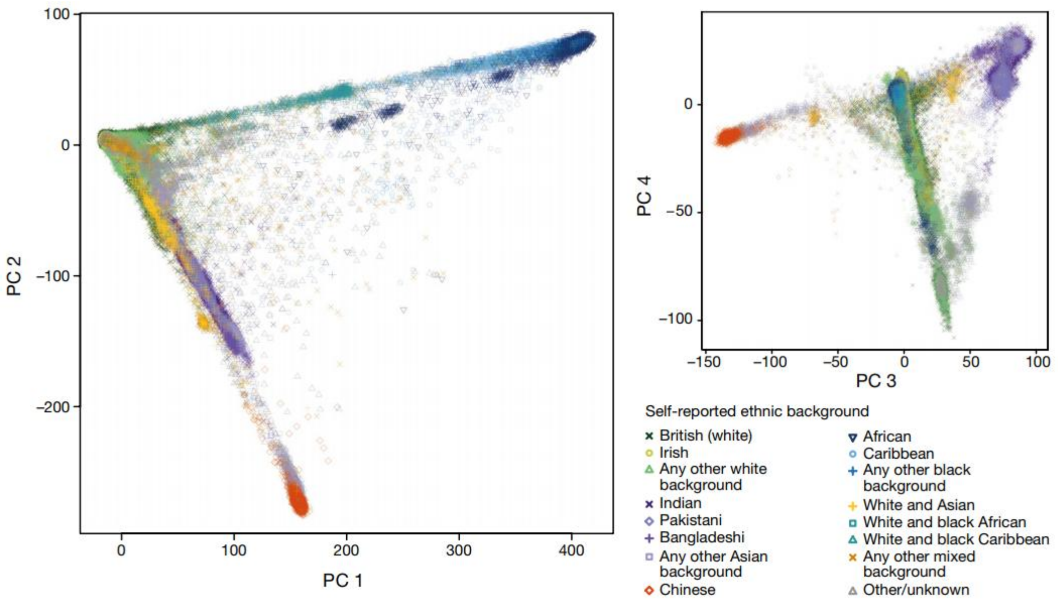


Figure 4-4 Principal components of genetic background of participants in UK Biobank. Reproduced with permission. The original source is covered by a Creative Commons Attribution License.

Given the dispersed distribution of genetic background of UK Biobank participants (**Figure 4-4**), further sample-based quality control procedures were conducted within the cases-control study (Project 7441). In particular, firstly participants with self-reported non-white origin were excluded; then individuals with evidence of non-European ancestry were identified and excluded by applying the same procedure as in the SOCCS study (described in section 4.2.1). Accordingly, individuals with discordant sex information, duplication or first-degree relatedness, high missing rate (>5%) and abnormal heterozygosity were also excluded.

Imputation

Genotype imputation for the UK Biobank cohort was conducted by the WTCHG. Similar to the imputation procedure of the SOCCS study, genotype data that passed quality control were first phased using the SHAPEIT (v3) software. With regard to genotype imputation, the WTCHG employed an updated version of the IMPUTE2 module—known as IMPUTE4 (URL4-8)—that applied the same statistical model but with improved computational efficiency. In line with the SOCCS study, genetic markers in UK Biobank were also coded based on the Genome Reference Consortium Human Reference 37 (GRCh37). Based on the imputed dataset, we excluded genetic variants with an information score less than 0.8. Eventually, a total of 9,067,367 autosomal genetic variants were included in the genotype dataset of the UK Biobank.

Phenotype data

Covariates

Extensive phenotyping of the UK Biobank cohort has been underway as of this writing. For instance, disease characteristics for patients diagnosed with certain disorders such as cancer have not been available for researchers. With regard to our study, clinic-pathological data including cancer stage and grade for CRC patients have not yet been released in November 2019. Other relevant variables available in UK Biobank are summarised in the **Table 4-3**. The variable 'Age when attended assessment centre' was used to differentiate incident and prevalent CRC cases. In particular, cases who were diagnosed with CRC within six months since attendance or before attendance were categorised as prevalent cases.

Table 4-3 Summary of non-genetic variables extracted from the UK Biobank cohort study

Variable	UK Biobank ID	Descriptions
Age at diagnosis	40008	Collected from the clinical records
Sex	31	Biological sex ascertained from genetic data
Age when attended assessment centre	21003	Documented at recruitment

Survival outcomes

UK Biobank retrieves death records based on the linkage to the National Death Registry; death records are updated quarterly. Since the records were directly synchronised with the National Death Registry, here we assumed no delay of data entry that could cause potential bias. The death status and date for our study (Project 7441) was last updated in February 2018. Therefore, the survival time for each individual was defined as the time span between the date of CRC diagnosis (UK Biobank variable ID: 40005, acquired from the Central Registry) to the date of death (UK Biobank variable ID: 40000) or February 1st 2018 for cases that were alive. As for the cause of death, UK Biobank assigned putative primary (UK Biobank variable ID:40001) and secondary causes of death (UK Biobank variable ID:40002) to each individual based on the death certificate. The causes of death were coded using ICD10 codes. In our study, death records were treated as CRC-related deaths in the analysis of CRC-specific survival if either primary or secondary causes of death contained ICD10 codes of CRC (**Table 4-2**) or any indications of conditions listed in the criteria applied to the SOCCS study.

4.2.3 Clinical trials datasets

I obtained summary-level data from three previously published clinical trials in order to validate potential discoveries from the genome-wide association study of associations between genetic variants and overall survival of CRC patients. The summary-level results were extracted and provided by Dr Claire Palles from the Institute of Cancer and Genomic Sciences, the University of Birmingham. More descriptions on the summary-level results will be presented in Chapter 5, and here basic characteristics of these three clinical trials will be briefly introduced.

The QUASAR2 trial

This trial explored potential added survival benefits of bevacizumab in addition to capecitabine in adjuvant chemotherapy of stage II and III CRC patients (Kerr *et al*, 2016). Bevacizumab is an antiangiogenic agent commonly used in the treatment of metastatic CRC. The study was approved by the West Midlands Research Ethics Committee (Edgbaston, Birmingham, UK; REC reference: 04/MRE/11/18). From April

25, 2005 to October 12, 2010, a total of 1,952 patients were recruited from 170 hospitals in seven countries with the central trial office based in Oxford, UK. Separate informed consent was obtained from all patients whose blood samples were to be taken for further genotyping.

The SCOT trial

This trial was an international collaborative effort aiming to establish the non-inferiority of a 3-month versus 6-month duration of adjuvant oxaliplatin plus fluoropyrimidine chemotherapy in stage II and III CRC patients (Iveson *et al*, 2018). The ethical approval for the trial was granted by the West Glasgow Research Ethics Committee and equivalent committees in participating countries (244 centres in six countries). Between March 27, 2008, and November 29, 2013, the SCOT trial eventually enrolled a total of 6,088 CRC patients, of which 5,244 patients were recruited in the UK. All participants provided informed consent including the use of blood samples for further research.

The VICTOR trial

This was a phase III randomised controlled trial investigating the efficacy of Rofecoxib in adjuvant chemotherapy of stage II and III CRC patients (Midgley *et al*, 2010). Rofecoxib is an inhibitor of Cyclooxygenase-2 (COX-2) which plays a key role in CRC tumourigenesis. In 2004, however, Rofecoxib was withdrawn globally due to growing concerns on increased risk of cardiovascular events. Therefore, the VICTOR trial was terminated after 2,434 patients had been entered between April 2002 and September 2004. This trial was approved by the Cancer Research Campaign, the Multicentre Research Ethics Committee, and committees at each participating centres in the UK (details presented in the original publication (Midgley *et al*, 2010)). Participants also provided informed consent for use of their blood samples.

As with the SOCCS and UK Biobank, participants with European ancestry from these three trials were genotyped and included in the latest GWAS meta-analysis on CRC risk by Law and colleagues (Law *et al*, 2019). Colorectal cancer cases in the SCOT trial were genotyped using the Illumina® Global Screening Array, whereas cases in the QUASAR2 and VICTOR trials were arrayed using the Illumina® Hap300 and Hap370. Quality control and genotype imputation for these genotyped samples were harmonised with the SOCCS and UK Biobank studies by Law et al. (Law *et al*, 2019).

In particular, the same reference panel as in the SOCCS study—the UK10k release (ALSPAC and TWINSUK studies, April 2014 release), and the 1000 Genome V3 (December 2013 release)—was used for genotype imputation of these three datasets, and all genetic variants were coded based on the GRCh37. Additional details of genotyping, quality control and imputation can be found in the published GWAS meta-analysis (Law *et al*, 2019).

4.3 Study design and data analysis

4.3.1 Candidate genetic association study

Study design

The two main study designs for investigating population-based genetic associations are candidate genetic association studies and genome-wide association studies. Candidate genetic association studies use a hypothesis-driven approach to test a prior hypothesis that a single or a group of genetic variants of interest are associated with certain disease outcomes. Hypotheses are made based on previous evidence, for example biological plausibility that can potentially link a genetic variant in a specific gene with the disease outcome under study. In this thesis, I conducted three candidate genetic association studies investigating three groups of common genetic variants: a) genetic variants previously reported to be associated with survival outcomes of CRC, b) genetic variants associated with CRC risk, c) genetic variants previously linked with survival outcomes of other cancers. The flow chart of study design for these three studies is presented in **Figure 4-5**.

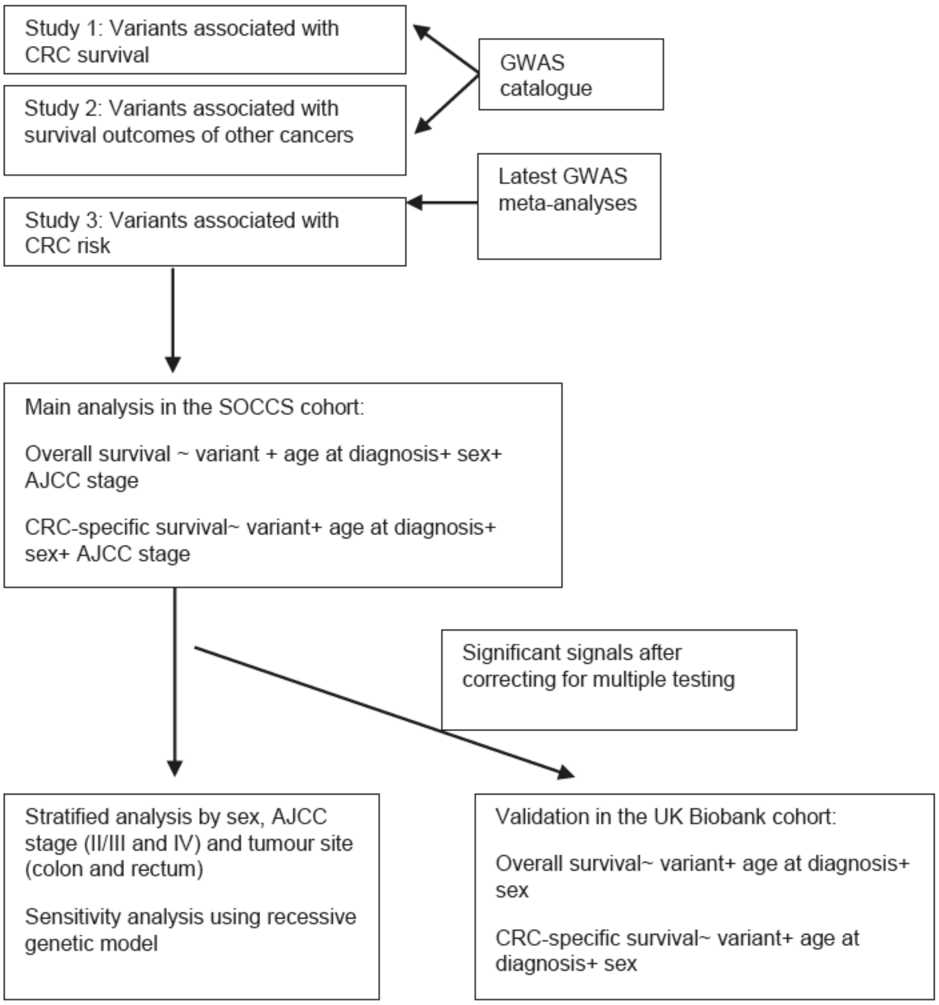

Study 1: Variants associated with CRC survival
Study 2: Variants associated with survival outcomes of other cancers
GWAS catalogue
Latest GWAS meta-analyses
Study 3: Variants associated with CRC risk
Main analysis in the SOCCS cohort:
Overall survival ~ variant + age at diagnosis+ sex+ AJCC stage
CRC-specific survival~ variant+ age at diagnosis+ sex+ AJCC stage
Significant signals after correcting for multiple testing
Stratified analysis by sex, AJCC stage (II/III and IV) and tumour site (colon and rectum)
Sensitivity analysis using recessive genetic model
Validation in the UK Biobank cohort:
Overall survival~ variant+ age at diagnosis+ sex
CRC-specific survival~ variant+ age at diagnosis+ sex

Figure 4-5 Flow chart of the study design for candidate association studies

Variant selection

Search in GWAS catalogue

Variants associated with cancer prognosis were identified by searching the GWAS catalogue (URL1-10). This online platform indexes all published variant-trait associations identified by rapidly accumulated GWASs. In particular, variant-trait associations with p-values less than 10^{-5} identified from either a single discovery study or a combined analysis of discovery and replication datasets are systematically searched and entered into the catalogue on a weekly basis. Detailed inclusion and exclusion criteria of the variant-trait associations indexed in the catalogue can be found at the URL1-10.

The catalogue assigns subject headings to all indexed traits, allowing search for genetic variants by trait of interest. As of December 10 h 2018, there were three subject headings corresponding to prognosis-related traits: 'Survival time', 'Mortality' and 'Disease prognosis measurement'. The traits 'Disease prognosis measurement' and 'Mortality' contained no prognostic outcomes of cancer, and therefore were excluded from the search. The trait 'Survival time' encompassed six secondary traits that included all outcomes related to cancer prognosis: 'disease free survival', 'distant metastasis free survival', 'event free survival time', 'metastasis free survival', 'overall survival', and 'progression free survival'. I retrieved all variant-trait associations under the subject heading 'Survival time' including the secondary traits. Associations between genetic variants and survival outcomes of non-cancer diseases were excluded. Leukaemia was also excluded considering its distinct pattern of progression compared with solid tumours.

CRC risk GWAS meta-analyses

Led by the Colon Cancer Genetics Group here in Edinburgh, the latest GWAS meta-analysis on CRC risk revealed 31 new CRC risk loci, bringing the total number of identified CRC risk variants to ~130 (Law *et al*, 2019). In this study, Law et al. summarised all previously reported CRC risk loci. Therefore, we extracted CRC-risk variants from this study directly, before it got indexed by the GWAS catalogue. Variants reported by another independent newly-published large GWAS meta-analysis (Huyghe *et al*, 2019) that had not been indexed in the catalogue were also extracted and merged with the variant list extracted from Law's study. In addition, we also searched the GWAS catalogue under the subject heading of 'Colorectal Cancer' to check if any previously reported risk loci were missed.

Accounting for linkage disequilibrium

Linkage disequilibrium (LD) reflects the fact that alleles at different positions (haplotypes) of the chromosome occur in a non-random manner—that is they are correlated to each other. The structure of LD is mainly determined by the population origin and genetic distance between two alleles. On the contrary, alleles are in linkage equilibrium if they are inherited independently. There have been two widely used metrics in genetic association studies to quantify the LD between two variants: D' and r^2. Assuming there are two biallelic genetic loci with alleles A_1A_2 and B_1B_2; the frequencies of these two loci are noted as p and q. The deviation (D) of the observed haplotype frequency (f) from the expected frequency can be calculated as:

$$\mathrm{D} = f(A_1B_1) - p_1q_1$$

D' is defined as the normalised D (D'= D/D_{max}), where D_{max} denotes the theoretical maximum of D (detailed formula can be found elsewhere (Lewontin, 1964)). Whilst D' is easy to be calculated, it has been shown to be sensitive to alleles of extreme frequencies (Ranganathan *et al*, 2018). Therefore, an alternative measure, r^2, is more frequently adopted in genetic association studies where risk loci of interest are often of low minor allele frequencies. The r^2 is expressed as:

$$r^2 == \frac{D^2}{p_1 \times p_2 \times q_1 \times q_2}$$

and it ranges from 0 to 1, where 0 means completely independent alleles and 1 means being perfect proxy for each other. In this thesis, the r^2 was used to measure the LD between two variants, and I calculated it in the British population by using the 1000 Genomes phase 3 GBR data from the Ensembl portal (URL4-9). I chose an r^2 of 0.2 as the threshold, and any pairs of variants with an r^2>0.2 were considered as being in LD. When LD was detected, the variant associated with the trait of interest (CRC risk or cancer survival) with the smaller p-value was retained.

Statistical analysis

Genetic model

Based on the assumed pattern of inheritance, each genetic variant can be coded in different ways, also known as genetic models. There are three main genetic models that are widely used in genetic association studies: dominant model, recessive model and co-dominant model. For a biallelic genetic variant with alleles A and a, the dominant model assumes that it takes only one risk allele (A) to exert the effect. That is, the model compares individuals with the genotype AA or Aa versus individuals with aa. The recessive model, however, assumes that an individual must have two copies of risk alleles (AA) to show the effect, and therefore it compares AA versus Aa + aa.

As for the co-dominant model, it assumes that the genetic risk conferred by Aa lies somewhere between AA and aa. The co-dominant model can be further divided into additive and multiplicative models according to the pattern of relative risk among these three genotypes. The conceptual framework of different genetic models is depicted in **Figure 4-6**.

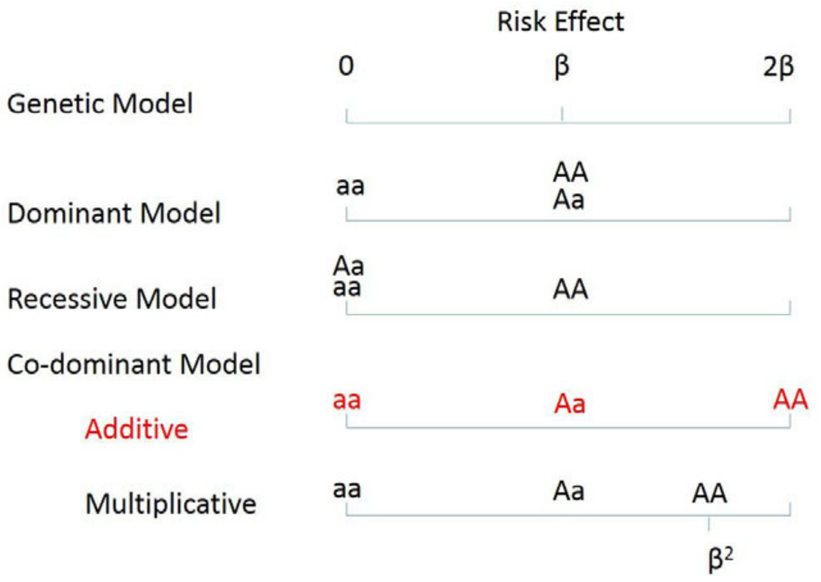


Figure 4-6 Conceptual framework of different genetic models

Among these genetic models, the additive model has been most widely adopted as the assumed model in both candidate and genome-wide association analyses. This model assumes a linear increase in risk for each allele copy. Evidence based on simulation studies indicated that the additive model also has acceptable statistical power to detect possible dominant effects in addition to additive effects of a certain genetic variant (Lettre *et al*, 2007). Moreover, the number of risk alleles of imputed genetic variants are presented as expected values in a continuous order (range from 0 to 2), and can be naturally modelled in an additive pattern. Therefore, I chose the additive genetic model as the primary model of analysis of genetic associations. However, simulation analysis suggests that the additive model has limited statistical power to identify a possible recessive effect (Lettre *et al*, 2007); thus I also rounded the imputed allele dosage to integers—that is, any imputed allele dosage that is greater than 1.5 was coded as 1, otherwise coded as 0—so as to fit a recessive model and investigated their associations with CRC survival. Notably, there has been no previous evidence indicating the mode of inheritance being recessive for any of these included candidate genetic variants in this thesis. Therefore, investigation using recessive model should be considered as sensitivity analysis.

Polygenic risk score

The previous section introduced genetic models used to code each single variant before analysing the genetic association with CRC survival. In addition, I also investigated the combined effect of a set of genetic variants, for example CRC-risk variants as a group, on survival outcomes by constructing a polygenic risk score (PRS) for each patient. I calculated the PRS using a naive unweighted approach due to lack of prior knowledge on the potential distribution of effects on survival for these candidate variants. This approach assumes equivalent effect sizes for all the variants included in the PRS. To create the PRS for each patient, I added the number of risk alleles of all the candidate variants in a group. To take variants associated with CRC risk as an example, I first harmonised their direction of effects on CRC risk, and identified genetic alleles responsible for increased CRC risk. Then the exact number or imputed dosage of risk-increasing alleles for each variant was summed up.

Time-to-event outcome

In contrast to common outcomes that are either categorical or continuous, survival outcomes have a special nature; that is, the outcome for each individual under study consists of two elements: event status (D) and survival time (T). In this thesis, the events of interest were defined as death of any cause for overall survival and CRC-related death for CRC-specific survival. The status of patients with event occurrence was coded as ‘1’ and patients who were alive until the last time of updating the death records (also known as right censored observations) were coded as ‘0’. As mentioned previously, the other element—survival time—was defined as the time span from the date of definitive treatment (SOCCS) or CRC diagnosis (UK Biobank) to the date of event or censoring; the time period was measured by years of follow-up. Finally, the survival outcome for each CRC patient was expressed as a combined measure: (T, D).

Kaplan-Meier estimator and log-rank test

In order to estimate the probability of surviving a certain amount of time (denoted as t) after CRC diagnosis, a few statistical concepts need to be introduced. At first, I define two random variables T_s for the survival time, and T_c for the censoring time in response to the aforementioned time element for the survival outcome of each patient. An assumption here is that T_s and T_c are independent of each other. Then, the cumulative distribution function (CDF) for T_s and T_c can be expressed as:

$$F_s = \mathrm{P}(T_s \leq \mathrm{t})$$

and

$$F_c = \mathrm{P}(T_c \leq \mathrm{t})$$

In the context of survival analysis, however, the complementary function of the CDF of T_s is commonly used as we are often more interested in the probability of patients surviving at least a certain time span. Hence, the survival and censoring function are defined as:

$$\mathrm{S(t)} = \mathrm{P}(T_s > \mathrm{t}) = 1 - F_s(t)$$

and

$$\mathrm{C(t)} = \mathrm{P}(T_c > \mathrm{t}) = 1 - F_c(t)$$

Since we are only interested in evaluating future risk of death for each patient, the primary focus of this section is concerning estimating the S(t).

Amongst many methods proposed to estimate the S(t), the Kaplan-Meier estimator has been the most widely-used statistic, especially in epidemiological and clinical research settings. The estimator was named after Kaplan and Meier who first proposed it in 1958 (Kaplan & Meier, 1958). The Kaplan-Meier estimator is defined by the following expression:

$$\mathrm{S(t)} = \prod_{i:t_1 \leq t} (1 - \frac{d_i}{n_i})$$

Here, t_i refers to a certain time point when at least one death occurs; d_i denotes the number of deaths that occur at t_i; and n_i refers to the size of risk set which is measured by the number of individuals that are still alive entering the time ti. Given the fact that each ti is a time point actually observed from the sample dataset, the distribution of estimated S(t) will not be continuous. Instead, there are 'jumps' at time points where events occur.

In this thesis, I calculated the Kaplan-Meier estimates for CRC patients and plotted the estimates against the follow-up time to make survival curves. Patients were grouped and their survival curves were stratified by prognostic factors such as AJCC stage for descriptive purposes and potential genetic variants as a graphic display for the study results. The Kaplan-Meier estimates were calculated using the packages 'survival' (URL4-10) and 'survminer' (URL4-11) in the R software (version 3.5.1, URL4-12).

In addition, a log-rank test was conducted to test the difference across multiple survival curves (Peto & Peto, 1972). Assuming a two-group comparison, I first used i=0 or 1 to denote the two groups, and j to define each time point for observed events. Let N be the number of individuals at risk and O be the number of observed events at the same time point. The null hypothesis assumes that the two groups have an identical hazard function h(t). Under this hypothesis, the expected number of events occurring at each time point can be calculated, and a test statistic *Z* is defined as follows:

$$Z_i = \frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{j} Var_{i,j}}}$$

It has been demonstrated that the statistic is asymptotically normal as j increases; a p-value can then be derived from a standard normal distribution (Peto & Peto, 1972). I conducted the log-rank test using the 'survival' (URL4-10) package.

Proportional hazards model

To estimate effects of prognostic factors, such as genetic variants and other non-genetic factors, on survival outcomes of CRC, regression techniques are needed to accommodate single or multiple variables simultaneously in a model. The proportional hazards model, also known as Cox regression model, is the most widely applied method to quantify the association between prognostic factors and survival outcomes. The basic ideas and principles of the Cox regression model will be briefly introduced here.

Proposed by Cox (Cox, 1972), this method is based on the proportional hazards assumption—the effect of a certain factor on the survival outcome remains constant over time. A few statistical concepts need to be defined to explain the method. Firstly, based on the survival function described in the previous section, the hazard (h) can be interpreted as the derivative of the survival function, or intuitively as the instantaneous risk of death for those individuals who are at risk. The hazard function is expressed as:

$$\mathrm{h(t)}d_t = \mathrm{P(T < t} + d_t|\mathrm{T \geq t)}$$

The left side of the of equation refers to the risk of death during a short time period d_t, and it is related to the survival function by the right side which denotes the probability of death within d_t given that the individual has survived up to the time point t. The Cox regression model relates potential prognostic factors under study to the hazard function by the following equation:

$$\mathrm{h(t)} = h_0(\mathrm{t})\mathrm{exp}(b_1X_1 + b_2X_2 + \cdots + b_nX_n)$$

It should be noted that in this expression, the hazard function h(t) is the expected (or predicted) hazard at the time point t instead of the observed hazard derived from the survival estimates. On the right side, coefficients b_1 to b_p are effects on the hazard at t per single unit change of each factor X (also known as covariates). The h_0(t) is the baseline hazard—the hazard function of individuals with all the factors at baseline level (all Xs equal to zero). In order to clearly explain the coefficients, or the prognostic effects estimated from the Cox model, we consider a simple univariable Cox model:

$$\mathrm{h(t)} = h_0(\mathrm{t})\exp(b_1 X_1)$$

Assuming the covariate X is a binary variable, such as sex, coded with values 0 or 1, the ratio of hazards between the two groups of patients with X=1 or X=0 is defined as the hazard ratio (HR) which can be expressed as:

$$\mathrm{HR} = \frac{h_0(\mathrm{t})\exp(b_1 X_{x=1})}{h_0(\mathrm{t})\exp(b_1 X_{x=0})}$$

namely:

$$\mathrm{HR} = \exp(b_1)$$

Notably, the exp(b_1) is a constant that does not depend on the time t, pointing to the proportional hazard assumption of the Cox regression model.

In this thesis, the Cox regression model was fitted using the ‘coxph’ function from the R package ‘survival’. Estimation of the regression coefficients when applying the ‘coxph’ function to fit a Cox model relies on a method called partial likelihood; a Wald statistic is calculated to generate a two-sided p-value for the inference of the statistical significance of the estimated coefficient (HR). Additional technical details about this method can be found in the original paper by Cox (Cox, 1972). As presented in the flow chart of study design, I investigated associations between three sets of candidate genetic variants and overall and CRC-specific survival outcomes using the SOCCS study. Age at diagnosis, sex and AJCC tumour stage at presentation for each patient were fitted in the Cox regression as covariates in addition to the genetic variant. For possible validation analysis using the UK Biobank cohort, covariates only included sex and age at CRC diagnosis due to unavailable data of AJCC stage. I also conducted Cox regression analysis in different strata of CRC patients stratified by sex (male and female), AJCC stage (stage II-III and stage IV) and tumour site (colon and rectum cancer).

Adjusting for multiple testing

I conducted three candidate genetic association studies with each study including a set of variants whose associations with CRC survival were tested simultaneously. If a normal threshold of $p<0.05$ was applied for each test, it would have caused remarkably higher false positive rates (or type I error) when considering a set of tests together. There are two approaches that have been widely used to adjust for this problem: the Bonferroni correction and the false discovery rate (FDR), also known as the Benjamini-Hochberg approach. The Bonferroni correction adjusts the significance threshold to α /n where the n denotes the number of tests to be conducted. The α here refers to the familywise error rate (FWER) which measures the probability of making at least one type I error (false positive) in the whole family of tests. Similar to a single statistical test, this threshold of the FWER is often set to be $\alpha=0.05$. Given that the Bonferroni correction directly adjusts the α level, it can be readily applied to estimate the statistical power using an adjusted α. Power calculation will be described in the following section.

Although effective in controlling type I error, the Bonferroni correction could lead to increased non-rejection of a false null hypothesis (type II error). Therefore, in addition to the Bonferroni correction, we also adopted a less conservative approach--the FDR approach (Benjamini & Hochberg, 1995)--to evaluate the significance of results and to screen for potential signals from the candidate genetic association studies that merited further validation. This approach is designed to control the expected proportion of false positive findings. To be specific, instead of adjusting for the α threshold to assess significance, the FDR approach adjusts all the p-values generated from the set of tests. It first sorts the n p-values in an ascending order, with p_1 the smallest and p_n the largest. With the largest p_n remains the same, the adjusted ith p-value (i<n) is expressed as:

$$p_i = \mathrm{Min}(\mathrm{adjusted}\ p_{i+1}, p_i(\frac{n}{i}))$$

Then a pre-specified α threshold (<0.05 for example) is used to evaluate the adjusted p-values.

Power estimation

Statistical power of a genetic association test estimates the probability of detecting significant association of a given variant with CRC survival when a genetic effect truly exists. Owzar et al. provided a widely-used approach to estimate the statistical power of a genetic association study on survival outcomes (Owzar *et al*, 2012). Some basic metrics used in this approach to calculate the power are introduced here. Firstly, the minor allele frequency (MAF) is the observed frequency of the second commonest allele of a given genetic variant (assuming the variant is biallelic) in a specific population. Another important metric is the effect size which in survival analysis is presented as the hazard ratio. The event rate, namely the proportion of overall or CRC-specific deaths in the study sample, is also needed for the power estimation. Combined with the sample size and the α level, these metrics were integrated into the formula provided in Owzar's paper to estimate the power (Owzar *et al*, 2012). Notably, the method developed by Owzar et al relies on a score statistic to make statistical inference which is different from the Wald statistic used in the 'coxph' function that we adopted to test the genetic associations. Nonetheless, these two statistics have been shown to be asymptotically identical as the sample size increases (Owzar *et al*, 2012). In order to adjust for multiple testing, we specified the α level as 0.05/n where n was the number of variants in each candidate genetic association study. We also chose the additive genetic model to calculate the power in response to our main association analyses. Power estimation was conducted using the 'survSNP' package (URL4-13). Additional technical details regarding the method can be found in the original publication (Owzar *et al*, 2012).

4.3.2 Predictive modelling

Study design

Following the previous section where I validated individual association between each genetic variant that had been linked to CRC prognosis and survival outcomes of CRC patients in the SOCCS study, here in this section, I focused on exploring the predictive value of variants that had been previously linked with CRC survival as a group in forecasting survival outcomes of CRC. As discussed in Chapter 3, a multivariable prediction model is an algorithm that provides an average estimate of the probability by combining all predictors together. In this study, I fitted Cox regression models to develop a genetic predictor and then evaluated its performance in predicting the survival rates of CRC patients.

In order to establish and validate a genetic predictor combining these variants together, two datasets—UK Biobank and the SOCCS study—were leveraged. As introduced in Chapter 1, a substantial amount of within-stage variation has been observed in relation to survival outcomes of CRC patients. From a clinical perspective, it has been of great interest to develop novel predictors to further improve prediction on the basis of tumour stage. Clinical interest also lies in predicting survival outcomes within specific stage, for example CRC patients at stage II-III, where the optimal treatment strategies for patients with varied prognostic profiles are still to be decided. Given that tumour stage is unavailable in UK Biobank, this dataset was used as the training set to develop the genetic predictor; external validation of the genetic predictor derived from the UK Biobank and within-stage prediction were then conducted in the SOCCS study. The flow chart of the study design is shown in the **Figure 4-7** below:

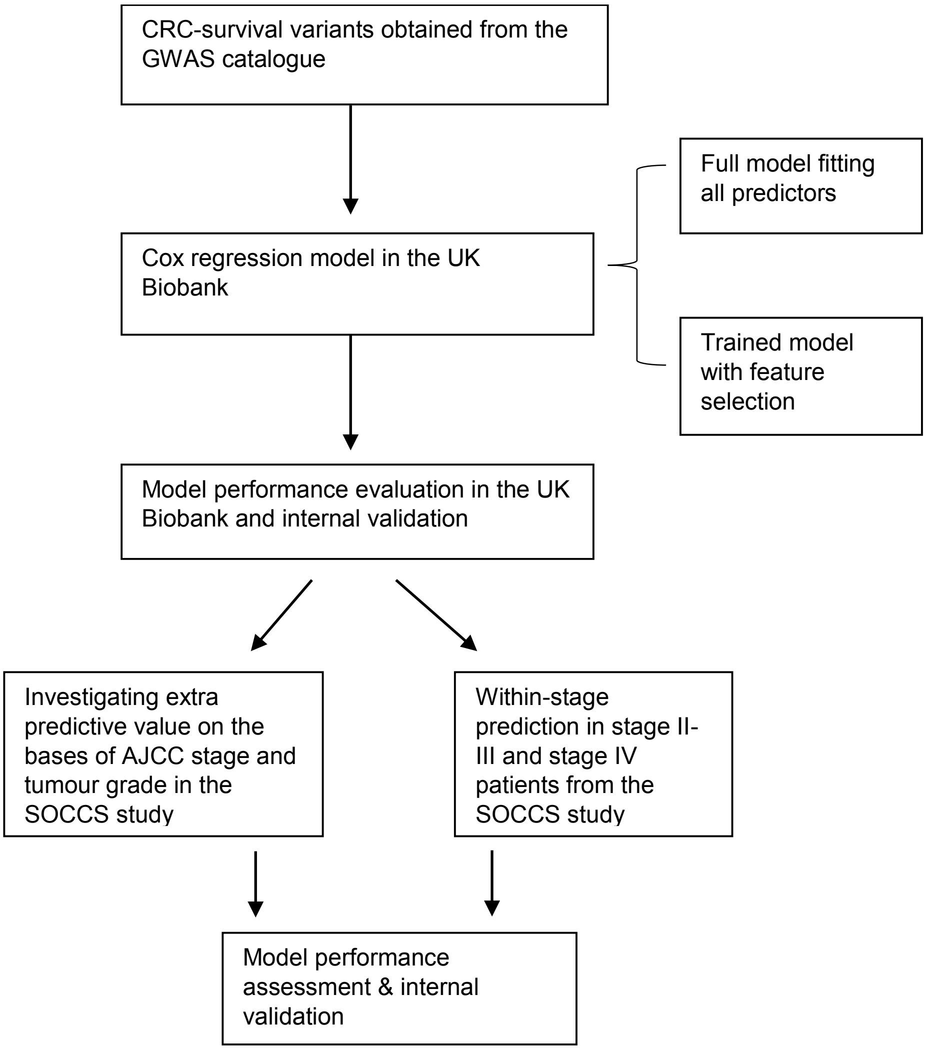

CRC-survival variants obtained from the GWAS catalogue
Cox regression model in the UK Biobank
Full model fitting all predictors
Trained model with feature selection
Model performance evaluation in the UK Biobank and internal validation
Investigating extra predictive value on the bases of AJCC stage and tumour grade in the SOCCS study
Within-stage prediction in stage II-III and stage IV patients from the SOCCS study
Model performance assessment & internal validation

Figure 4-7 Flow chart of study design for prediction modelling

Developing the genetic predictor

Full model

In order to apply the predictor across different datasets, I first harmonised the reference allele of each genetic variant across the published GWASs and our two study cohorts--the SOCCS and UK Biobank study. In contrast to the analysis approach used in the preceding section where a single variant was fitted along with age, sex and AJCC stage, here I added all the 43 variants that were reportedly associated with CRC survival into one Cox regression model in the UK Biobank cohort to estimate the coefficient for each variant. All variants were coded under an additive genetic model. The linear predictor (LP) combining the 43 genetic variants is expressed as:

$$\mathrm{LP} = \sum_{i=1}^{n} \beta_i \mathrm{X}_i$$

Here β_i denotes the coefficient of the i_{th} variant extracted from the fitted Cox model, and X_i is the allele count.

Feature selection

As discussed in Chapter 3, it has been widely accepted that an increased number of predictors leads to higher events per variable (EPV) which further results in increased risk for over-optimism, where the developed model is over-fitted to the data under study and therefore incurs poor generalisability when applied to external data. Therefore, feature selection techniques have been developed and widely employed to reduce the model dimension.

In this study, I first used a least absolute shrinkage and selection operator (LASSO) for feature selection. This method reduces over-fitting by penalising the Cox regression coefficients towards zero. In particular, the LASSO regression seeks to minimise the following expression:

$$\sum_{i=1}^{n} (\mathrm{y}_i - \sum_{j} \mathrm{x}_{i,j} \beta_j)^2 + \lambda \sum_{j=1} |\beta_j|$$

The former part of the expression refers to the sum of squared errors in general linear models. The later part is the penalty term with the λ as the tuning factor that controls the strength of penalisation. The addition of the penalty term shrinks the coefficients and some of them can be reduced to zero, and therefore are excluded from the final model. In practice, a range of lambdas can be used to generate different models. Breiman et al. suggested an 'one-standard-error' approach utilising cross-validation to assist selection of the tuning factor (Breiman *et al*). Cross-validation is a resampling technique that randomly splits the study sample into k groups (also called k-fold cross-validation). Each time, the model takes out one group as a test dataset and uses the remaining groups as a training dataset. Then the statistical model is fitted repeatedly and the parameters of interest are summarised afterwards. Following this method, we used a 10-fold cross-validation to identify the lambda in response to the most parsimonious model where the cross-validation prediction error is within one standard error of the minimum. Technical details of the method can be found in the original publication (Breiman *et al*). The LASSO regression was conducted using the ‘glmnet’ R package (URL4-14).

In addition to the LASSO approach, I also conducted backward selection to screen for potential useful variants in predicting CRC survival. This method starts with a full model with all the candidate variables included. The model then removes the least significant variable (evaluated by its p-value) and iterates this process until the all the remaining variables are below the pre-defined p-value threshold. Here in this thesis, I set the p-value threshold at 0.15 based on recommendations from a previous publication which conducted optimisation analysis to determine this threshold (Heinze *et al*, 2018). The backward variable selection was conducted using the ‘validate’ function in the ‘rms’ R package (URL4-15).

Model performance assessment

Evaluating the performance of a given prediction model is to quantify how concordant the predictions made from the model are with the observed outcomes. As opposed to the previous section where I focused on the relative risk effect (hazard ratio) of a single genetic variant, here the absolute probabilities of survival were used in the setting of risk prediction. To quantitatively evaluate the concordance, both the observed and predicted survival estimates are needed. The method used to generate Kaplan-Meier estimates for the observed survival rates has been introduced in the previous section. I obtained the predicted survival estimates by one minus predicted death risk which had been generated in the Cox regression. There are two major properties relevant to the model performance: discrimination and calibration. In this thesis, I adopted multiple metrics to quantify the model performance based on these two aspects.

The discriminative ability of a given prediction model reflects how good the model is to distinguish between individuals with and without the outcome of interest. In the case of the time-to-event outcome, the Harrell's C statistic is the mostly commonly used measure of the discriminative ability (Harrell *et al*, 1996). The C statistic calculates the proportion of randomly selected pairs of subjects in which an individual with longer observed survival time also exhibits higher predicted probability of survival. A C statistic of 0.5 suggests random predictions with null predictive value. Notably, pairs with an individual who has a shorter censoring time than the survival time of the other individual are not counted when calculating the C statistic. I calculated the C statistic using the "rcorr.cens" function in the 'rms' package (URL4-15). It is worth mentioning that the "rccor.cens" function derives the C statistic along with its standard error from another measure called Somers' D statistic. This statistic is defined as:

$$D_{xy} = \mathrm{E}(\mathrm{sign}(\mathrm{X}_1 - \mathrm{X}_2)\mathrm{sign}(\mathrm{Y}_1 - \mathrm{Y}_2))$$

Where (X_1,Y_1) and (X_2,Y_2) are two randomly selected pairs of random variables denoting the predicted and observed survival probability respectively. Ranging from -1 to 1, the Somers' D is related to the C statistic by:

$$\mathrm{D} = 2 \times (\mathrm{C} - 0.5)$$

In order to identify the added predictive value of genetic variants on the basis of other variables, I compared the C statistics along with their confidence intervals derived from the model with and without genetic predictors. Moreover, I conducted a U-statistic based test to determine whether the predicted probabilities from one model were more concordant with the observed estimates than the other model (Harrell Jr, 2015). This test quantifies the proportion of subject pairs where there is less difference between predicted and observed estimates in one model than the other. I performed the U-statistic based test using the 'rcorrp.cens' function in the 'Hmisc' R package (URL4-16). A p-value <0.05, which means there is 95% of chance that one model is more concordant than the other, was considered as statistically significant improvement.

The other important property to evaluate the model performance is calibration—the overall agreement between predicted and observed probability. In the context of survival analysis where both the predicted and observed survival probabilities vary by time, the calibration of the model is therefore usually evaluated at a fixed time point. From a clinical point of view, I employed the 5-year survival as the observation time to compare the agreement of predicted and actual survival outcomes. I first grouped the individuals using the quartiles of the linear predictor of the prediction model under study. For each group, the observed Kaplan-Meier survival estimates were derived and contrasted with the mean predicted survival estimates. The calibration plot was constructed by plotting the predicted probabilities against the observed estimates for each risk group of patients. This plot allowed a visual assessment of the departure from the ideal calibration line with a slope of 1 and intercept of 0.

In addition to evaluating the model calibration graphically, I also performed a Hosmer–Lemeshow (HL) test to assist determining whether the model was well-calibrated (Hosmer Jr *et al*, 2013). The HL test statistic (H) is given by the following equation:

$$\mathrm{H} = \sum_{g=1}^{G} \left(\frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}} \right)$$

where G refers to the number of groups categorised by the linear predictor, O_1 and O_0 denote the number of observed events and non-events, and E_1 and E_0 represent the expected number assuming that the proportion of observed and predicted events are the same across all the risk groups (G). It has been shown that the test statistic asymptotically follows a chi square distribution from which the p-value is generated. I employed the ‘hoslem.test’ function in the ‘ResourceSelection’ R package (URL4-17) to conduct the HL test and examine the model calibration at the 5th year after diagnosis. A p-value<0.05 was deemed as significant departure from calibration.

Adjusting for over-optimism

Adjusting the model performance

Model performance metrics, such as the C statistic, are commonly too optimistic if they were estimated from the dataset used to develop the model. This often leads to poorer prediction accuracy when applying the model to new patients. Hence, methods have been proposed to adjust this over-optimism; this procedure is also commonly known as internal validation of the developed model. In this case, since I developed the genetic predictor in the UK Biobank cohort, over-optimism was expected in the C statistic derived from the same cohort. With regard to the SOCCS cohort, the first step of validating the genetic predictor was immune to over-fitting as the model was fixed, and the SOCCS cohort was used as an external source. However, over-optimism arose in the second step when the genetic predictor was combined with other variables and the model was re-fitted. Therefore, internal validation was also performed for the new model established in the SOCCS study.

I adopted a bootstrapping approach in this thesis to correct for potential over-optimism of the C statistic. In contrast to cross-validation which divides the sample into equally sized groups, bootstrapping is a re-sampling technique that draws samples from the dataset under study and therefore it allows replacement. I generated 200 bootstrap samples and repeatedly calculated 200 C statistics following the procedure described above. Then the corrected C statistic along with its confidence interval was derived from the bootstrap sample distribution. Bootstrapping was conducted using the 'validate' function in the 'rms' R package (ULR 4.15).

Adjusting the model presentation

As stated in the previous sections, I re-fitted the Cox model by combining the genetic predictor and other variables, which again introduced over-optimism into the model. Although the C statistic was adjusted to evaluate the model performance, the coefficients estimated from the newly-fitted model still remained inflated and therefore needed to be corrected before the model being presented. This was done by calculating and applying a metric called 'shrinkage factor' (Harrell Jr, 2015; Steyerberg, 2019). In particular, I first took a bootstrap sample, fitted the same Cox model, and obtained a new set of coefficients. Then the new coefficients were utilised to create a new linear predictor for each individual in the cohort. I re-fitted the Cox model using the new linear predictor as the covariate and extracted the coefficient of the predictor. This process was repeated in 200 bootstrap samples and the average of these coefficients was retrieved as the shrinkage factor (SF). I then applied this factor to shrink the original coefficients. Based on the expression of Cox regression model introduced previously, the ultimate model predicting the probability of surviving time t was presented as:

$$\mathrm{S(t)} = S_0(\mathrm{t})^{\exp((SF\times(\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n))}$$

S_0 refers to the baseline survival probability which can be derived from the baseline hazard estimated from the Cox regression. Betas are the regression coefficients from the original fitted model. Further details regarding the model shrinkage can be found in references of (Harrell Jr, 2015; Steyerberg, 2019).

4.3.3 Genome-wide association study

Study design

As described in Chapter 2, the final part of the thesis is a GWAS investigating the whole genome to identify potential novel variants associated with CRC survival. Overall, we employed a two-step approach for this section. Firstly, the SOCCS cohort was used as the discovery dataset to perform the main GWA analysis. Secondly, any genetic variants identified at GWAS significance ($p<5x10^{-8}$) were further validated by a meta-analysis combining datasets of summary statistics from the UK Biobank cohort, the QUASAR2 trial, and the pooled dataset of the VICTOR and SCOT trial. With respect to covariates, age at diagnosis, sex and AJCC stage were adjusted for the SOCCS cohort and the three clinical trial datasets. For the UK Biobank cohort, only age at diagnosis and sex were adjusted as AJCC stage was unavailable. For the outcome of CRC-specific survival, the validation analysis was only conducted in the UK Biobank cohort due to data availability. In addition, the summary statistics derived from the GWAS in the SOCCS were utilised for gene and gene-set based enrichment analysis to explore putative genes and biological pathways associated with CRC survival. I also performed GWA analysis separately within stage II/III and stage IV CRC patients in the SOCCS study. **Figure 4-8** presents the overall study design for this section.

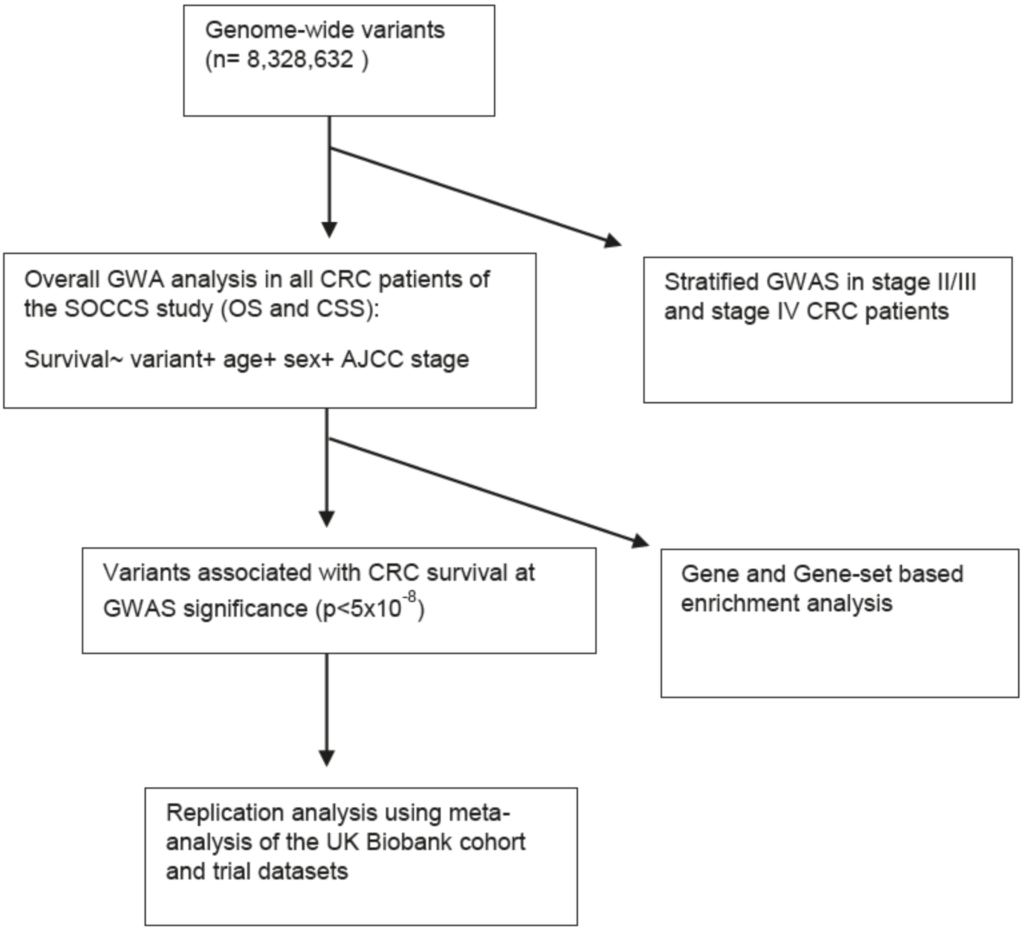

Genome-wide variants (n= 8,328,632)
Overall GWA analysis in all CRC patients of the SOCCS study (OS and CSS):
Survival~ variant+ age+ sex+ AJCC stage
Stratified GWAS in stage II/III and stage IV CRC patients
Variants associated with CRC survival at GWAS significance (p<5x10^{-8})
Gene and Gene-set based enrichment analysis
Replication analysis using meta-analysis of the UK Biobank cohort and trial datasets

Figure 4-8 Flow chart of study design for the genome-wide association study

Survival outcomes transformation and analysis

I used a standard Cox regression model in preceding sections to estimate the association between a single variant or multiple variants on survival outcomes of CRC. Although the standard Cox model has been widely used in survival analysis of genetic associations, it involves heavy computational burden particularly in the context of GWAS where millions of associations are tested. To improve computational efficiency, here I transformed the survival outcomes and employed a modified approach based on the standard Cox model which has been adopted by previous published GWASs (Joshi *et al*, 2017; Timmers *et al*, 2019). To be specific, I first transformed the right-censored survival outcomes including both the overall survival and CRC-specific survival to Martingale residuals for each patient in the SOCCS study. The Martingale residual is defined as the difference between the observed number of events for the i^th^ individual in the cohort (either 0 or 1) and the expected number of events for the same individual at time t (Therneau *et al*, 1990). Mathematically it is expressed as:

$$\widehat{M_\iota} = \delta_i - \widehat{\Lambda}_0(\tau_i)e^{\widehat{\Upsilon}_1 Z_1 + \cdots + \widehat{\Upsilon}_k Z_k}$$

where Λ0 refers to the baseline cumulative hazard at time point τi, and δi denotes the observation event (0 or 1). Z1 to Zk are covariates excluding the genetic variant of interest. In this study, Martingale residuals were calculated by fitting a standard Cox regression model with age at diagnosis, sex and AJCC as covariates using the 'coxph' function. It has been demonstrated that the Martingale residual is linearly related to the potential effect of the remaining variable, that is, the genetic variant under analysis (Therneau *et al*, 1990). In concordance with previous studies (Joshi *et al*, 2017; Therneau *et al*, 1990), the calculated Martingale residuals were then scaled up by 1/(the proportion of events) to generate a 1:1 correspondence with the regression coefficients of genetic variants to be analysed. Finally, we regressed the scaled residuals on the genetic variant (coded under the additive model) by fitting a univariable linear regression model as follows:

$$P = \beta X + e$$

Where β is the estimated genetic effect—approximated log-transformed hazard ratio of each variant. A Ward test was performed to obtain the two-sided p-value to examine if the β was significantly different from the null. I repeated the linear regression model for each of the eight million variants throughout the genome using the SNPtest (v2.50) software (URL4-18). Although it has been widely-accepted that the Martingale residual approach is a reasonable approximation to the standard Cox model with considerably reduced computational load, simulation studies found that there is a 2%-4% loss of statistical power for the residual approach (Reynisson, 2018). Therefore, I re-fitted the standard Cox regression models for genetic variants identified from the Martingale residual based approach at a relatively lenient threshold $p<5x10^{-7}$, and variants at GWAS significance ($p<5x10^{-8}$) from the re-fitted Cox models were considered as significant GWAS signals and were passed on for replication.

Considering that metastatic CRCs (stage IV) may have different genetic components in contrast with locally advanced CRCs (stage II and III), stratified GWA analyses were conducted following the same procedure above in stage II/III and stage IV CRC patients separately.

To visualise potential signals of genetic effects across the genome, I created Manhattan plots on a genomic scale by stacking the log-base-10 of the p-values for associations of all genetic variants grouped by chromosomes in genomic order. Genetic variants with strong associations, for example with $p< 5x10^{-8}$, tend to rise up high in the plot. Manhattan plots were generated using the 'Manhattan' function in the 'qqman' R package (URL4-19). In addition to Manhattan plots, I also constructed quantile-quantile (QQ) plots to visualise the genetic effect of all tested variants. The quantiles of log-base-10 of observed p-values obtained from the GWAS was plotted against the quantiles of log-base-10 of p-values sampled from a theoretical uniform distribution ranging from 0 to 1. A straight line with a slope of 1 is expected if overall there is no genetic effect on the outcome or if the study is underpowered. Due to the LD among variants, a real genetic effect manifests as an upward tail in a QQ plot. I created QQ plots using the 'qqPlot' function in the 'GWASTools' R package (URL4-20).

Unadjusted confounding can lead to systematic inflation in GWAS results—that is, a global excess of higher observed p-values than the theoretical distribution. This is often caused by unaccounted population structure and can be visualised by a systematic upward deviation from the diagonal of the QQ plot. In addition to the graph, I also quantified the extent of possible inflation by calculating the inflation factor denoted by lambda (λ). According to the method proposed by Aulchenko et al. (Aulchenko *et al*, 2007), I calculated the genome-wide inflation factor by regressing the observed p-values on the theoretical distribution using the 'estlambda' function in the 'GenABEL' R package (URL4-21). A lambda value>1.1 was considered as presence of inflation (Yang *et al*, 2011).

I adopted the same method as in previous candidate genetic association studies to estimate the statistical power of this GWAS (Owzar *et al*, 2012). An α level at GWAS significance ($5x10^{-8}$) was employed along with the same set of other metrics including the sample size, proportion of events, effect sizes and minor allele frequencies.

Replication analysis

Associations of genetic variants with CRC survival that reached GWAS significance ($p<5x10^{-8}$) were validated by meta-analysis combining three datasets. Standard Cox regression models adjusted for age at diagnosis, sex and AJCC stage were conducted in the QUASAR2, VICTOR and SCOT trials to investigate associations between variants discovered in the SOCCS and overall survival of CRC patients. Both overall and CRC-specific survival were included as outcomes in the UK Biobank cohort without adjusting for AJCC stage due to unavailable data. I then extracted the summary statistics from the regression models fitted in these datasets including the regression coefficients along with their standard errors of the genetic variants. Given the concordant ethnicity across all included study cohorts, I implemented a fixed-effect model meta-analysis, which has been widely used in GWAS meta-analyses, to obtain pooled estimates of genetic effects of these variants. In contrast with the random-effects model used in the meta-analysis in Chapter 3, the fixed-effect model assumes a constant effect of the factor under study across all included datasets. I employed an inverse-variance weighted (IVW) estimator to combine effect estimates extracted from each validation study. To be specific, let X_i be the ith effect estimate, namely the regression coefficient of the ith study. The IVW approach defines the weight of each included study as:

$$\mathrm{W}_i = \frac{1}{V_i}$$

Where V_i is the variance of X_i. Weighted by W_i, the IVW estimator of the pooled effect size X_{FE} is then expressed as:

$$\mathrm{X}_{FE} = \frac{\sum W_i X_i}{\sum W_i}$$

With the variance as:

$$\mathrm{V}_{FE} = \frac{1}{\sum W_i}$$

X_i is normally distributed with a sufficiently large sample size, so is the X_{FE}. Therefore, a Z score can be generated to infer the significance of X_{FE}. In addition, I used the I^2 statistic to evaluate potential heterogeneity of effect sizes from each included study (Higgins *et al*, 2003). A two-sided p-value<0.05 was considered as a statistically significant association that was successfully replicated in the three independent datasets. I carried out the meta-analyses and created forest plots displaying the individual and pooled effect estimates using the ‘metagen’ function in the ‘meta’ R package (URL4-22).

Gene and Gene-set based enrichment analysis

By performing a GWAS, I interrogated individual effects of approximately eight million genetic variants on CRC survival outcomes. However, survival of CRC patients is a complex trait resulting from polygenic effects of variations in multiple genomic regions. Each genetic variant may contribute only a very small fraction of the effect which can be undetectable by a GWAS with limited statistical power. Therefore, gene and gene-set based pathway analysis has been proposed to group millions of genetic variants based on their known biological function, and then to test the joint effect of these variants as a group on the outcome of interest (de Leeuw *et al*, 2015). This type of analysis can not only indicate functional implications by identifying significant signals enriched in genes or sets of genes involved in key biological pathways, but also provide higher statistical power due to the reduced number of statistical tests.

Here I employed the Multi-marker Analysis of GenoMic Annotation (MAGMA) approach to conduct gene and gene-set based pathway analyses (de Leeuw *et al*, 2015) using the summary statistics of the main GWAS including all SOCCS patients. The MAGMA analysis was implemented using the online portal of Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA, URL4-23) (Watanabe *et al*, 2017). For the gene based test, genetic variants included in the preceding GWAS were first annotated and mapped to a total of 35,808 genes curated in the Ensembl genome database (build 85) (URL4-9). Then the genetic effects on CRC survival outcomes of variants within the same gene were aggregated. To be specific, the MAGMA approach proposed a SNP-wise model leveraging the summary statistics of the GWAS results. In particular, the summary statistics including the effect estimates, standard errors and p-values for all genetic variants mapped in a specific gene were used to re-construct their test statistics (MAGMA uses Chi-square statistic). Then the mean statistic of these variants in the gene was estimated to infer the statistical significance of the gene-wise effect. It is worth noting that the linkage disequilibrium needs to be adjusted in order to model the distribution of the variants' test statistics. In this study, the 1000 Genomes Project phase 3 was adopted as the reference panel to create an LD matrix and account for the LD structure among the variants. Similar to the GWAS, I created Manhattan plots to show the gene based signals.

Results of gene based tests were utilised in subsequent gene-set based pathway analysis. Mapped genes were aggregated and assigned to ~ 4,728 gene-sets involved in different biological pathways curated in The Molecular Signatures Database (MSigDB v5.3)(Liberzon *et al*, 2015; Watanabe *et al*, 2017). Leeuw et al. proposed a competitive model to analyse the gene-set based effect. This test examines whether associations of genes encompassed in the set with CRC survival outcomes are generally stronger than genes outside the set by harnessing p-values from the gene based tests above (de Leeuw *et al*, 2015). To be specific, the method first creates a statistic Z from the gene based p-values by:

$$\mathrm{Z} = \Phi^{-1}(1 - p)$$

Where it takes the inverse cumulative standard normal distribution of 1-p. Then a linear regression model is fitted as:

$$\mathrm{Z} = \beta_{0S}L + S_s\beta_s + \epsilon$$

Where β_0L is the intercept, and S_s denotes whether the gene is included in the set s (S_s=1 or 0). A one-sided test is then conducted to determine if β_s is significantly larger than 0. In order to adjust for gene-gene correlations, the competitive model derives a gene-gene correlation matrix from correlations of genetic variants in each pairs of genes (de Leeuw *et al*, 2015). With respect to evaluating statistical significance, the FUMA implements a Bonferroni correction to adjust the α level to 0.05/n where n denotes the number genes and gene sets.

4.4 Summary

This chapter describes the study cohorts and presents details of the study design and statistical methods implemented in the thesis. The SOCCS cohort was used as the main discovery dataset for the genetic association studies and the GWAS, whilst the UK Biobank cohort along with three clinical trial datasets were employed to replicate possible discoveries. The Cox regression model was used to estimate effect of each genetic variant on overall and CRC-specific survival outcomes of CRC patients. For the predictive modelling, I used the UK Biobank cohort to develop the genetic predictor and tested its predictive performance as well as added predictive value of the genetic predictor on the bases of other non-genetic factors using the SOCCS cohort. The model performance was evaluated by examining the discriminative ability and the model calibration. The results of each part of analyses are presented in Chapter 5.

Chapter 5 Results

5.1 Introduction

This chapter presents the results of the analysis in response to the objectives listed in Chapter 2. The first section summarises the basic characteristics of the included variables and survival outcomes of patients from the SOCCS and UK Biobank cohorts. Descriptive analysis of the associations between these variables and survival outcomes is also presented. In the second section, the main results of the validation study on previously reported genetic variants associated with CRC survival are described. Next in the third section, I present the prediction models combining these previously reported variants and other variables. Results of model performance in predicting CRC survival are also presented. The fourth section features the results from candidate association studies testing two groups of genetic variants—variants associated with CRC risk and variants reportedly linked with survival outcomes of other types of cancers. Finally, the fifth section presents the results of the genome-wide association study using the SOCCS cohort and replication analysis based on a meta-analysis of the summary statistics from the UK Biobank and the clinical trial datasets.

5.2 Descriptive analysis of the study cohorts

This section presents basic characteristics of the participants included in the analysis. Detailed information regarding the definition and obtainment of study variables can be found in Chapter 4. To be specific, the results of patient selection are presented firstly. Then characteristics of covariates including missing data are reported. Genetic association analyses in this thesis were conducted on the basis of complete case analysis; therefore, no missing data were present in covariates used in these analyses including age at diagnosis, sex and AJCC stage. Missing values were only identified in other variables (tumour grade and site). Finally, descriptive analysis of survival outcomes of the study cohorts and associations between the covariates and survival outcomes are presented in this section.

5.2.1 Patient selection

Study of Colorectal Cancer in Scotland

As mentioned previously, the survival time of each CRC patient was calculated from the date of definitive treatment to the date of death (or alive until January 1st 2018). Therefore, I was unable to construct the survival time if the date of definitive treatment was missing; these patients were excluded from the analysis (N=46). I also excluded 527 patients with missing covariates including age at diagnosis, sex and AJCC stage. In addition, ten patients with stage 0 designation were also excluded because they could be a mixture of patients with polyps or with pathologic complete response (no residual tumour identified) to neoadjuvant therapy. I further excluded 94 cases diagnosed with appendix and endocrine tumour. Presented in **Figure 5-1** is the diagram of patient selection for the SOCCS study. A total of 5,675 CRC patients were eligible for the final analysis.

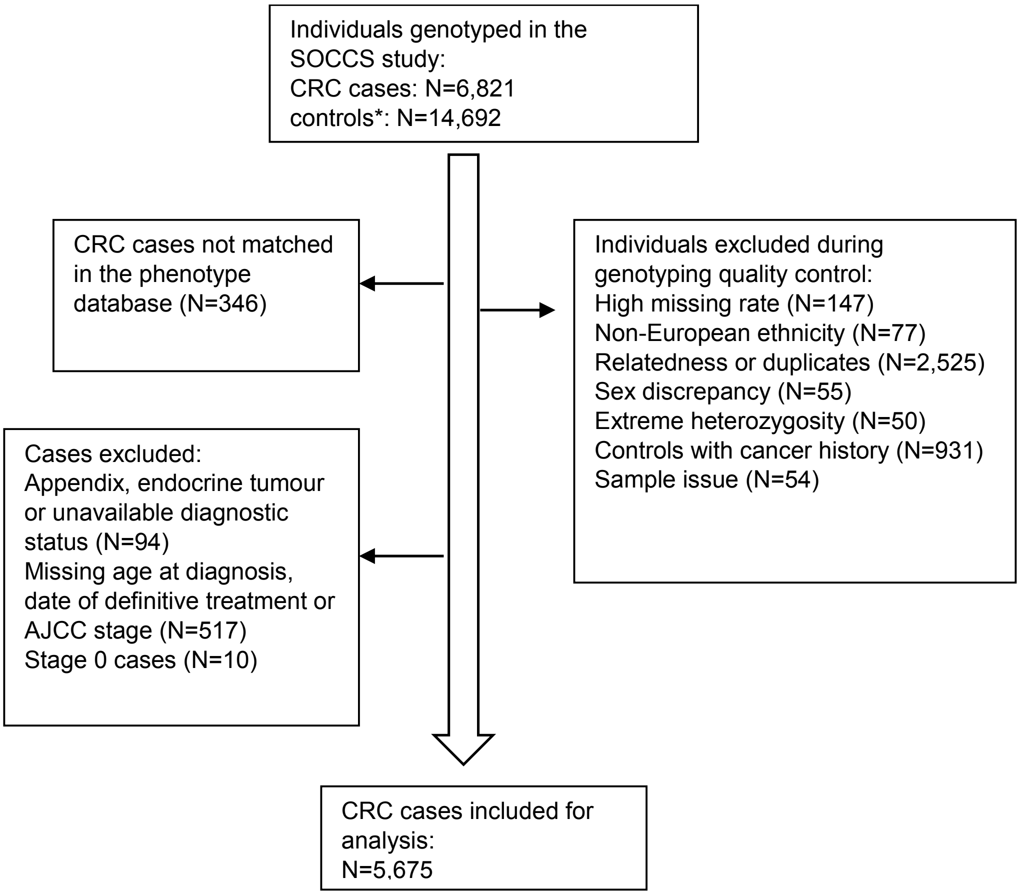

Individuals genotyped in the SOCCS study:
CRC cases: N=6,821
controls*: N=14,692
CRC cases not matched in the phenotype database (N=346)
Individuals excluded during genotyping quality control:
High missing rate (N=147)
Non-European ethnicity (N=77)
Relatedness or duplicates (N=2,525)
Sex discrepancy (N=55)
Extreme heterozygosity (N=50)
Controls with cancer history (N=931)
Sample issue (N=54)
Cases excluded:
Appendix, endocrine tumour or unavailable diagnostic status (N=94)
Missing age at diagnosis, date of definitive treatment or AJCC stage (N=517)
Stage 0 cases (N=10)
CRC cases included for analysis:
N=5,675

Figure 5-1 Diagram of patient selection for the Study of Colorectal Cancer in Scotland. Adapted from the previous publication (He *et al*, 2019b) with permission. The original source is an open access article under the terms of the Creative Commons Attribution License.

*controls were only used for QC, not included in analysis of this thesis

UK Biobank

Given that disease characteristics such as AJCC stage were still unavailable in the UK Biobank cohort, no disease-based quality control, for example exclusion of stage 0 patients, was conducted. Therefore, I included all CRC cases with genotype data that passed genotyping quality control from the case-control study (Project 7411). Twelve patients with missing information on age at diagnosis were excluded. **Figure 5-2** shows the diagram of patient selection for the UK Biobank study dataset. Finally, a total of 4,887 CRC patients were included.

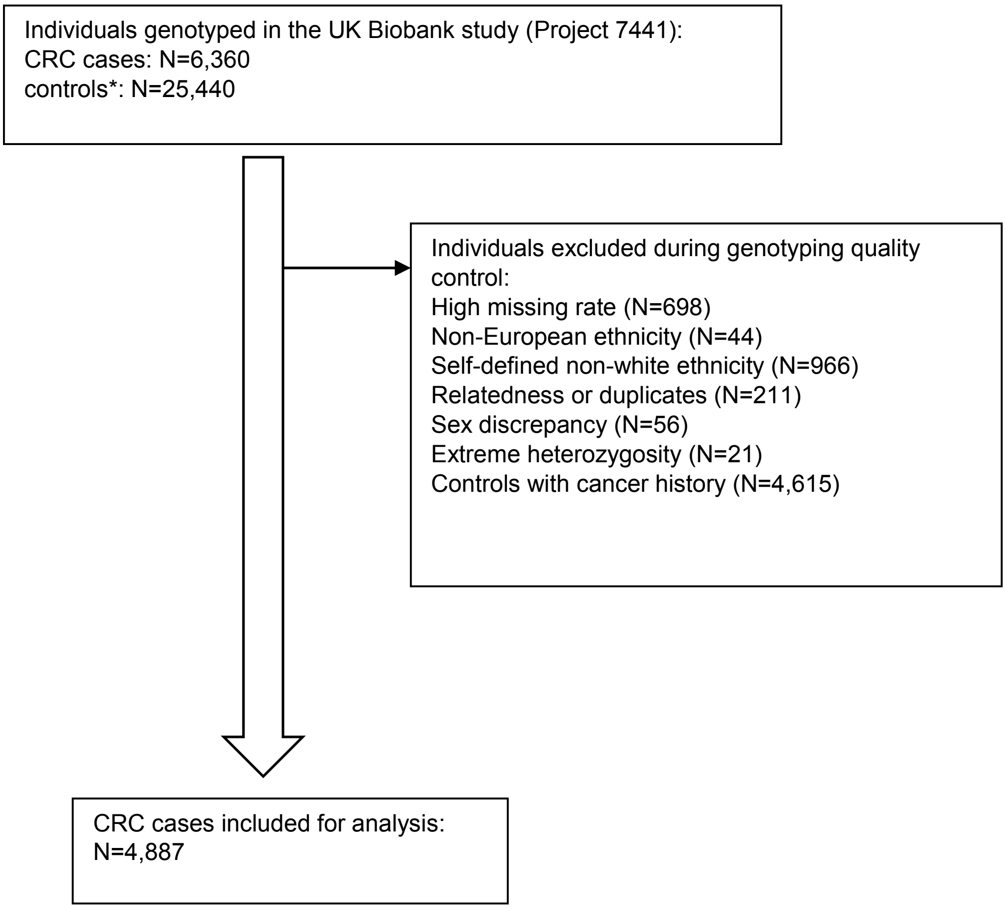

Individuals genotyped in the UK Biobank study (Project 7441):
CRC cases: N=6,360
controls*: N=25,440
Individuals excluded during genotyping quality control:
High missing rate (N=698)
Non-European ethnicity (N=44)
Self-defined non-white ethnicity (N=966)
Relatedness or duplicates (N=211)
Sex discrepancy (N=56)
Extreme heterozygosity (N=21)
Controls with cancer history (N=4,615)
CRC cases included for analysis:
N=4,887

Figure 5-2 Patient selection procedure from the UK Biobank study

*controls were only used for QC, not included in analysis of this thesis

5.2.2 Covariates

As described in Chapter 4, the SOCCS cohort included a total of 5,675 patients diagnosed with CRC and the UK Biobank cohort included 4,887 patients (incident and prevalent CRC cases). Based on the recommendations from the Canadian Cancer Society on prognostic factors of CRC (introduced in Chapter 1), the AJCC stage and the tumour grade were available in the SOCCS cohort and therefore they were extracted along with other basic characteristics including age at CRC diagnosis, sex and tumour site (colon or rectum). Given that enhanced cancer data had not been released by the UK Biobank, only age at diagnosis and sex were included as covariates for the study cohort in this thesis. Descriptive statistics along with the number of missing values of these variables are summarised in **Table 5-1**. Categorical or binary variables are presented as exact numbers and percentages; continuous variables are presented as median and interquartile range (IQR).

Table 5-1 Basic characteristics and descriptive statistics of eligible non-genetic variables from the SOCCS (n=5,675) and UK Biobank study (n=4,887)

Variables*	SOCCS	Missing	UK Biobank	Missing
Age at diagnosis (years)	64.5(54.6-71.6)	0	62.2(55.8-67.2)	0
Sex		0		0
Male	3,235(57%)		2,063(42%)	
Female	2,440(43%)		2,824(58%)	
AJCC stage		0		NA
I	1,005(17.7%)		NA	
II	1,891(33.3%)			
III	1,995(35.2%)			
IV	784(13.8%)			
Tumour grade		839(14.8%)		NA
1 (well)	185(3.3%)		NA	
2 (moderate)	3,954(69.7%)			
3 (poor/undifferentiated)	697(12.3%)			
Tumour site		66(1.2%)		NA
Colon	3,392(59.8%)		NA	
Rectum	2,201(38.8)			
Colon and rectum	16(0.3%)			

*Continuous variables are presented with median and interquartile range.
AJCC, the American Joint Committee on Cancer; NA, not available.

5.2.3 Survival outcomes

Kaplan-Meier estimates—probabilities of CRC patients in SOCCS and UK Biobank cohorts surviving over 1, 3, 5 and 10 years—were calculated and listed in **Table 5-2**.

The survival rates of the UK Biobank cohort (N=4,887) were mostly higher than rates of the SOCCS cohort (N=5,675) at each observation time. Results of log-rank test also suggested a statistically significant difference in terms of the survival estimates between the two study cohorts (p<0.001 for both overall and CRC-specific survival).

Table 5-2 Kaplan-Meier estimates for overall and CRC-specific survival of patients in the SOCCS and UK Biobank cohort

	Kaplan-Meier estimates (95%CI)	
Time (years)	SOCCS	UK Biobank
Overall survival		
1	0.958(0.953-0.963)	0.957(0.951-0.962)
3	0.824(0.814-0.834)	0.886(0.877-0.895)
5	0.738(0.726-0.751)	0.838(0.828-0.849)
10	0.610(0.595-0.625)	0.766(0.754-0.780)
CRC-specific survival		
1	0.963(0.958-0.968)	0.964(0.958-0.969)
3	0.845(0.835-0.855)	0.908(0.900-0.916)
5	0.775(0.763-0.787)	0.872(0.863-0.882)
10	0.705(0.691-0.719)	0.828(0.817-0.840)

CI, confidence interval; CRC, colorectal cancer

Notably, the UK Biobank cohort consisted of 2,474 incident and 2,080 prevalent CRC cases. It is widely accepted that prevalent cases suffer from potential selection bias—patients with less severe disease are more likely to be selected into the study, leading to the appearance of higher observed survival rates than expected. Therefore, I plotted survival curves of Kaplan-Meier estimates of the incident and prevalent CRC cases in the UK Biobank separately along with survival estimates of patients in the SOCCS cohort to visualise this potential difference caused by the inclusion of prevalent cases in the UK Biobank. The outcome of overall survival is presented in **Figure 5-3** and CRC-specific survival is shown in **Figure 5-4**. The occurrence of censorings as well as the number of patients at risk for each observation time is also presented in the figures.

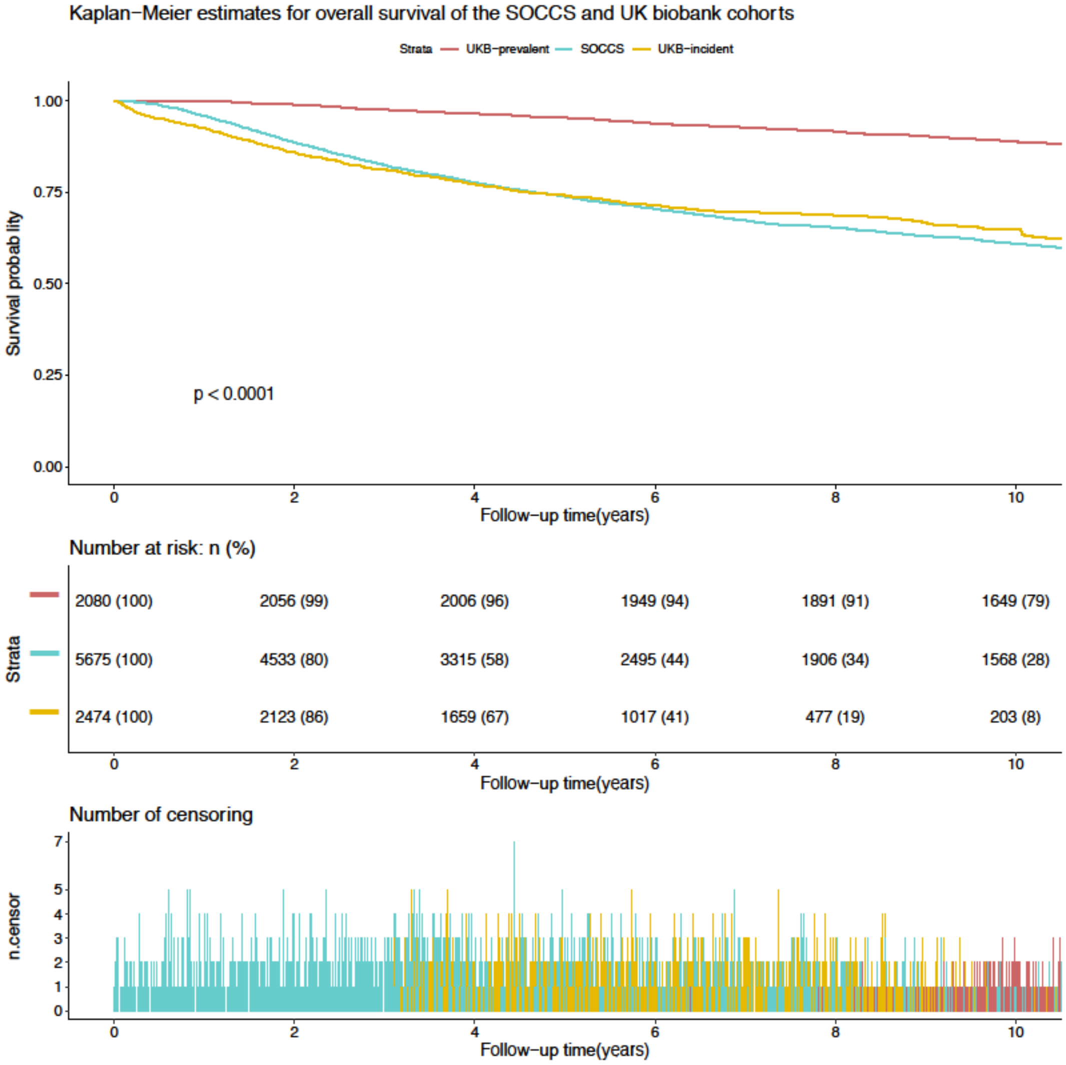
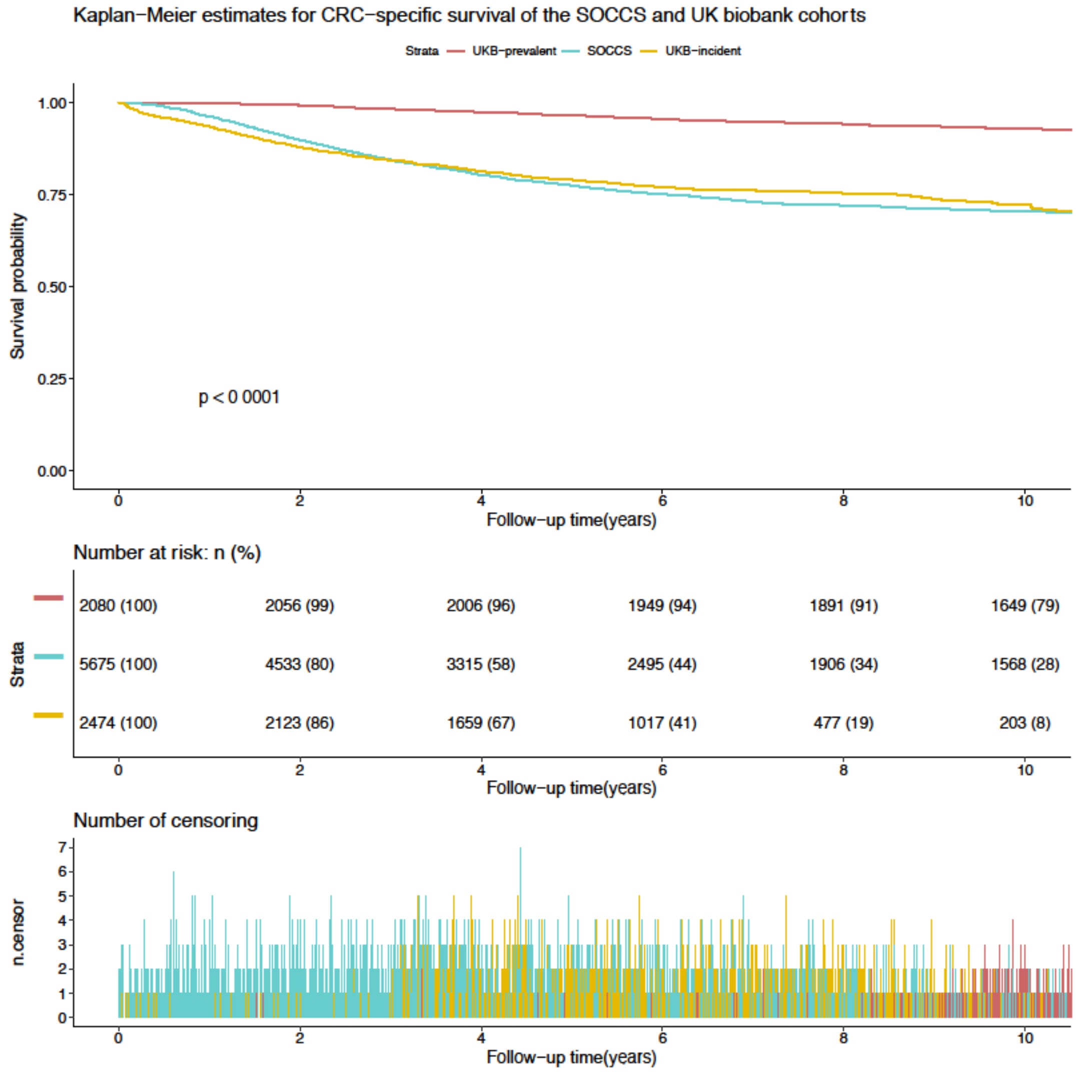


Figure 5-3 Kaplan-Meier estimates for overall survival of patients from the SOCCS and the UK Biobank study (incident and prevalent CRC cases)



Strata	0	2	4	6	8	10
🟥	2080 (100)	2056 (99)	2006 (96)	1949 (94)	1891 (91)	1649 (79)
🟦	5675 (100)	4533 (80)	3315 (58)	2495 (44)	1906 (34)	1568 (28)
🟨	2474 (100)	2123 (86)	1659 (67)	1017 (41)	477 (19)	203 (8)

Figure 5-4 Kaplan-Meier estimates for CRC-specific survival of patients from the SOCCS and the UK Biobank study (incident and prevalent CRC cases)

As expected, prevalent CRC cases in the UK Biobank had significantly favoured overall and CRC-specific survival outcomes compared with incident cases in the UK Biobank and patients in the SOCCS study (log-rank p<0.001). There were more censorings in the prevalent cases enriched in later years after diagnosis compared with the other two groups of patients. The survival curves also indicated comparable survival estimates between the SOCCS cohort and incident cases in the UK Biobank. Separate log-rank tests were conducted to compare survival estimates of these two groups of patients and no significant difference was detected for either overall (p=0.50) or CRC-specific survival (p=0.55). The salient divergence between survival estimates of prevalent CRC cases and incident cases in the UK Biobank hinted to selection bias present amongst the prevalent CRC cases. In order to eliminate this bias, subsequent survival analysis of the UK Biobank was only conducted using incident CRC cases.

After excluding prevalent cases, the median follow-up time of the UK Biobank cohort was 5.3 years (quartiles: 3.5-7.5). There had been 765 deaths (30.9%) of any causes until the censoring time; among them, 587 patients (76.7% of all deaths) died from CRC-related causes. With regard to the SOCCS study, the median follow-up time of the SOCCS cohort was 5.1 years (IQR: 2.4-11.4). During the follow-up time, an aggregate of 1,918(33.8%) patients died, and 1,358 (70.8% of all deaths) of them died from CRC-related causes.

5.2.4 Associations between covariates and CRC survival

Associations between these aforementioned variables and survival outcomes of CRC in both cohorts were examined by fitting univariable Cox models. **Table 5-3** presents detailed effect estimates and p-values for each association analysis. No stringent correction for multiple testing was applied to evaluate statistical significance at this point. In particular, male CRC patients showed significantly favoured overall and CRC-specific survival outcomes, and this association was consistent across the two study cohorts. With respect to age at diagnosis, significant association between increased age and poorer survival was only observed for overall survival of the SOCCS cohort. However, younger age at diagnosis was significantly associated with inferior CRC-specific survival for both cohorts.

Table 5-3 Summary of associations between covariates and survival outcomes of CRC using univariable Cox regression

	SOCCS(n=5,675)		**UK Biobank (n=2,474)**	
Variables	HR(95%CI)	p	HR(95%CI)	p
Overall survival				
Age at diagnosis	1.015(1.011-1.019)	5.33E-12	0.995(0.984-1.006)	0.342
Sex (Male vs Female)	0.864(0.788-0.946)	0.002	0.733(0.632-0.850)	4.0E-5

	SOCCS(n=5,675)		UK Biobank (n=2,474)	
AJCC stage	2.194(2.075-2.320)	3.3E-168	NA	
Tumour grade	1.334(1.186-1.502)	1.74E-6	NA	
Tumour site(rectum vs. colon)	1.009(0.920-1.106)	0.850	NA	
CRC-specific survival				
Age at diagnosis	0.992(0.988-0.997)	0.002	0.984(0.972-0.996)	0.009
Sex (Male vs Female)	0.894(0.802-0.997)	0.043	0.807(0.682-0.953)	0.012
AJCC stage	3.240(3.018-3.477)	1.60E-232	NA	
Tumour grade	1.630(1.421-1.870)	3.11E-12	NA	
Tumour site(Rectum vs. Colon)	1.026(0.920-1.144)	0.647	NA	

HR, hazard ratio; CI, confidence interval; BMI, body mass index; CRP, C-reaction protein; AJCC, the American Joint Committee on Cancer; NA, not available.

As for the pathological variables (tumour stage and grade), analysis was only conducted in the SOCCS cohort due to data availability. In concordance with previous evidence, more advanced tumour stage and grade were both strongly associated with inferior survival outcomes of CRC patients (p<0.001) in the SOCCS dataset. Kaplan-Meier survival curves stratified by AJCC stage and tumour grade are plotted in **Figure 5-5** and **Figure 5-6** respectively. As shown in **Figure 5-5**, the 5-year overall survival rates for stage I to IV CRC patients in the SOCCS study were 93%, 84%, 72% and 27%. The 5-year CRC-specific survival rates were 97%, 89%, 75% and 28% for stage I to IV patients respectively. In respect to tumour grade, patients diagnosed with grade 1 to 3 tumours had 5-year overall survival rates of 81%, 76% and 67%. The 5-year CRC-specific survival rates were 88% for grade 1, 80% for grade 2 and 70% for grade 3 patients.

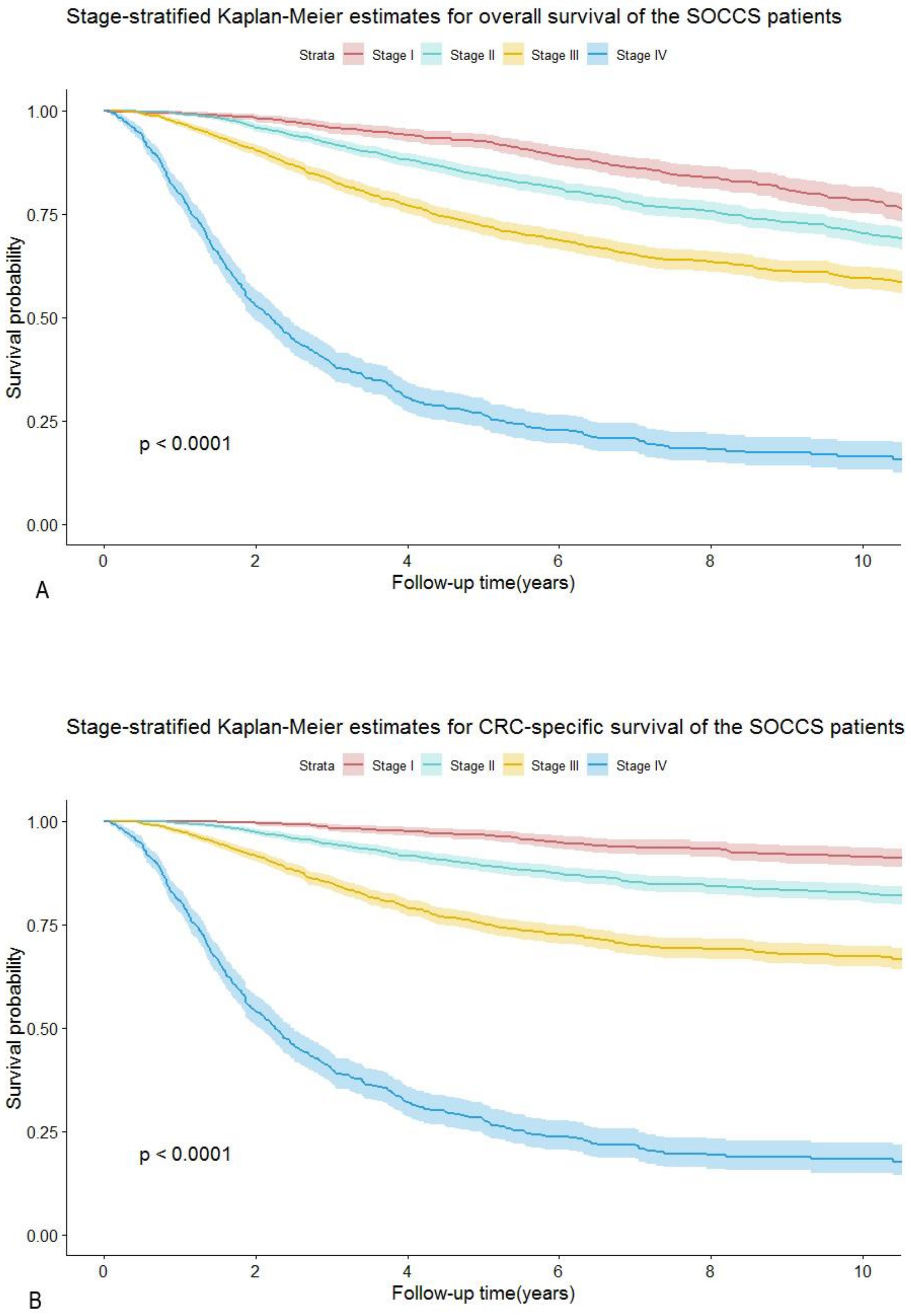


Figure 5-5 Stage-stratified Kaplan-Meier estimates of CRC patients in the SOCCS study. (A for overall survival and B for CRC-specific survival).

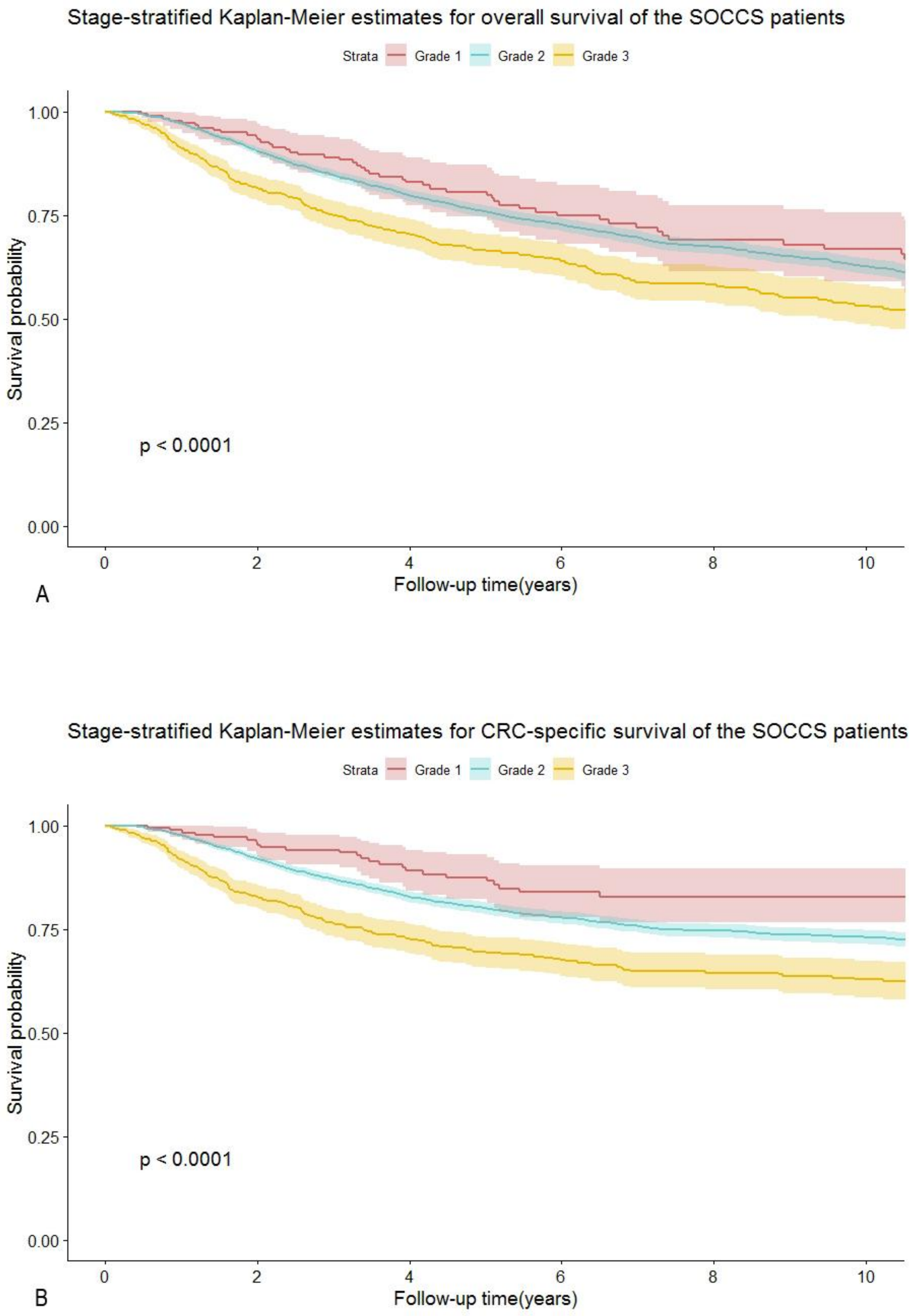


Figure 5-6 Grade-stratified Kaplan-Meier estimates of CRC patients in the SOCCS study. (A for overall survival and B for CRC-specific survival).

5.3 Candidate association studies

5.3.1 Validation of genetic variants previously linked with CRC survival

Eligible variants

The initial search in the GWAS catalogue yielded three GWASs reporting significant associations ($p<10^{-5}$) between 51 autosomal genetic variants and CRC survival outcomes (Pander *et al*, 2015; Phipps *et al*, 2016; Xu *et al*, 2015). Eight variants were excluded due to linkage disequilibrium with other identified variants ($r^2>0.2$). Eventually, a total of 43 variants were included in the analysis, and their basic characteristics are listed in **Table 5-4**. Among the included 43 variants, two variants—rs209489 and rs885036—were reported to be associated with CRC survival at GWAS significance level ($p<5x10^{-8}$) (Pander *et al*, 2015) (Phipps *et al*, 2016). These two variants were identified in a subgroup of metastatic CRC patients. The variant rs209489 was associated with disease-free survival whereas rs885036 was associated with progression-free survival. With respect to different survival outcomes of CRC patients, 24 out of the 43 variants were reported to be associated with disease-free survival; 15 variants with overall survival; and the remaining three variants with progression-free survival of CRC patients. Xu et al. conducted stratified GWAS analysis by tumour site (Xu *et al*, 2015), 10 variants were identified specifically in colon cancer patients, and 17 variants were associated with survival outcomes of rectal cancer patients (**Table 5-4**). Additional details of the effect estimates for each listed variant can be found in the GWAS catalogue.

Table 5-4 Summary details of the included genetic variants previously associated with CRC survival

Variant	locus	MA	MAF	Gene	Reported outcomes	Reference
rs10921219	1q31.2	A	0.46	*AL390957.1*	Colon cancer(OS)	Xu,2015
rs6720296	2p21	C	0.48	*LINC01121*	Colorectal cancer (MSI-L/S) (DFS)	Xu,2015
rs885036	2q11.2	A	0.45	*MGAT4A*	Colorectal cancer(metastatic)(PFS)	Pander,2015
rs17048372	2q14.1	T	0.15	*DPP10*	Colon cancer(OS)	Xu,2015
rs4377367	2q21.1	C	0.19	*ARHGEF4*	Colorectal cancer(metastatic)(PFS)	Pander,2015

Variant	locus	MA	MAF	Gene	Reported outcomes	Reference
rs16867335	2q31.3	T	0.2	*AC009478.1*	Rectal cancer(OS)	Xu,2015
rs6854845	4q13.3	T	0.12	*Intergenic*	Rectal cancer(OS)	Xu,2015
rs17026425	4q31.23	A	0.08	*IQCM*	Rectal cancer(OS)	Xu,2015
rs13180087	5p12	C	0.18	*HCN1*	Colon cancer(OS)	Xu,2015
rs10040610	5p15.31	C	0.21	*Intergenic*	Colorectal cancer (MSI-L/S) (OS)	Xu,2015
rs157411	5q13.1	G	0.33	*Intergenic*	Rectal cancer(OS)	Xu,2015
rs1493383	5q33.2	T	0.1	*GRIA1*	Colorectal cancer (MSI-L/S) (OS)	Xu,2015
rs17057166	5q33.3	T	0.09	*LINC01847*	Rectal cancer(DFS)	Xu,2015
rs12187751	5q34	G	0.09	*AC113414.1*	Colorectal cancer (MSI-L/S) (OS)	Xu,2015
rs4868304	5q35.2	T	0.14	*LINC01484*	Rectal cancer(DFS)	Xu,2015
rs209489	6p12.1	C	0.08	*ELOVL5*	Colorectal cancer (metastastic) (DFS)	Phipps,2016
rs2073016	6p21.1	C	0.18	*APOBEC2*	Colorectal cancer(metastatic)(PFS)	Pander,2015
rs1573948	6p25.1	C	0.1	*Intergenic*	Rectal cancer(OS)	Xu,2015
rs4959799	6p25.2	C	0.14	*SLC22A23*	Rectal cancer(DFS)	Xu,2015
rs17087282	6q25.3	A	0.11	*Intergenic*	Colorectal cancer (MSI-L/S) (OS)	Xu,2015
rs10275272	7p21.1	T	0.28	*TWIST1*	Rectal cancer(DFS)	Xu,2015
rs2936519	8p23.1	A	0.27	*Intergenic*	Colorectal cancer(metastatic)(PFS)	Pander,2015
rs7004484	8q24.22	C	0.3	*EFR3A*	Rectal cancer(OS)	Xu,2015
rs1998584	9p23	T	0.41	*Intergenic*	Colorectal cancer (MSI-L/S) (OS)	Xu,2015
rs11138220	9q21.31	G	0.13	*Intergenic*	Rectal cancer(DFS)	Xu,2015
rs1407508	9q22.33	C	0.02	*AL136084.1*	Colorectal cancer (MSI-L/S) (DFS)	Xu,2015
rs1555895	10p15.3	A	0.5	*Intergenic*	Rectal cancer(OS)	Xu,2015
rs1570271	10q25.3	A	0.14	*Intergenic*	Rectal cancer(DFS)	Xu,2015
rs9419702	10q26.3	C	0.26	*Intergenic*	Rectal cancer(DFS)	Xu,2015
rs12224794	11p12	A	0.42	*LRRC4C*	Colorectal cancer(OS)	Phipps,2016
rs3781663	11q13.3	G	0.31	*ANO1*	Rectal cancer(DFS)	Xu,2015
rs912294	13q12.3	A	0.47	*Intergenic*	Colorectal cancer (MSI-L/S) (DFS)	Xu,2015
rs17280262	14q32.2	T	0.04	*Intergenic*	Colon cancer(DFS)	Xu,2015
rs1075232	15q13.3	A	0.1	*AC104759.1*	Colorectal cancer (non-metastatic)(OS)	Phipps,2016
rs8035094	15q13.3	C	0.1	*Intergenic*	Colon cancer(DFS)	Xu,2015
rs10152207	15q14	A	0.09	*Intergenic*	Rectal cancer(OS)	Xu,2015
rs338389	15q23	G	0.49	*Intergenic*	Rectal cancer(OS)	Xu,2015
rs3794924	18q12.1	A	0.1	*DSG3*	Colon cancer(OS)	Xu,2015
rs1372474	18q21.2	G	0.07	*LINC01919*	Colorectal cancer(metastatic)(OS)	Phipps,2016
rs6105057	20p12.1	G	0.28	*ISM1*	Colon cancer(OS)	Xu,2015
rs658495	20p13	G	0.04	*C20orf27*	Colon cancer(DFS)	Xu,2015
rs4812219	20q13.33	T	0.14	*Intergenic*	Colon cancer(OS)	Xu,2015

Variant	locus	MA	MAF	Gene	Reported outcomes	Reference
rs139156	22q13.31	A	0.26	*PARVG*	Colon cancer(OS)	Xu,2015

CRC, colorectal cancer; CC, colon cancer; RC, rectal cancer; mCRC, metastatic colorectal cancer; OS, overall survival; DFS, disease-free survival; PFS, progression-free survival. MSI-L/S, microsatellite instability-low/stable

Statistical power

Given that there were 43 independent candidate variants to be investigated, a Bonferroni corrected α level of 0.001 was employed to estimate the statistical power for this study. As shown in **Table 5-4**, the minor allele frequency (MAF) of included genetic variants ranged from 0.02 to 0.50; therefore, a range of MAFs starting from 0.01 to 0.50 was used for power estimation. In combination with the sample size (n=5,675) and the number of events in the SOCCS cohort—34% for overall survival and 24% for CRC-specific survival, statistical power was estimated based on a range of potential genetic effect sizes (hazard ratio) from 1.05 to 1.60 under the additive genetic model. Using the formula provided by Owzar et al. (Owzar *et al*, 2012), this study had a power of 85% to detect a hazard ratio of 1.25 on overall survival for 72% (31/43) of included variants (MAF> 0.15), and the power for CRC-specific survival was 66%. Power curves at different levels of hazard ratios and MAFs are plotted in **Figure 5-7**. As shown, this study had limited power (<0.5) to detect an effect as large as 1.6 for the variant with the lowest MAF of 0.01.

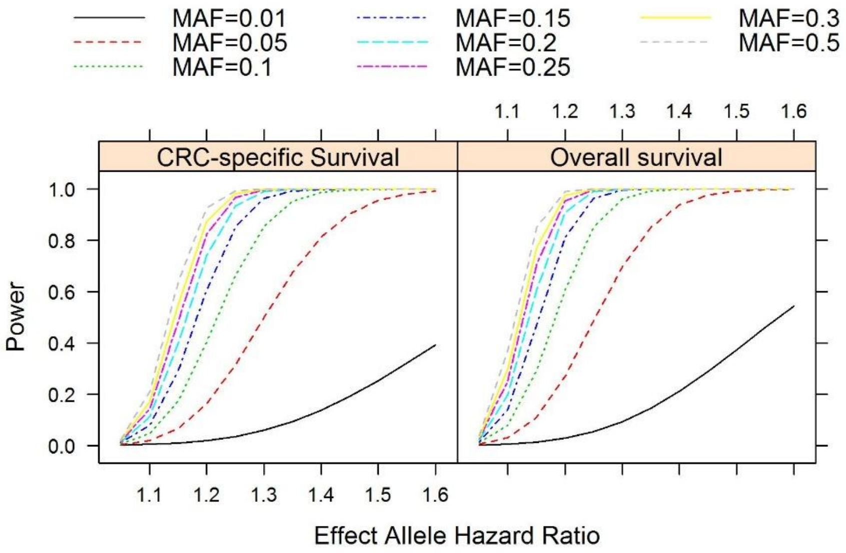


Figure 5-7 Power curves for the SOCCS validation study of genetic variants previously linked with CRC survival

Main results

I fitted Cox regression models adjusting for age at diagnosis, sex and AJCC stage to investigate associations between each of the 43 variants and survival outcomes of CRC. After correcting for multiple testing using the FDR approach, no genetic variants were statistically significantly associated with overall survival (FDR corrected p-value<0.05). With regard to results not corrected for multiple testing, four variants (rs11138220, rs17026425, rs6854845, rs17057166) were associated with overall survival at p<0.05. In particular, three variants (rs17026425, rs6854845, rs17057166) showed the same direction of effects as the previous GWAS—the minor alleles of these three variants conferred higher overall death hazard. However, an opposite effect for the variant rs11138220 was found when compared to the original GWAS (Xu *et al*, 2015). Our results suggested a favourable overall survival for the minor allele (G) (HR=0.88, 95%CI=0.79-0.98, uncorrected p=0.016), whereas Xu et al. reported that the G allele was associated with worse disease-free survival for rectal cancer patients (HR=2.76, 95%CI=1.77-4.31, p=8.0x10^{-6}). The effect estimates (HR) along with both uncorrected and FDR-corrected p-values of these 43 variants on overall survival are presented in **Table 5-5**.

Table 5-5 Summary of associations between 43 variants previously linked with CRC survival and overall survival of CRC patients in the SOCCS study (N=5,675).

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs11138220	G	0.13	*Intergenic*	0.88(0.79-0.98)	0.016	0.462
rs17026425	A	0.08	*IQCM*	1.16(1.01-1.33)	0.039	0.462
rs6854845	T	0.12	*Intergenic*	1.14(1.01-1.29)	0.040	0.462
rs17057166	T	0.09	*LINC01847*	1.14(1.00-1.29)	0.042	0.462
rs17087282	A	0.11	*Intergenic*	1.14(1.00-1.29)	0.054	0.467
rs17280262	T	0.04	*Intergenic*	0.86(0.74-1.01)	0.064	0.467
rs6720296	C	0.48	*LINC01121*	1.05(0.99-1.12)	0.122	0.689
rs912294	A	0.47	*Intergenic*	0.95(0.89-1.01)	0.127	0.689
rs1570271	A	0.14	*Intergenic*	1.08(0.97-1.21)	0.172	0.689
rs139156	A	0.26	*PARVG*	0.92(0.82-1.04)	0.176	0.689
rs16867335	T	0.20	*AC009478.1*	0.95(0.87-1.03)	0.183	0.689
rs2073016	C	0.18	*APOBEC2*	1.05(0.97-1.14)	0.189	0.689
rs1372474	G	0.07	*LINC01919*	1.08(0.96-1.21)	0.212	0.689
rs3794924	A	0.10	*DSG3*	0.94(0.85-1.04)	0.219	0.689
rs13180087	C	0.18	*HCN1*	1.06(0.96-1.16)	0.271	0.751
rs1407508	C	0.02	*AL136084.1*	0.93(0.82-1.07)	0.313	0.751
rs10040610	C	0.21	*Intergenic*	1.05(0.95-1.16)	0.347	0.751
rs3781663	G	0.31	*ANO1*	1.03(0.96-1.11)	0.365	0.751
rs4377367	C	0.19	*ARHGEF4*	0.96(0.89-1.05)	0.373	0.751
rs2936519	A	0.27	*Intergenic*	0.95(0.86-1.06)	0.381	0.751
rs12224794	A	0.42	*LRRC4C*	0.97(0.91-1.04)	0.387	0.751
rs4959799	C	0.14	*SLC22A23*	0.94(0.82-1.08)	0.395	0.751
rs17048372	T	0.15	*DPP10*	0.97(0.89-1.05)	0.422	0.751
rs10152207	A	0.09	*Intergenic*	1.04(0.94-1.14)	0.428	0.751
rs1573948	C	0.10	*Intergenic*	1.04(0.94-1.15)	0.441	0.751
rs885036	A	0.45	*MGAT4A*	0.98(0.91-1.04)	0.444	0.751
rs4812219	T	0.14	*Intergenic*	1.04(0.92-1.17)	0.554	0.899
rs1075232	A	0.10	*AC104759.1*	1.04(0.90-1.20)	0.572	0.899
rs6105057	G	0.28	*ISM1*	0.98(0.91-1.06)	0.602	0.913
rs10921219	A	0.46	*AL390957.1*	0.98(0.92-1.05)	0.637	0.925
rs338389	G	0.49	*Intergenic*	0.99(0.93-1.05)	0.682	0.925
rs1493383	T	0.10	*GRIA1*	0.98(0.90-1.07)	0.689	0.925
rs658495	G	0.04	*C20orf27*	1.03(0.87-1.22)	0.714	0.925
rs1555895	A	0.50	*Intergenic*	1.01(0.95-1.08)	0.715	0.925
rs9419702	C	0.26	*Intergenic*	1.01(0.94-1.09)	0.759	0.945
rs209489	C	0.08	*ELOVL5*	0.98(0.87-1.11)	0.773	0.945
rs12187751	G	0.09	*AC113414.1*	1.01(0.90-1.14)	0.826	0.956
rs4868304	T	0.14	*LINC01484*	1.01(0.92-1.10)	0.866	0.968

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs10275272	T	0.28	*TWIST1*	1.01(0.91-1.12)	0.880	0.968
rs7004484	C	0.30	*EFR3A*	1.00(0.93-1.09)	0.942	0.979
rs157411	G	0.33	*Intergenic*	1.00(0.93-1.07)	0.944	0.979
rs1998584	T	0.41	*Intergenic*	1.00(0.94-1.07)	0.960	0.979
rs8035094	C	0.10	*Intergenic*	1.00(0.88-1.13)	0.979	0.979

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values corrected using the false positive rate approach.

With respect to results of CRC-specific survival, again no statistically significant associations between any of the 43 variants and CRC-specific survival were identified in the SOCCS cohort after correcting for multiple testing using the FDR approach. The only variant with a p-value<0.05 prior to FDR-correction was rs11138220. Similar to its effect on overall survival, a potential favourable effect on CRC-specific survival was identified to be linked with the minor allele G (HR=0.85, 95%CI=0.75-0.97, uncorrected p=0.016), which was discordant with the original finding. Detailed effect estimates and p-values can be found in **Table 5-6**:

Table 5-6 Summary of associations between 43 variants previously linked with CRC survival and CRC-specific survival of CRC patients in the SOCCS study (N=5,675).

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs11138220	G	0.13	*Intergenic*	0.85(0.75-0.97)	0.016	0.700
rs17280262	T	0.04	*Intergenic*	0.84(0.70-1.01)	0.069	0.767
rs1372474	G	0.07	*LINC01919*	1.13(0.99-1.29)	0.078	0.767
rs139156	A	0.26	*PARVG*	0.89(0.77-1.02)	0.087	0.767
rs2073016	C	0.18	*APOBEC2*	1.09(0.99-1.19)	0.087	0.767
rs16867335	T	0.20	*AC009478.1*	0.92(0.84-1.02)	0.112	0.771
rs6720296	C	0.48	*LINC01121*	1.05(0.98-1.14)	0.187	0.771
rs3781663	G	0.31	*ANO1*	1.05(0.96-1.14)	0.276	0.771
rs12187751	G	0.09	*AC113414.1*	1.08(0.94-1.23)	0.289	0.771
rs17087282	A	0.11	*Intergenic*	1.09(0.93-1.27)	0.290	0.771
rs912294	A	0.47	*Intergenic*	0.96(0.89-1.04)	0.299	0.771
rs6854845	T	0.12	*Intergenic*	1.08(0.93-1.26)	0.303	0.771
rs12224794	A	0.42	*LRRC4C*	0.96(0.89-1.04)	0.308	0.771

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs1570271	A	0.14	*Intergenic*	1.06(0.93-1.21)	0.347	0.771
rs10040610	C	0.21	*Intergenic*	1.06(0.94-1.19)	0.349	0.771
rs6105057	G	0.28	*ISM1*	0.96(0.87-1.05)	0.379	0.771
rs885036	A	0.45	*MGAT4A*	0.97(0.90-1.04)	0.384	0.771
rs1407508	C	0.02	*AL136084.1*	0.93(0.79-1.09)	0.389	0.771
rs10275272	T	0.28	*TWIST1*	1.05(0.93-1.19)	0.393	0.771
rs1573948	C	0.10	*Intergenic*	1.05(0.94-1.18)	0.394	0.771
rs17057166	T	0.09	*LINC01847*	1.07(0.92-1.24)	0.395	0.771
rs17048372	T	0.15	*DPP10*	0.96(0.88-1.06)	0.469	0.771
rs209489	C	0.08	*ELOVL5*	0.95(0.83-1.09)	0.491	0.771
rs2936519	A	0.27	*Intergenic*	0.96(0.84-1.09)	0.494	0.771
rs10152207	A	0.09	*Intergenic*	1.04(0.93-1.16)	0.494	0.771
rs157411	G	0.33	*Intergenic*	0.97(0.90-1.05)	0.494	0.771
rs338389	G	0.49	*Intergenic*	0.97(0.90-1.05)	0.495	0.771
rs13180087	C	0.18	*HCN1*	1.04(0.93-1.17)	0.502	0.771
rs4377367	C	0.19	*ARHGEF4*	0.97(0.88-1.07)	0.508	0.771
rs4959799	C	0.14	*SLC22A23*	0.95(0.81-1.12)	0.540	0.772
rs10921219	A	0.46	*AL390957.1*	0.98(0.90-1.06)	0.544	0.772
rs8035094	C	0.10	*Intergenic*	0.97(0.84-1.11)	0.634	0.849
rs1493383	T	0.10	*GRIA1*	0.98(0.88-1.08)	0.640	0.849
rs4812219	T	0.14	*Intergenic*	1.03(0.90-1.18)	0.656	0.849
rs1555895	A	0.50	*Intergenic*	1.01(0.94-1.09)	0.776	0.947
rs3794924	A	0.10	*DSG3*	0.99(0.87-1.11)	0.808	0.947
rs658495	G	0.04	*C20orf27*	0.98(0.81-1.18)	0.808	0.947
rs9419702	C	0.26	*Intergenic*	0.99(0.91-1.08)	0.818	0.947
rs7004484	C	0.30	*EFR3A*	0.99(0.90-1.09)	0.897	0.964
rs1075232	A	0.10	*AC104759.1*	0.99(0.83-1.17)	0.898	0.964
rs1998584	T	0.41	*Intergenic*	1.00(0.92-1.07)	0.898	0.964

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs17026425	A	0.08	*IQCM*	1.00(0.83-1.19)	0.956	0.984
rs4868304	T	0.14	*LINC01484*	1.00(0.90-1.11)	0.984	0.984

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values corrected using the false positive rate approach.

Then I extracted the reported effect estimates from the original studies and compared them with our results using the SOCCS cohort; the comparison of the four variants that were associated with CRC survival (uncorrected p<0.05) in our study is summarised in **Table 5-7**. Compared with the original reports, our results tended to show smaller effect sizes although different survival outcomes were used. According to the results of power estimation (**Figure 5-7**), this study would have a power of 100% to detect associations of these variants with survival outcomes in the SOCCS cohort assuming the same effect sizes with previous findings.

Table 5-7 Comparison of estimates in original reports and the SOCCS of variants associated with CRC survival (p<0.05 in the SOCCS) (N=5,675)

			Effect estimate	
Variant	**Gene**	**Reported outcomes**	**Reported HRs**	**HRs in SOCCS**
rs17026425	*IQCM*	OS(RC)	5.06(2.67-9.60)	1.16(1.01-1.33)
rs17057166	*LINC01847*	DFS(RC)	5.56(2.91-10.64)	1.14(1.00-1.29)
rs6854845	*Intergenic*	DFS(RC)	4.12(2.34-7.26)	1.14(1.01-1.29)
rs11138220	*Intergenic*	DFS(RC)	2.76(1.77-4.31)	0.88(0.79-0.98)
				0.85(0.75-0.97)*

*CRC-specific survival in the SOCCS. Other estimates are for overall survival.
HR, hazard ratio, DFS, disease-free survival; RC, rectal cancer

In addition to testing individual effects of included variants, I also created a polygenic risk score (PRS) by counting the total number of risk alleles of the 43 variants for each patient in the SOCCS study. The risk allele of each variant that was detrimental for CRC survival was ascertained from the original GWASs. Colorectal cancer patients in the SOCCS study carried an average of 17.1 risk alleles (standard deviation=3.3). The distribution of the number of risk alleles carried by individuals from the SOCCS is plotted in **Figure 5-8**. As shown, the PRS of variants previously linked with CRC survival were approximately normally distributed. A Cox regression model with the same group of covariates was performed, and the result indicated no significant effect on either overall (HR=1.00, 95% CI=0.99-1.02, p=0.775) or CRC-specific survival (HR=1.00, 95% CI=0.98-1.01, p=0.91) of the PRS.

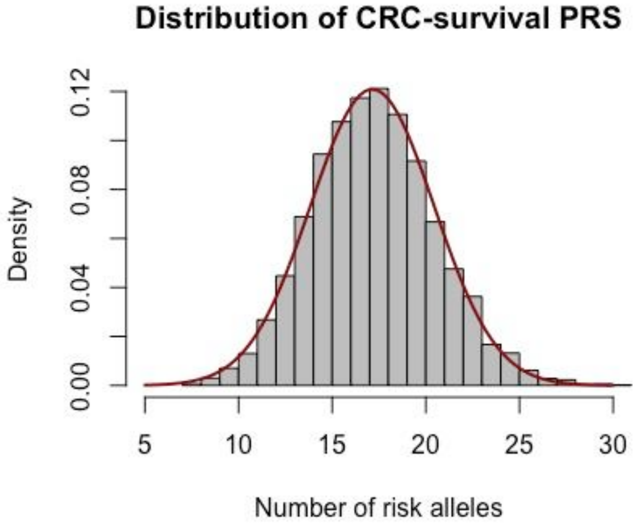


Figure 5-8 Distribution of polygenic risk score of variants associated with CRC survival in the SOCCS study

Stratified analysis

I also conducted genetic association analyses stratified by sex, stage (II/III and IV) and tumour site (colon and rectum) under the additive genetic model. Overall, none of these 43 genetic variants showed significant association with either overall or CRC-specific survival after correction for multiple testing within any strata. Nonetheless, a number of variants were identified to be associated with survival outcomes of CRC at nominal significance (uncorrected p<0.05).

To be specific, in male CRC patients (N=3,225), two variants were associated with overall survival (rs6854845: HR=1.26, 95%CI=1.08-1.48, uncorrected p=0.004, Pfdr=0.086; rs3794924: HR=0.87, 95%CI=0.76-0.99, uncorrected p=0.038, Pfdr=0.497), whereas another variant rs1573948 was observed to be associated with CRC-specific survival in male CRC patients (HR=1.16, 95%CI=1.00-1.33, uncorrected p=0.043, Pfdr=0.710). With respect to female patients (N=2,440) in the SOCCS study, five variants were identified to be associated with CRC survival at p<0.05, and among them, I found associations of the variant rs17280262 with both overall and CRC-specific survival. Effect estimates of these aforementioned variants (p<0.05) are summarised in **Table 5-8**. A full list of 43 variants along with the summarised effects are presented in **Appendix Table S6**.

Table 5-8 Summary of associations (p<0.05) between variants previously linked with CRC survival and outcomes of CRC patients in the SOCCS study stratified by sex

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
Female (N=3,225)						
OS	rs17280262	T	0.04	0.69(0.53-0.89)	0.005	0.237
	rs1998584	T	0.41	0.90(0.81-1.00)	0.044	0.559
	rs1407508	C	0.02	0.80(0.64-1.00)	0.048	0.559
CSS	rs17280262	T	0.04	0.65(0.47-0.89)	0.008	0.309
	rs11138220	G	0.13	0.76(0.61-0.95)	0.014	0.309
	rs3781663	A	0.31	1.14(1.00-1.30)	0.048	0.584
Male (N=2,440)						
OS	rs6854845	T	0.12	1.26(1.08-1.48)	0.004	0.086
	rs3794924	A	0.10	0.87(0.76-0.99)	0.038	0.497
CSS	rs1573948	C	0.10	1.16(1.00-1.33)	0.043	0.710

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

With respect to analyses stratified by stage, a total of 3,886 stage II/III and 784 stage IV CRC patients were included. For stage II/III patients, the A allele of the variant rs17087282 was associated with inferior overall survival (HR=1.18, 95% CI=1.01-1.39, uncorrected p=0.038, Pfdr=0.645). As for CRC-specific survival, a copy of G allele of the variant rs1372474 was associated with 19% higher hazards of death of any causes (HR=1.19, 95% CI=1.01-1.42, uncorrected p=0.041, Pfdr=0.938). However, for stage IV patients, I did not identify any associations with either overall or CRC-specific survival with p-values less than 0.05. The full list of results is presented in **Appendix Table 7.**

In relation to tumour site, a total of 3,392 patients with colon cancer and 2,201 patients with rectal cancer were included in analysis. For colon cancer patients, I found five variants associated with survival outcomes with uncorrected p<0.05 (presented in **Table 5-9**); among them, the variants rs11138220 and rs3794924 were detected in associations with both overall and CRC-specific survival. Regarding rectal cancer patients, six variants were associated with overall survival whereas no variants were related to CRC-specific survival at p<0.05 (**Table 5-9**). Detailed information of all results can be found in **Appendix Table 8**.

Table 5-9 Summary of associations (p<0.05) between variants previously linked with CRC survival and outcomes of CRC patients in the SOCCS study stratified by tumour site.

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
Colon (N=3,392)						
OS	rs3794924	A	0.10	0.82(0.71-0.94)	0.006	0.266
	rs17087282	A	0.11	1.24(1.05-1.46)	0.012	0.266
	rs11138220	G	0.13	0.85(0.74-0.97)	0.018	0.266
	rs16867335	T	0.20	0.90(0.80-1.00)	0.050	0.547
CSS	rs1372474	G	0.07	1.22(1.02-1.45)	0.029	0.531
	rs3794924	A	0.10	0.84(0.71-1.00)	0.046	0.531
	rs11138220	G	0.13	0.84(0.71-1.00)	0.046	0.531
Rectum (N=2,201)						
OS	rs12224794	A	0.42	0.88(0.79-0.97)	0.010	0.339
	rs6854845	T	0.12	1.24(1.02-1.50)	0.027	0.339
	rs13180087	C	0.18	1.18(1.02-1.37)	0.029	0.339
	rs139156	C	0.26	0.82(0.69-0.98)	0.031	0.339
	rs17026425	A	0.08	1.25(1.01-1.56)	0.044	0.339

Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
rs1570271	G	0.14	1.19(1.00-1.41)	0.046	0.339

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

Sensitivity analysis

In order to test potential genetic effects of recessive pattern, I examined the associations of these 43 variants with CRC survival by comparing individuals carrying two copies of risk alleles versus those with zero or one risk allele. In total, four genetic variants (rs16867335, rs17057166, rs17280262, rs6854845) were identified to be associated with both overall and CRC-specific survival with p<0.05. Three out of these four variants (rs17057166, rs17280262, rs6854845) remained significantly associated both survival outcomes after correction for multiple testing using the FDR approach. I summarise the effect estimates of these four variants in **Table 5-10**. Additional details regarding the results of other 39 variants can be found in **Appendix Table 9.**

Table 5-10 Summary of associations (p<0.05) between variants previously linked with CRC survival and outcomes of CRC patients in the SOCCS study under recessive model (N=5,675)

	Variant	MG	MGF	HR(95%CI)	P(uncorrected)	Pfdr
OS	rs17280262	TT	0.01	0.55(0.42-0.70)	1.98E-6	8.71E-5
	rs17057166	TT	0.01	0.60(0.47-0.75)	1.58E-5	3.48E-4
	rs6854845	TT	0.02	0.63(0.50-0.79)	8.42E-5	0.001
	rs16867335	TT	0.05	0.66(0.49-0.89)	0.007	0.072
CSS	rs17280262	TT	0.01	0.54(0.41-0.71)	8.48E-6	3.73E-4
	rs17057166	TT	0.01	0.57(0.43-0.74)	2.19E-5	4.55E-4
	rs6854845	TT	0.02	0.57(0.44-0.74)	3.50E-5	4.55E-4
	rs16867335	TT	0.05	0.58(0.40-0.83)	0.003	0.038

MG, minor genotype; MGF, minor genotype frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival.

For the three significant associations that survived the FDR correction, I further validated the associations between these variants and CRC survival in the UK Biobank cohort (N=2,474) under the recessive model. Overall, I failed to observe any significant effects (p<0.05) of these four variants on either overall or CRC-specific survival (presented in **Table 5-11**). Notably, for the variant rs17057166, I was unable to estimate its effect on CRC-specific survival in the UK Biobank cohort due to the absence of CRC-related deaths observed in CRC patients carrying the effect genotype. Although no variants were successfully replicated with statistical significance, after harmonising the reference genotype for each variant, concordant direction of effects were observed for all of the four variants (**Table 5-11**).

Table 5-11 Summarised associations in the UK Biobank cohort between variants identified from the SOCCS study under recessive model (N=2,474)

	Variant	**MG**	**MGF**	**HR(95%CI)**	**P**
OS	rs16867335	TT	0.05	0.85(0.56-1.27)	0.420
	rs17057166	TT	0.01	0.79(0.20-3.17)	0.739
	rs17280262	TT	0.01	0.93(0.23-3.75)	0.924
	rs6854845	TT	0.02	0.92(0.43-1.93)	0.816
CSS					
	rs16867335	TT	0.05	0.69(0.41-1.16)	0.160
	rs6854845	TT	0.02	0.68(0.25-1.82)	0.442
	rs17280262	TT	0.01	0.58(0.08-4.12)	0.584
	rs17057166	TT	0.01	NA	NA

MG, minor genotype; MGF, minor genotype frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival.

5.3.2 Predictive modelling of CRC-survival variants

Genetic predictor development in the UK Biobank dataset

Full model with 43 genetic variants

I included the same set of 43 variants used in the preceding validation study and fitted a multi-variable Cox regression model on overall and CRC-specific survival using incident CRC cases (N=2,474) in the UK Biobank cohort. The estimated coefficients are presented in **Table 5-12**. We observed 18 (42%) variants with effect alleles that were associated with favourable overall survival (regression coefficient<0) and another group of 18 variants associated with favourable CRC-specific survival (**Table 5-12**). These variants were therefore modelled with opposite direction of effects compared to the original published GWASs, given that the effect allele for each candidate variant had been harmonised. I then created a genetic linear predictor (GP) by summing up the number of effect alleles multiplied by their corresponding regression coefficients. Notably, the genetic predictor was developed in a similar approach compared to the polygenic risk score except that effect alleles were weighted by coefficients derived from the fitted Cox model to form the genetic predictor.

Table 5-12 Regression coefficients for the 43 genetic variants reportedly associated with CRC survival in the UK Biobank (N=5,675)

Variant	MA	MAF	Gene	Regression coefficients	
				OS	**CSS**
rs10040610	C	0.21	*Intergenic*	0.0587	0.0819
rs10152207	A	0.09	*Intergenic*	-0.0997	-0.1187
rs10275272	T	0.28	*TWIST1*	0.0118	0.1203
rs1075232	A	0.10	*AC104759.1*	0.1076	0.1044
rs10921219	A	0.46	*AL390957.1*	-0.0199	-0.0742
rs11138220	G	0.13	*Intergenic*	0.0027	0.0050
rs12187751	G	0.09	*AC113414.1*	-0.0221	0.1233
rs12224794	A	0.42	*LRRC4C*	-0.0121	0.0104

Variant	MA	MAF	Gene	Regression coefficients	
rs13180087	C	0.18	*HCN1*	-0.1610	-0.0768
rs1372474	G	0.07	*LINC01919*	0.1244	0.1367
rs139156	A	0.26	*PARVG*	-0.1112	-0.0804
rs1407508	C	0.02	*AL136084.1*	-0.2627	-0.2622
rs1493383	T	0.10	*GRIA1*	0.0052	0.0868
rs1555895	A	0.50	*Intergenic*	0.0281	0.0117
rs1570271	A	0.14	*Intergenic*	0.0411	0.0586
rs1573948	C	0.10	*Intergenic*	0.0724	0.0950
rs157411	G	0.33	*Intergenic*	-0.0044	0.0068
rs16867335	T	0.20	*AC009478.1*	0.0276	-0.0538
rs17026425	A	0.08	*IQCM*	0.0783	0.0365
rs17048372	T	0.15	*DPP10*	-0.0251	-0.0082
rs17057166	T	0.09	*LINC01847*	-0.0296	-0.1161
rs17087282	A	0.11	*Intergenic*	0.0318	-0.0254
rs17280262	T	0.04	*Intergenic*	0.1019	0.0410
rs1998584	T	0.41	*Intergenic*	0.0024	0.0147
rs2073016	C	0.18	*APOBEC2*	-0.0033	-0.0398
rs209489	C	0.08	*ELOVL5*	0.1335	0.1211
rs2936519	A	0.27	*Intergenic*	0.0735	0.0835
rs338389	G	0.49	*Intergenic*	-0.0056	-0.0095
rs3781663	G	0.31	*ANO1*	0.0869	0.1069
rs3794924	A	0.10	*DSG3*	0.1082	0.0927
rs4377367	C	0.19	*ARHGEF4*	0.0968	0.0976
rs4812219	T	0.14	*Intergenic*	0.0596	0.1114
rs4868304	T	0.14	*LINC01484*	0.1214	0.1178
rs4959799	C	0.14	*SLC22A23*	0.0845	0.0669
rs6105057	G	0.28	*ISM1*	0.0004	-0.0070
rs658495	G	0.04	*C20orf27*	-0.0934	-0.1347

Variant	MA	MAF	Gene	Regression coefficients	
rs6720296	C	0.48	*LINC01121*	-0.0770	-0.1284
rs6854845	T	0.12	*Intergenic*	-0.1194	-0.1932
rs7004484	C	0.30	*EFR3A*	0.0495	0.0667
rs8035094	C	0.10	*Intergenic*	0.0142	0.0342
rs885036	A	0.45	*MGAT4A*	-0.0022	-0.0516
rs912294	A	0.47	*Intergenic*	-0.0490	-0.0127
rs9419702	C	0.26	*Intergenic*	-0.1049	-0.1049

MA, minor allele; MAF, minor allele frequency; OS, overall survival, CSS, CRC-specific survival

Feature selection

A LASSO regression model was fitted so as to select variants with potential predictive value from these 43 candidates. After applying a 10-fold cross-validation, the tuning factor lambda was identified at 0.0461 for overall survival and 0.0249 for CRC-specific survival corresponding to the most parsimonious model with the cross-validation prediction error within one standard error (SE) of the minimum. However, no variants were selected into the ultimate model. Similarly, I identified no variants remaining in the final model after performing backward selection (p<0.15). Given the absence of any variants that survived feature selection, only the full model with all variants was further evaluated.

Predictive performance of the genetic predictor after internal validation

The Harrell's concordance index (C statistic) was used to evaluate the discriminative performance of the genetic predictor developed from the UK Biobank cohort. Both the observed and predicted survival estimates were obtained in order to derive the C statistic. The observed survival estimates were extracted directly from the Kaplan-Meier estimates, whilst the predicted survival rates were derived from the LP of the 43 genetic variants and the baseline hazard function estimated from the fitted Cox model. Based on the observed and predicted survival estimates in the UK Biobank cohort, the genetic predictor yielded an apparent C statistic (without internal validation) of 0.558 (95%CI=0.502-0.595) for overall survival and 0.570 (95%CI= 0.502-0.613) for CRC-specific survival.

As introduced in Chapter 4, over-optimism could be generated if the performance is estimated directly from the same dataset in which the predictor is developed. Therefore, I conducted internal validation by re-evaluating and summarising the C statistics in 200 bootstrap samples. The results showed that the C statistic was reduced to 0.510 (95% CI=0.498-0.521) for overall survival and 0.518 (95% CI=0.498-0.530) for CRC-specific survival after bootstrapping. The fact that the confidence intervals of C statistic after bootstrapping included the null value (0.5) indicated a statistically non-significant discriminative ability for the full model of 43 variants after internal validation. The decreased C statistics for both outcomes supported the presence of over-fitting for the genetic predictor. As described in section 4.3.2, over-fitting could be quantified by the shrinkage factor (SF) and the model could be adjusted by applying the SF. From the bootstrap samples, I obtained a shrinkage factor (SF) of 0.4729 for overall survival and 0.4868 for CRC-specific survival. Then the regression coefficients listed in **Table 5-12** were multiplied with the SF to generate two new sets of coefficients to calculate the shrunken LP and corresponding predicted survival estimates for each individual. The observed survival rates and the updated predicted survival rates are presented in **Table 5-13**. I then performed the Hosmer-Lemeshow test to assess the accordance of the observed and predicted estimates derived from the adjusted model. The results of the test showed significant departure of the predicted rates from the observed estimates (p<0.001). Additionally, to visualise the relationship between the observed and predicted estimates, calibration curves are plotted in **Figure 5-9**. As depicted, the adjusted model after internal validation showed inaccurate model calibration for both overall and CRC-specific survival.

Table 5-13 Comparison between observed and predicted 5-year survival estimates stratified by the quartiles of the shrunken 43-variant genetic predictor in the UK Biobank (N=2,474)

	5-year Survival estimates		
Quartiles (GP)			
Overall survival	Observed	Predicted	HL-p
Q1	0.835	0.748	1.8E-8
Q2	0.810	0.723	

	5-year Survival estimates		
Q3	0.799	0.706	
Q4	0.719	0.681	
CRC-specific survival			
Q1	0.833	0.828	3.2E-6
Q2	0.813	0.807	
Q3	0.809	0.791	
Q4	0.709	0.768	

Q, quartile; CRC, colorectal cancer; HL-p, p-value of the Hosmer-Lemeshow test; GP, genetic predictor.

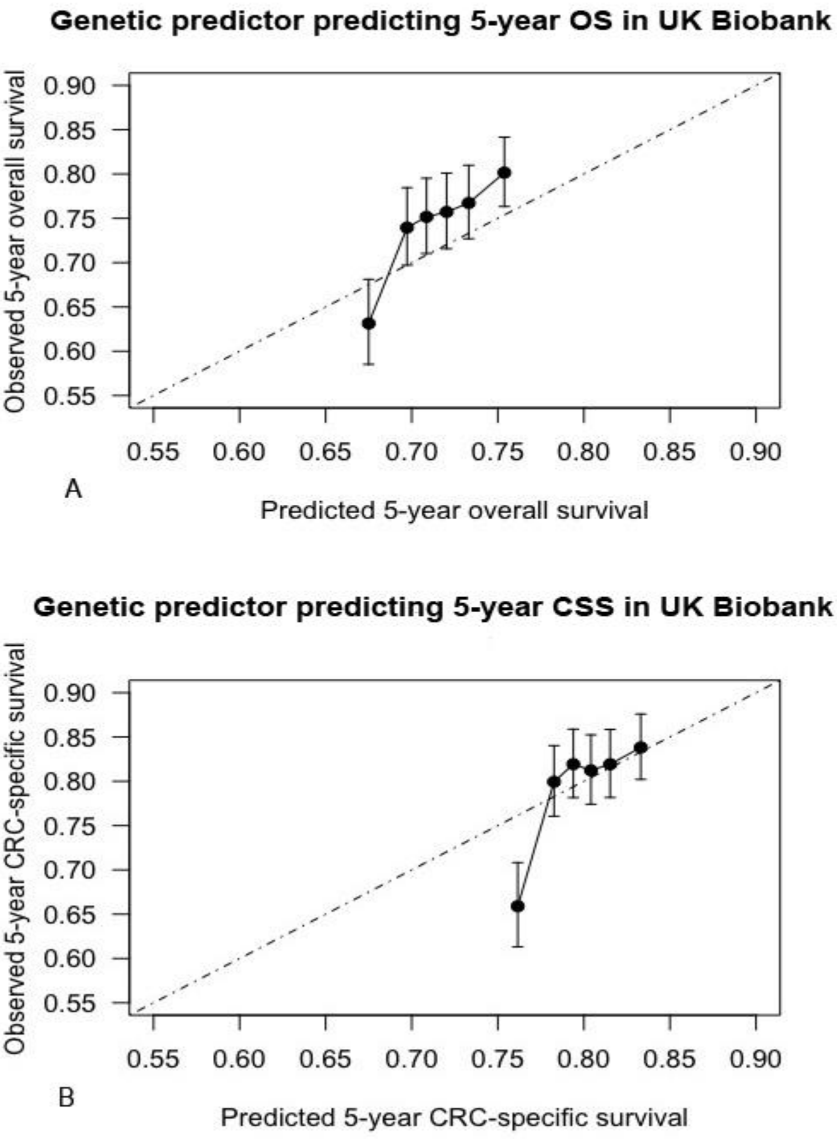


Figure 5-9 Calibration plots of the adjusted model of 43-variant genetic predictor predicting 5-year overall (A) and CRC-specific survival (B).

I regressed the observed survival estimates on the predictive value in a linear regression model to obtain the coefficient along with the intercept that were used to recalibrate the predictive survival rates (overall survival: β=1.9010, intercept=-0.6172; CRC-specific survival: β=2.1757, intercept=-0.9459). Based on the shrunken model, a 5-year baseline survival rate of 0.7156 for overall survival and 0.7997 for CRC-specific survival was derived. Combining the baseline survival and the shrunken regression coefficients, the predicted 5-year survival rates are expressed as follows:

Predicted 5-year overall survival:

$$1.9010 \times \left(0.7156^{\exp(0.4729 \times GP)}\right) - 0.6172$$

Predicted 5-year CRC-specific survival:

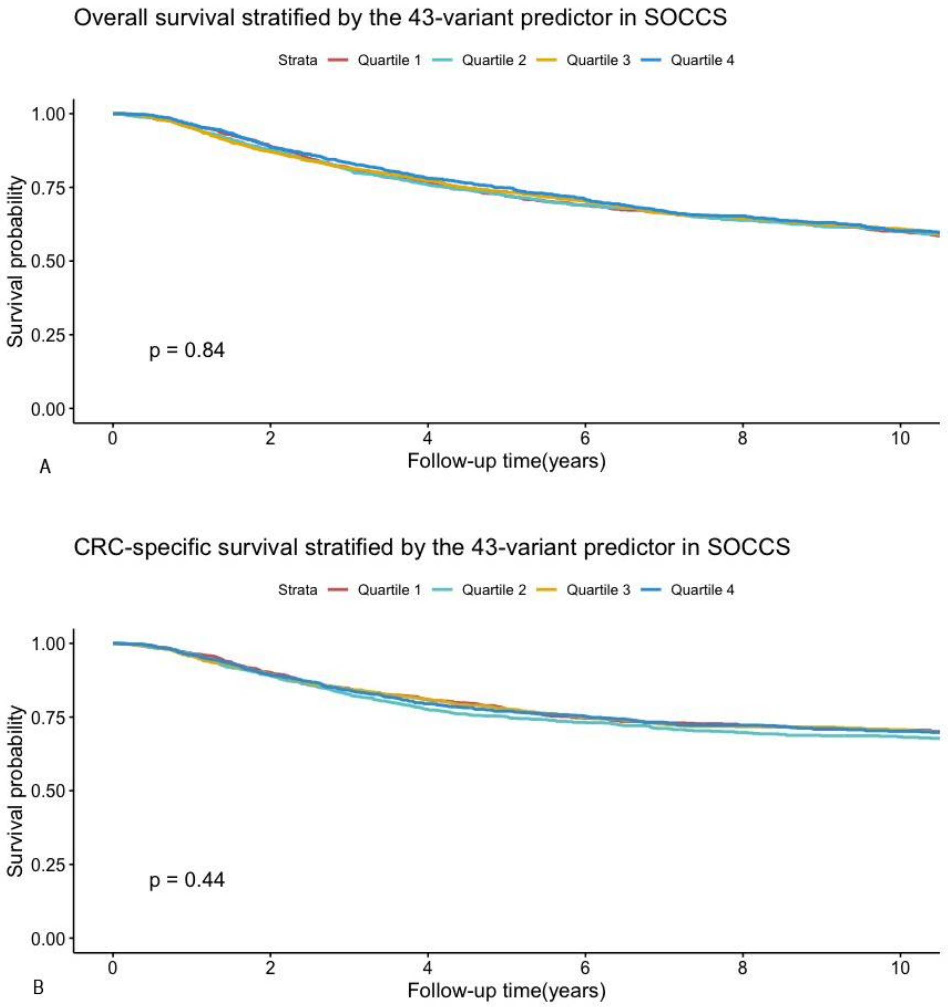
$$2.1757 \times \left(0.7997^{\exp(0.4868 \times GP)}\right) - 0.9459$$

Following the formula above, the genetic predictor was then externally validated in the SOCCS dataset.

Predictive modelling in the SOCCS study

External validation of the genetic predictor

I first investigated the association between the genetic predictor created for each patient in the SOCCS and survival outcomes in a univariable Cox model. However, no significant association was found for either overall (HR=0.93, 95%CI=0.76-1.13, p=0.457) or CRC-specific survival (HR=1.05, 95%CI=0.86-1.28, p=0.622). The stratified overall and CRC-specific survival curves by quartiles of the genetic predictor are plotted in **Figure 5-10**. As shown by the figure, the genetic predictor was unable to efficiently differentiate CRC patients in the SOCCS of varied observed survival outcomes. With respect to the discriminative performance, a positive yet insignificant C statistic was observed when applying the genetic predictor to predict overall survival (C=0.512, 95%CI=0.480-0.544) of CRC patients in the SOCCS study. However, for CRC-specific survival, a negative point estimate of C statistic (0.499) was obtained with the 95% confidence interval also including the null (95%CI=0.464-0.534).


Overall survival stratified by the 43-variant predictor in SOCCS

CRC-specific survival stratified by the 43-variant predictor in SOCCS

Figure 5-10 Kaplan-Meier estimates of overall (A) and CRC-specific (B) survival stratified by the 43-variant genetic predictor in SOCCS

To evaluate the model calibration, I extracted the predicted and observed 5-year survival rates for each quartile of the genetic predictor. These estimates are listed in **Table 5-14**. Although the predicted survival rates decreased in sequence, the corresponding observed rates almost remained unchanged, indicating poor prediction accuracy of the genetic predictor. This could also be seen from the calibration plots of overall and CRC-specific survival (**Figure 5-11**).

Table 5-14 Comparison between observed and predicted 5-year survival estimates stratified by the quartiles of the 43-variant genetic predictor in the SOCCS (N=5,675)

	5-year survival estimates		
Quartiles (GP)			
Overall survival	Observed	Predicted	HL-p
Q1	0.747	0.809	4.1E-13
Q2	0.721	0.764	
Q3	0.736	0.733	
Q4	0.748	0.682	
CRC-specific survival			
Q1	0.771	0.871	1.9E-13
Q2	0.751	0.829	
Q3	0.778	0.795	
Q4	0.770	0.744	

Q, quartile; CRC, colorectal cancer; HL-p, p-value of the Hosmer-Lemeshow test; Gp, genetic predictor

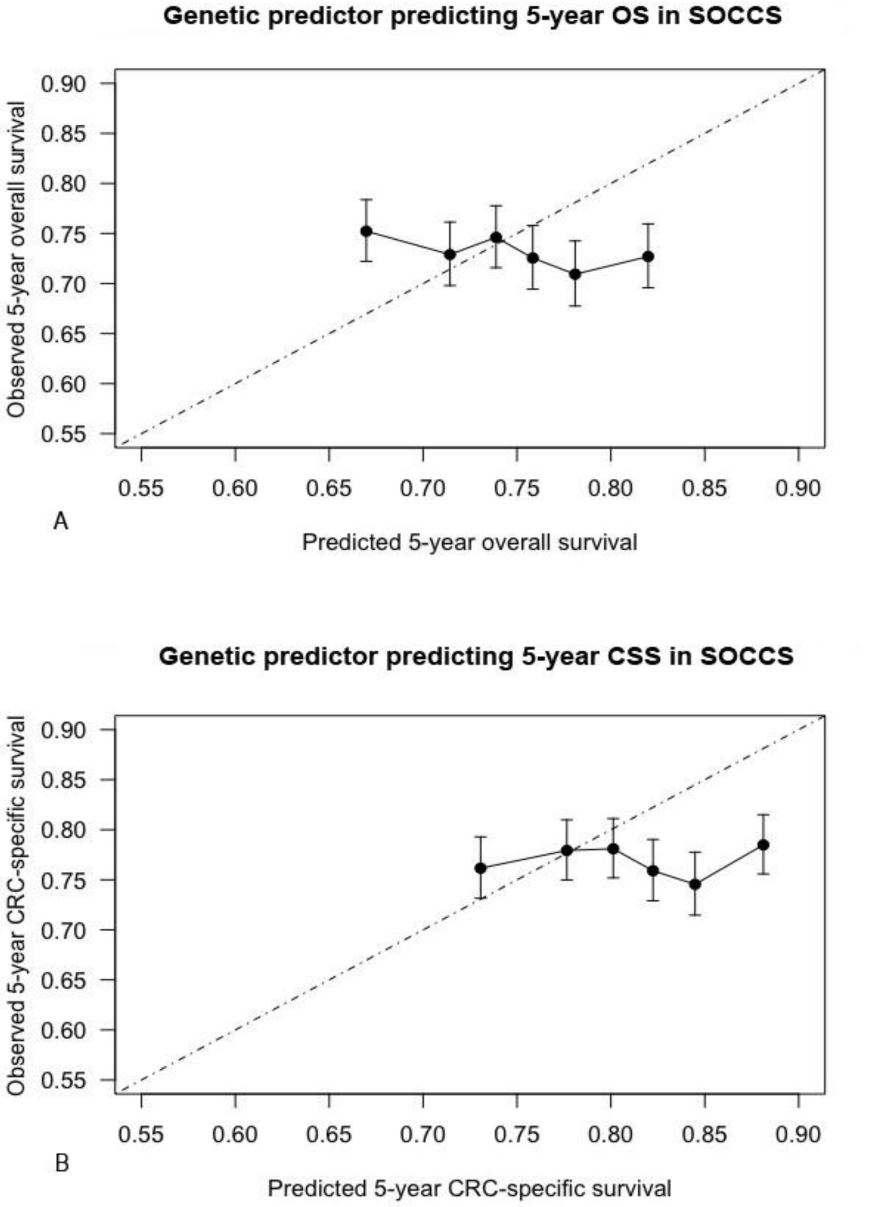


Figure 5-11 Calibration plots of the 43-variant genetic predictor predicting 5-year overall (A) and CRC-specific survival (B) in the SOCCS

Combining the genetic predictor with other variables

I further assessed the potential added predictive value of the 43-variant genetic predictor on the basis of a baseline model including other non-genetic predictors. Given that tumour site was not significantly associated with either overall or CRC-specific survival (**Table 5-3**), two Cox regression models were fitted—the baseline model (Model 1) included non-genetic predictors of age at diagnosis, sex, AJCC stage and tumour grade, and the Model 2 added the genetic predictor to the baseline model. The regression coefficients of these two models are summarised in **Table 5-15**. In the multivariable model, age and AJCC stage were significantly associated with overall survival (p<0.05), whereas AJCC stage and tumour grade were associated with CRC-specific survival.

Table 5-15 Summarised coefficients of Cox regression models with or without the genetic predictor in SOCCS (N=4,836)

	Model 1*		**Model 2****	
Variables	**HR(95%CI)**	**P**	**HR(95%CI)**	**P**
Overall survival				
Age	1.023(1.018-1.028)	5.9E-20	1.023(1.018-1.028)	1.1E-19
Sex (Male)	1.174(1.059-1.302)	0.002	1.176(1.060-1.304)	0.002
Stage	2.113(1.891-2.253)	3.5E-119	2.111(1.980-2.251)	2.0E-115
Grade	1.064(0.941-1.204)	0.322	1.067(0.944-1.207)	0.300
GP	NA		0.849(0.684-1.054)	0.138
CRC-specific survival				
Age	1.003(0.997-1.008)	0.345	1.003(0.997-1.008)	0.347
Sex (Male)	1.102(0.972-1.249)	0.128	1.102(0.973-1.249)	0.128
Stage	3.100(2.855-3.366)	4.7E-163	3.100(2.855-3.366)	1.3E-159
Grade	1.231(1.065-1.423)	0.005	1.231(1.065-1.424)	0.005
GP	NA		0.992(0.793-1.240)	0.941

*Model 1 includes non-genetic predictors of age at diagnosis, sex, AJCC stage and tumour grade
**Model 2 includes the genetic predictor and predictors in Model 1
HR, hazard ratio; CI, confidence interval; GP, genetic predictor; NA, not available

The potential added discriminative performance of the genetic predictor was evaluated by comparing the C statistics of the two models after internal validation using the same bootstrapping procedure in UK Biobank. I first calculated the apparent C statistic using the same SOCCS cohort from which the two models had been derived. For the outcome of overall survival, the baseline model with age, sex, AJCC stage and tumour grade showed a significant C statistic of 0.703(95%CI=0.675-0.734) after bootstrapping. The addition of the genetic predictor led to an almost equivalent point estimate of C statistic (0.704) along with a confidence interval (95%CI=0.677-0.731). The U-statistic based test found no significant increase of model concordance with the addition of the genetic predictor (p=0.46). Pertaining to the other outcome of CRC-specific survival, the baseline model yielded a C statistic of 0.761 (95%CI=0.732-0.790) after internal validation. A slightly reduced point estimate of C statistic was identified in the model 2 after adding the genetic predictor (C=0.760, 95%CI=0.729-0.789). Similarly, the U-statistic based test did not detect a significant change in the discriminative ability with the addition of the genetic predictor (p=0.50).

I then applied the same procedure and re-estimated the shrinkage factors in 200 bootstrap samples to quantify the potential over-optimism. For the baseline model 1, a shrinkage factor was identified showing only slight optimism for both overall (SF=0.994) and CRC-specific survival (SF=0.998). Similarly, shrinkage factors of the model 2 with the genetic predictor were also close to 1 (overall survival: SF=0.996, CRC-specific survival: SF=0.991). The shrinkage factors were applied to adjust and re-fit the models. Both the observed and predicted 5-year survival estimates were calculated to evaluate the model calibration. Results of the Hosmer-Lemeshow test found no significant deviation of predicted 5-year overall (Model 1: p=0.110; Model 2: p=0.251) and CRC-specific survival (Model 1: p=0.132, Model 2: p=0.290) from the observed estimates for the two models. The calibration plots of the baseline Model 1 and Model 2 for both the 5-year overall and CRC-specific survival in the SOCCS are presented in **Figure 5-12** and **Figure 5-13** respectively. As shown, prediction models in SOCCS of non-genetic prognostic factors with or without the genetic predictor were generally well-calibrated except for patients with moderate 5-year survival rates (50%-70%) where the predicted survival rates tended to be lower than the observed survival rates.

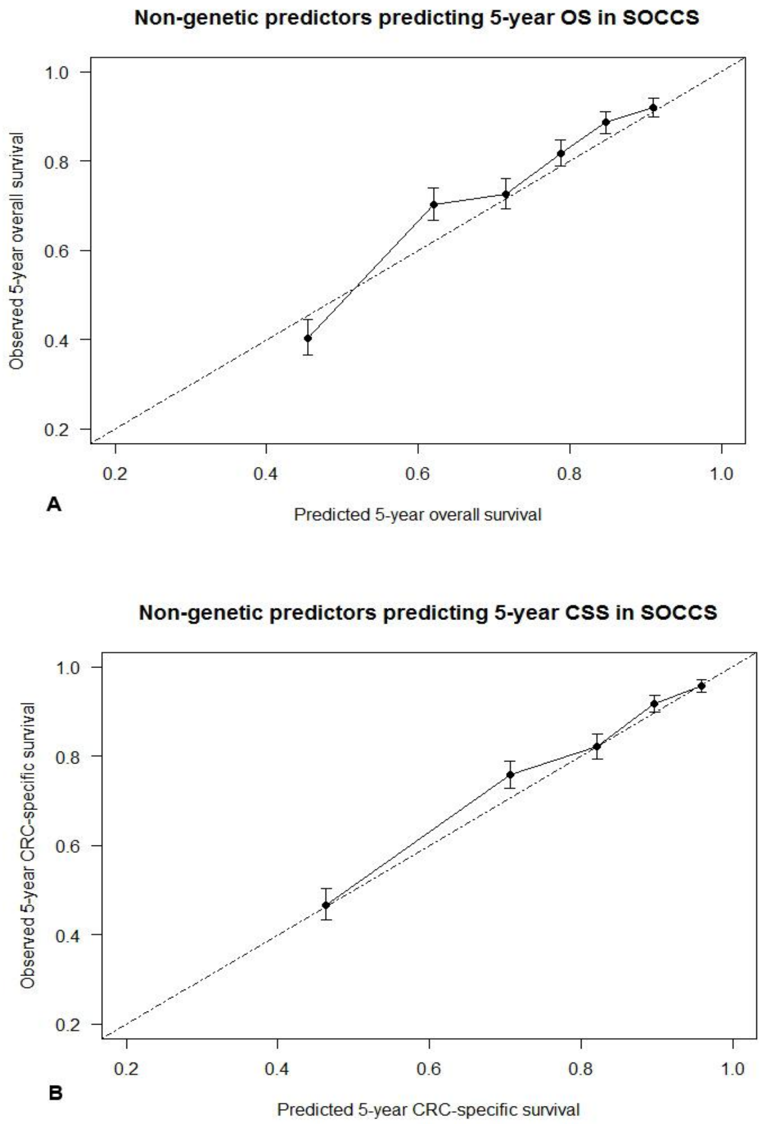


Figure 5-12 Calibration plots of the prediction model of non-genetic factors (age, sex, AJCC stage and tumour grade) predicting 5-year overall (A) and CRC-specific survival (B) in the SOCCS cohort

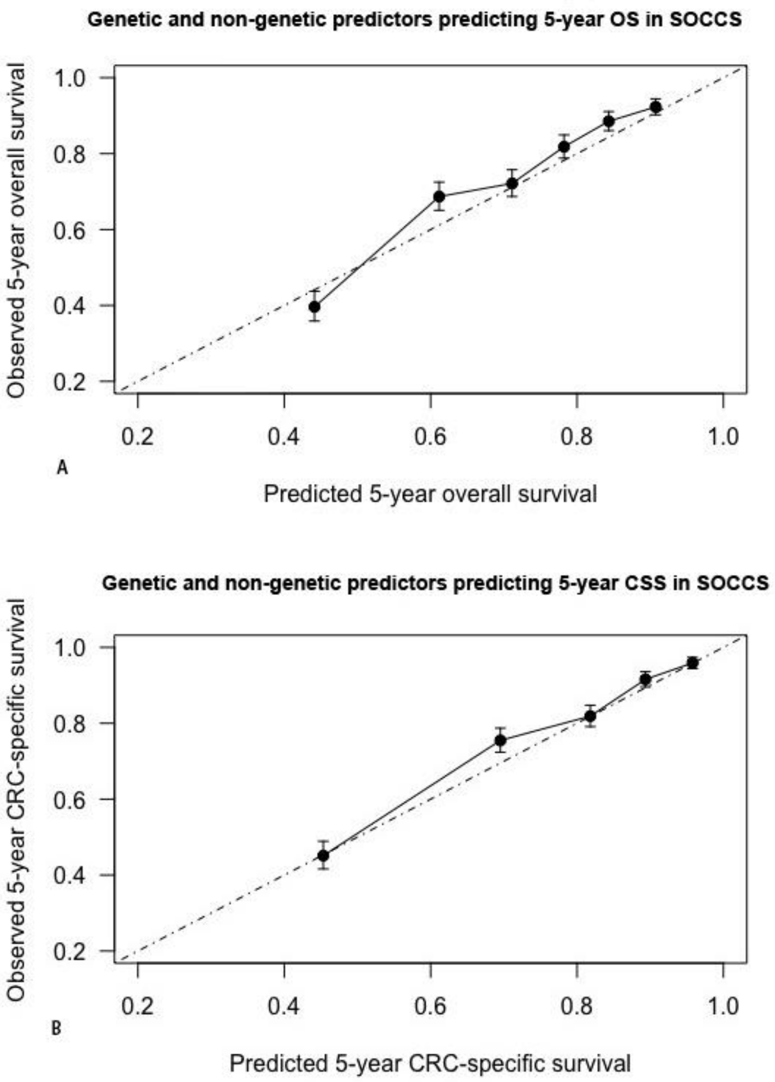


Figure 5-13 Calibration plots of the prediction model combining the genetic and non-genetic factors (age, sex, AJCC stage and tumour grade) predicting 5-year overall (A) and CRC-specific survival (B) in the SOCCS cohort.

Predictive modelling in locally advanced (stage II/III) CRC

Although I failed to detect significant increase of model performance by adding the genetic predictor into the model, the genetic predictor was still retained to develop prediction models in stage II/III (N=3,886) CRC patients in the SOCCS cohort.

Firstly, I fitted a Cox regression model including the genetic predictor, age at diagnosis, sex, AJCC stage (II or III) and tumour grade as covariates. The effect estimates of covariates are summarised in **Table 5-16**. I observed that age, sex and stage were significantly associated with overall survival (p<0.05), whilst sex, stage and tumour grade were significant prognostic factors for CRC-specific survival (p<0.05) in the multivariable model.

Table 5-16 Summarised coefficients of Cox regression models for stage II/III CRC patients in SOCCS (N=3,886)

Variables	**HR(95%CI)**	**P**	**HR(95%CI)**	**P**
Overall survival			**CRC-specific survival**	
Age	1.028(1.022-1.035)	3.83E-19	1.005(0.998-1.012)	0.177
Sex **(Male vs Female)**	1.248(1.098-1.419)	7.27E-4	1.189(1.014-1.394)	0.034
Stage **(III vs II)**	1.625(1.429-1.847)	1.18E-13	2.320(1.960-2.747)	1.52E-22
Grade	1.148(0.989-1.334)	0.070	1.355(1.132-1.622)	9.49E-4
GP	0.895(0.683-1.173)	0.422	1.045(0.788-1.387)	0.760

HR, hazard ratio; CI, confidence interval; GP, genetic predictor; NA, not available

A linear predictor combining these factors weighted by their regression coefficients was created for each individual, and the Kaplan-Meier estimates of each quartile of the linear predictor are plotted in **Figure 5-14** to display the survival outcomes of stage II/III patients stratified by the multivariable model. The summarised C statistic derived from the model after bootstrapping was 0.604(95%CI=0.563-0.643) for overall survival and 0.636(95%CI=0.592-0.680) for CRC-specific survival.

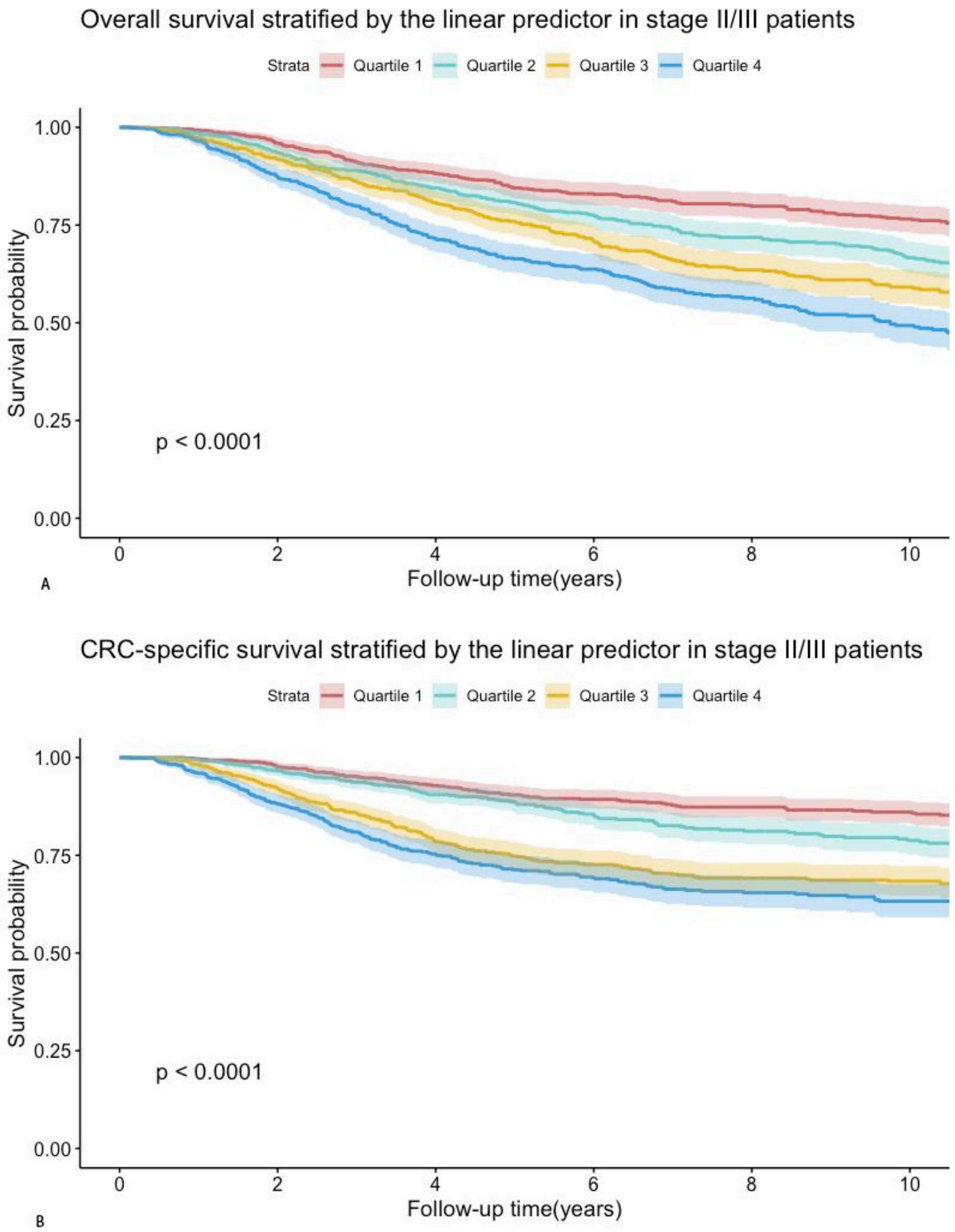


Figure 5-14 Kaplan-Meier estimates of stage II/III patients in the SOCCS stratified by the linear predictor (A: overall survival; B: CRC-specific survival)

The shrinkage factor was 0.969 for overall survival and 0.978 for CRC-specific survival respectively. Then the regression coefficients listed in **Table 5-16** were multiplied with the shrinkage factor to adjust the model and obtained updated coefficients to predict 5-year survival probabilities. The predicted 5-year survival rates estimated from the developed model and observed 5-year survival estimates extracted from the Kaplan-Meier curves grouped by quartiles of the linear predictor are presented in the **Table 5-17**. The result of Hosmer-Lemeshow test did not suggest significant departure of the predicted 5-year survival rates from the observed estimates for both overall and CRC-specific survival (**Table 5-17**). Calibration plots (**Figure 5-15**) also showed that these two adjusted models were generally well-calibrated, except that the model might be less sensitive in predicting CRC-specific survival for patients of relatively good survival outcome with an observed survival rate of approximately 90% (**Figure 5-15B**).

Table 5-17 Comparison between observed and predicted 5-year survival estimates in stage II/III patients of the SOCCS study stratified by the quartiles of the linear predictor (N=3,886)

	5-year survival estimates		
Quartiles (LP)			
Overall survival	Observed	Predicted	HL-p
Q1	0.845	0.854	0.280
Q2	0.807	0.792	
Q3	0.760	0.737	
Q4	0.663	0.640	
CRC-specific survival			
Q1	0.904	0.890	0.188
Q2	0.886	0.859	
Q3	0.748	0.759	
Q4	0.713	0.701	

Q, quartile; CRC, colorectal cancer; HL-p, p-value of the Hosmer-Lemeshow test; LP, linear predictor of genetic and non-genetic factors

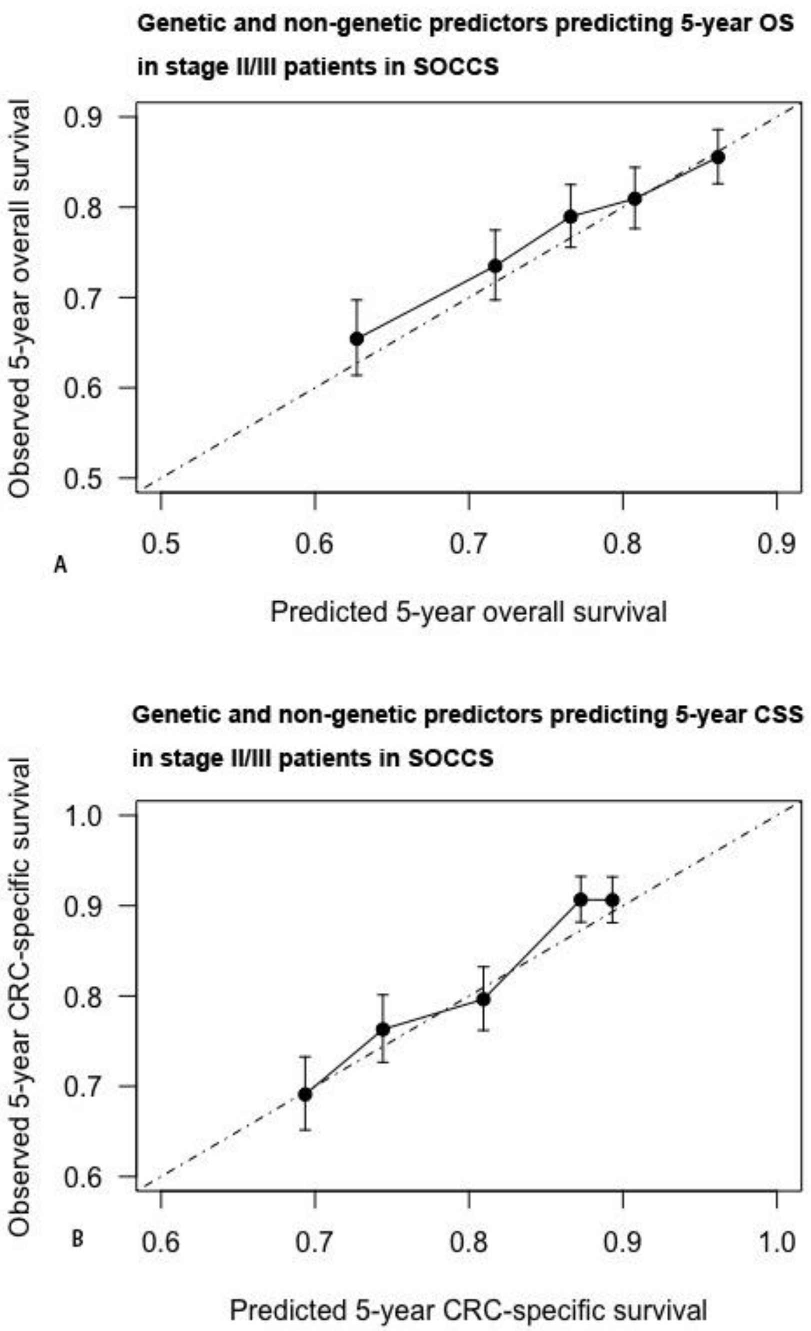


Figure 5-15 Calibration plots of the prediction models predicting 5-year overall (A) and CRC-specific survival (B) of stage II/III patients in the SOCCS cohort.

Similar to the previous procedure, I regressed the observed 5-year survival estimates on the predicted values in a univariable linear regression model and obtain the coefficient along with the intercept to further re-calibrate the model (overall survival: β=0.8593, intercept=0.1192; CRC-specific survival: β=0.9014, intercept=0.1344). The 5-year baseline survival was also retrieved from the adjusted model (0.7680 for overall survival and 0.8165 for CRC-specific survival). The re-calibrated 5-year survival probability for a given stage II/III CRC patient is then expressed as:

Predicted 5-year overall survival:

$$0.8593 \times \left(0.7680^{\exp\left(0.969\times(0.028\times Age+0.2215\times Sex(1\ for\ Male)+0.4845\times AJCC\ stage+0.1382\times tumour\ grade-0.1107\times GP)\right)}\right) + 0.1192$$

Predicted 5-year CRC-specific survival:

$$0.9014 \times \left(0.8165^{\exp\left(0.978\times(0.005\times Age+0.1728\times sex(1\ for\ Male)+0.8417\times AJCC\ stage+0.3038\times tumour\ grade+0.044\times GP)\right)}\right) + 0.1344$$

In clinical practice, there has been interest in identifying high/low risk stage II/III CRC patients for tailored treatment strategy. Therefore, in addition to predicting the absolute survival probability, I further explored potential performance of these factors to characterise patients of different risk profiles within the stage II and stage III strata separately. Cox regression models were fitted in stage II and stage III patients with age, sex, tumour grade and the genetic predictor as covariates. Then I assigned the patients into high- and low-risk groups by the median of the linear predictor derived from the fitted model. The linear predictors along with the cut-off median values obtained from the models are summarised in the **Table 5-18** below. I then plotted the Kaplan-Meier survival curves for stage II and III patients with high and low risk of deaths in **Figure 5-16**. As shown in the plots, the risk score combining age, sex, tumour grade and the genetic predictor might be less sensitive in stratifying stage III CRC patients. However, once externally validated, it might be useful to assist in identifying low risk stage II patients with potential favoured overall survival (not for CRC-specific survival) as candidates who might be exempt from intensive treatment strategies.

Table 5-18 Summarised score rules of characterising stage II/III patients with high/low risk profiles

	Risk profiling score	**Cut-off median of the score**
Stage II		
Overall survival	0.046*Age+0.219*Sex+0.025*grade-0.309*GP	0.0565
CRC-specific survival	0.017*Age+0.241*Sex+0.369*grade-0.261*GP	0.0152
Stage III		
Overall survival	0.017*Age+0.227*Sex+0.196*grade-0.019*GP	0.0198
CRC-specific survival	0.00016*Age+0.143*Sex+0.281*grade +0.159*GP	-0.0063

GP, genetic predictor, CRC, colorectal cancer

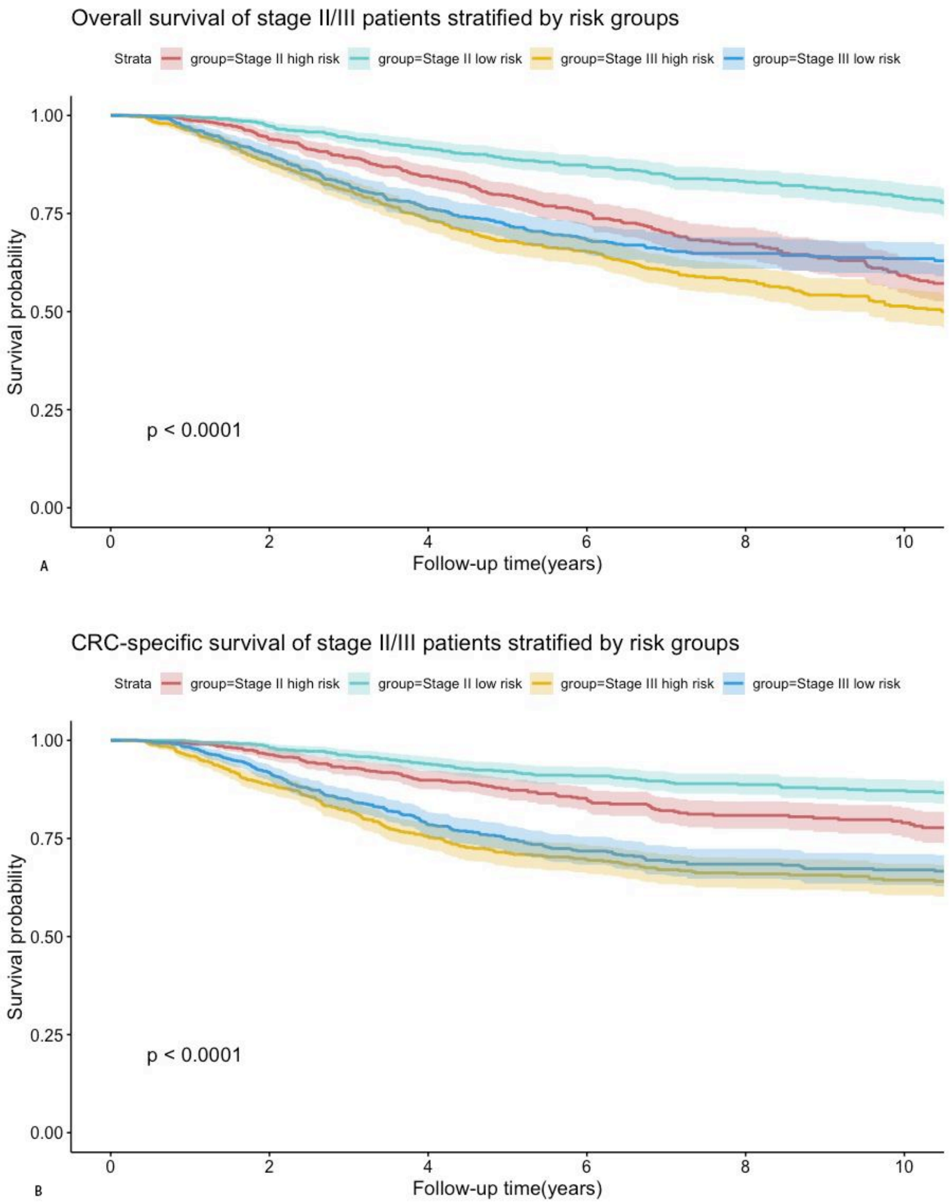


Figure 5-16 Kaplan-Meier estimates of low and high-risk stage II/III patients stratified by the linear predictor (A: overall survival; B: CRC-specific survival)

5.3.3 Variants associated with CRC risk

Eligible variants

As stated in section 4.3.1, genetic variants associated with CRC risk were extracted directly from the two latest meta-analysis of GWASs (to date) that also include all previously known CRC-risk variants (Huyghe *et al*, 2019; Law *et al*, 2019). Initially, a total of 141 autosomal genetic variants associated with CRC risk at p<5x10^{-8} were extracted from these two meta-analyses. No additional independent variants were retrieved after searching the GWAS catalogue. After excluding variants in linkage disequilibrium with other variants, a total of 128 genetic variants were eligible for this candidate genetic association study at last. These 128 variants were originally discovered and reported in 24 GWASs (Al-Tassan *et al*, 2015; Dunlop *et al*, 2012; Huyghe *et al*, 2019; Jia *et al*, 2013; Jiang *et al*, 2015; Law *et al*, 2019; Peters *et al*, 2013; Real *et al*, 2014; Schmit *et al*, 2018; Schmit *et al*, 2014; Study *et al*, 2008; Tanikawa *et al*, 2018; Tenesa *et al*, 2008; Tomlinson *et al*, 2007; Tomlinson *et al*, 2011; Tomlinson *et al*, 2008; Wang *et al*, 2014; Wang *et al*, 2016; Whiffin *et al*, 2014; Zeng *et al*, 2016; Zhang *et al*, 2014). Included genetic variants were located in chromosomes 1 to 20 (details presented in **Table 5-19**).

Table 5-19 Summary details of the included genetic variants previously associated with CRC risk

Variant	Locus	MA	MAF	Gene	Reference
rs12143541	1p32.3	G	0.07	*TTC22*	Law, 2019
rs61776719	1p34.3	C	0.38	*FHL3*	Law, 2019
rs72647484	1p36.12	T	0.03	*Intergenic*	Law, 2019
rs10911251	1q25.3	A	0.37	*LAMC1*	Law, 2019
rs6658977	1q41	T	0.21	*LINC02257*	Law, 2019
rs11692435	2q11.2	G	0.02	*ACTR1B*	Law, 2019
rs448513	2q24.2	C	0.50	*TANC1*	Huyghe, 2019
rs11903757	2q32.3	C	0.12	*Intergenic*	Law, 2019
rs11893063	2q33.1	A	0.30	*AC019330.1*	Law, 2019
rs7593422	2q33.1	T	0.35	*SATB2*	Law, 2019
rs13020391	2q35	C	0.30	*PNKD*	Law, 2019
rs2279290	3p14.1	G	0.11	*LRIG1*	Law, 2019
rs9831861	3p21.1	G	0.43	*AC096887.1*	Law, 2019
rs35360328	3p22.1	A	0.08	*Intergenic*	Law, 2019
rs12635946	3q13.2	C	0.29	*Intergenic*	Law, 2019
rs72942485	3q13.2	G	0.05	*BOC*	Huyghe, 2019
rs10049390	3q22.2	A	0.43	*SLCO2A1*	Huyghe, 2019
rs10936599	3q26.2	C	0.27	*MYNN*	Law, 2019
rs1370821	4q22.2	T	0.26	*Intergenic*	Law, 2019
rs1391441	4q24	A	0.34	*TET2*	Huyghe, 2019

Variant	Locus	MA	MAF	Gene	Reference
rs17035289	4q24	T	0.25	*Intergenic*	Law, 2019
rs3987	4q26	G	0.39	*LINC02264*	Law, 2019
rs75686861	4q31.21	A	0.03	*HHIP*	Law, 2019
rs186722897	4q32.2	T	0.10	*Intergenic*	Law, 2019
rs35509282	4q32.2	A	0.23	*Intergenic*	Law, 2019
rs1445011	5p13.1	C	0.13	*Intergenic*	Law, 2019
rs7708610	5p13.1	A	0.32	*Intergenic*	Huyghe, 2019
rs2735940	5p15.33	A	0.47	*TERT*	Law, 2019
rs77776598	5p15.33	C	0.03	*SLC6A18*	Law, 2019
rs12522693	5q23.3	G	0.11	*Intergenic*	Law, 2019
rs639933	5q31.1	C	0.31	*C5orf66*	Law, 2019
rs647161	5q31.1	A	0.46	*C5orf66*	Law, 2019
rs62404966	6p12.1	C	0.13	*BMP5*	Law, 2019
rs4711689	6p21.1	A	0.23	*TFEB*	Law, 2019
rs6933790	6p21.1	T	0.20	*TFEB*	Law, 2019
rs1321310	6p21.2	C	0.28	*Intergenic*	Law, 2019
rs16878812	6p21.31	A	0.13	*FKBP5*	Law, 2019
rs2516420	6p21.32	C	0.06	*HCP5*	Huyghe, 2019
rs9271770	6p21.32	A	0.29	*HLA-DQA1*	Law, 2019
rs3131043	6p21.33	G	0.49	*HCG20*	Law, 2019
rs2070699	6p24.1	T	0.36	*EDN1*	Law, 2019
rs6928864	6q21	C	0.29	*Intergenic*	Law, 2019
rs10951878	7p12.3	C	0.45	*AC004870.4*	Law, 2019
rs3801081	7p12.3	G	0.25	*TNS3*	Law, 2019
rs12672022	7p13	T	0.13	*TBRG4*	Huyghe, 2019
rs16892766	8q23.3	C	0.08	*Intergenic*	Law, 2019
rs4313119	8q24.21	G	0.28	*Intergenic*	Huyghe, 2019
rs6983267	8q24.21	G	0.39	*CASC8*	Law, 2019
rs1412834	9p21.3	T	0.31	*CDKN2B-AS1*	Law, 2019
rs34405347	9q22.33	T	0.17	*Intergenic*	Huyghe, 2019
rs10980628	9q31.3	C	0.16	*LPAR1*	Huyghe, 2019
rs10795668	10p14	G	0.23	*RNA5SP299*	Law, 2019
rs10994860	10q11.23	C	0.16	*A1CF*	Law, 2019
rs704017	10q22.3	G	0.45	*ZMIZ1-AS1*	Law, 2019
rs1035209	10q24.2	T	0.12	*Intergenic*	Law, 2019
rs4919687	10q24.32	G	0.19	*CYP17A1*	Law, 2019
rs11196171	10q25.2	G	0.5	*TCF7L2*	Law, 2019

Variant	Locus	MA	MAF	Gene	Reference
rs12241008	10q25.2	C	0.19	*VTI1A*	Law, 2019
rs4450168	11p15.4	C	0.08	*SBF2*	Law, 2019
rs174537	11q12.2	G	0.3	*MYRF*	Law, 2019
rs3824999	11q13.4	G	0.33	*POLD3*	Law, 2019
rs4944940	11q13.4	G	0.03	*CHRDL2*	Law, 2019
rs2186607	11q22.1	T	0.37	*TRPC6*	Huyghe, 2019
rs3087967	11q23.1	T	0.27	*C11orf53*	Law, 2019
rs2238126	12p13.2	G	0.23	*ETV6*	Law, 2019
rs10849432	12p13.31	T	0.17	*Intergenic*	Law, 2019
rs10849438	12p13.31	G	0.17	*Intergenic*	Law, 2019
rs11064437	12p13.31	C	0.16	*TPI1/RPL13P5*	Law, 2019
rs10774214	12p13.32	T	0.45	*CCND2-AS1*	Law, 2019
rs3217810	12p13.32	T	0.05	*CCND2*	Law, 2019
rs3217874	12p13.32	T	0.42	*CCND2*	Huyghe, 2019
rs11610543	12q12	G	0.48	*Intergenic*	Huyghe, 2019
rs11169552	12q13.13	C	0.25	*ATF1*	Law, 2019
rs4759277	12q13.3	A	0.37	*LRP1*	Huyghe, 2019
rs7398375	12q13.3	C	0.32	*LRP1*	Law, 2019
rs3184504	12q24.12	C	0.15	*SH2B3*	Law, 2019
rs72013726	12q24.21	C	0.48	*Intergenic*	Law, 2019
rs73208120	12q24.22	G	0.05	*NOS1*	Law, 2019
rs10161980	13q13.2	C	0.35	*AL139383.1*	Law, 2019
rs9537521	13q13.2	G	0.18	*AL139383.1*	Law, 2019
rs12427600	13q13.3	C	0.20	*SMAD9*	Law, 2019
rs45597035	13q22.1	A	0.20	*KLF5*	Law, 2019
rs78341008	13q22.1	C	0.02	*Intergenic*	Huyghe, 2019
rs1330889	13q22.3	C	0.13	*LINC00446*	Law, 2019
rs7993934	13q34	T	0.42	*COL4A2*	Law, 2019
rs1570405	14q22.2	G	0.50	*Intergenic*	Law, 2019
rs35107139	14q22.2	C	0.48	*BMP4*	Law, 2019
rs17094983	14q23.1	G	0.10	*LINC01500*	Huyghe, 2019
rs11632715	15q13.3	A	0.45	*Intergenic*	Law, 2019
rs16959063	15q13.3	A	0.01	*FMN1*	Law, 2019
rs16969681	15q13.3	T	0.20	*SCG5*	Law, 2019
rs17816465	15q13.3	A	0.12	*FMN1*	Law, 2019
rs73376930	15q13.3	G	0.32	*GREM1*	Law, 2019

Variant	Locus	MA	MAF	Gene	Reference
rs4776316	15q22.31	A	0.22	*SMAD6*	Law, 2019
rs56324967	15q22.33	C	0.39	*SMAD3*	Huyghe, 2019
rs10152518	15q23	G	0.37	*Intergenic*	Law, 2019
rs7495132	15q26.1	T	0.15	*CRTC3*	Law, 2019
rs9929218	16q22.1	G	0.26	*CDH1*	Law, 2019
rs61336918	16q23.2	A	0.39	*Intergenic*	Law, 2019
rs2696839	16q24.1	G	0.34	*Intergenic*	Law, 2019
rs847208	16q24.1	A	0.33	*LINC01081*	Law, 2019
rs899244	16q24.1	T	0.21	*AC009154.1*	Law, 2019
rs1078643	17p12	A	0.45	*TMEM238L*	Law, 2019
rs12603526	17p13.3	C	0.05	*NXN*	Law, 2019
rs73975588	17p13.3	A	0.07	*NXN*	Law, 2019
rs17836917	17q12	G	0.05	*ASIC2*	Law, 2019
rs983318	17q24.3	A	0.12	*LINC00511*	Huyghe, 2019
rs75954926	17q25.3	G	0.47	*AC144831.1*	Huyghe, 2019
rs4939827	18q21.1	T	0.35	*SMAD7*	Law, 2019
rs285245	19p13.11	T	0.12	*AC020911.2*	Law, 2019
rs10411210	19q13.11	C	0.26	*RHPN2*	Law, 2019
rs1800469	19q13.2	G	0.37	*TMEM91*	Law, 2019
rs12979278	19q13.33	T	0.24	*MAMSTR*	Law, 2019
rs73068325	19q13.43	T	0.16	*MZF1-AS1*	Huyghe, 2019
rs2423279	20p12.3	C	0.36	*AL031679.1*	Law, 2019
rs28488	20p12.3	T	0.29	*BMP2*	Huyghe, 2019
rs6085661	20p12.3	T	0.30	*Intergenic*	Law, 2019
rs961253	20p12.3	A	0.29	*Intergenic*	Law, 2019
rs994308	20p12.3	C	0.50	*Intergenic*	Huyghe, 2019
rs2295444	20q11.22	C	0.39	*PIGU*	Law, 2019
rs2179593	20q13.12	A	0.32	*TOX2*	Law, 2019
rs6065668	20q13.12	C	0.28	*Intergenic*	Law, 2019
rs1810502	20q13.13	C	0.46	*Intergenic*	Law, 2019
rs4811050	20q13.13	A	0.19	*Intergenic*	Law, 2019
rs6066825	20q13.13	A	0.49	*PREX1*	Law, 2019
rs6091213	20q13.13	C	0.35	*Intergenic*	Law, 2019
rs1741640	20q13.33	C	0.33	*LAMA5*	Law, 2019
rs3787089	20q13.33	C	0.38	*RTEL1*	Law, 2019

MA, minor allele; MAF, minor allele frequency

Statistical power

This study included 128 CRC-risk variants. An approximated Bonferroni corrected α level of 5x10^{-4} was adopted for this study. Similar to the power estimation for the validation study described in section 5.3.1, I used the same group of parameters including a sample size of 5,675 for the SOCCS cohort, the proportion of events (34% for deaths of any causes and 24% for CRC-related deaths) and a range of MAFs (from 0.01 to 0.50) to estimate the power for this study. This had a power of 81% and 60% for overall and CRC-specific survival in the SOCCS cohort in order to detect an effect of 1.25 for 80% (108/128) of the CRC-risk variants. The power curves for this study are plotted in **Figure 5-17**. As suggested by the curves, this study was underpowered (<50%) to identify a small to moderate effect for any rare variants (MAF<0.01).

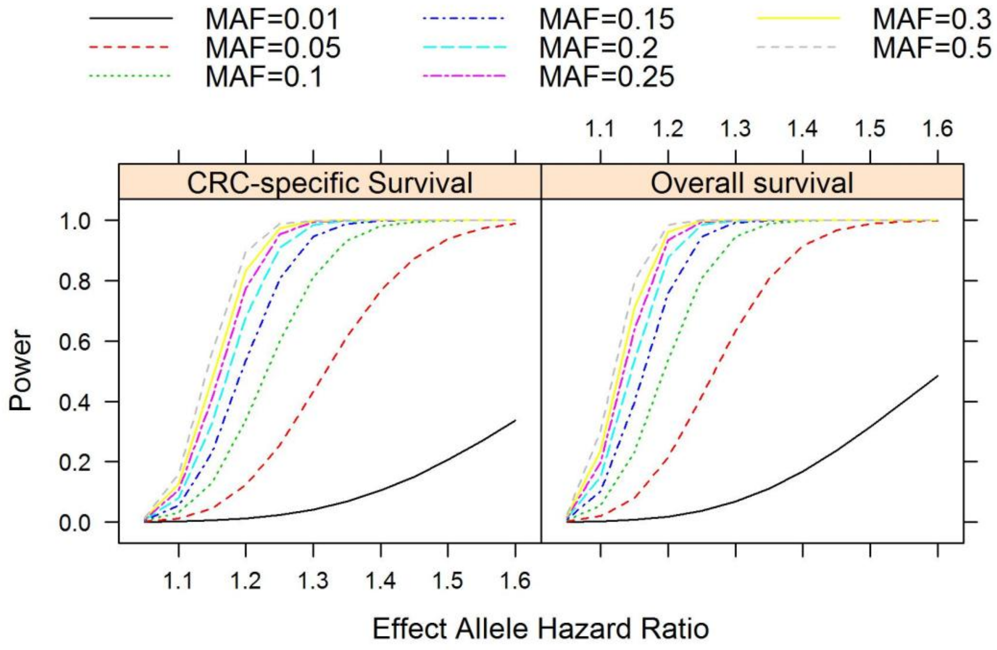


Figure 5-17 Power curves for the two candidate genetic association studies of genetic variants previously linked with CRC risk in the SOCCS cohort

Main analysis

Similar to the procedure mentioned in section 5.3.1, I estimated effect estimates of the 128 genetic variants on CRC survival by fitting Cox regression models adjusting for age at diagnosis, sex and AJCC stage in the SOCCS cohort. Overall, after correcting for multiple testing, none of the 128 variants were significantly associated with either overall or CRC-specific survival (Pfdr<0.05). The effect estimates of the 128 variants on overall survival are summarised in **Table 5-20**. Fifteen genetic variants were found to be associated with overall survival at p<0.05 prior to FDR correction; among them, I identified eight variants of which the CRC-risk increasing alleles showed detrimental effects on overall survival (rs12143541, rs3087967, rs3217810, rs3217874, rs34405347, rs4759277, rs6065668, rs9929218).

Table 5-20 Summary of associations between 128 CRC-risk variants and overall survival of CRC patients in the SOCCS study (N=5,675).

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs6065668	T	0.28	*Intergenic*	0.90(0.83-0.96)	0.003	0.250
rs12143541	G	0.07	*TTC22*	1.13(1.04-1.24)	0.006	0.250
rs34405347	G	0.17	*Intergenic*	0.84(0.74-0.95)	0.007	0.250
rs9537521	A	0.18	*AL139383.1*	1.10(1.02-1.17)	0.008	0.250
rs10994860	T	0.16	*A1CF*	1.11(1.02-1.20)	0.016	0.250
rs3217810	T	0.05	*CCND2*	1.13(1.02-1.25)	0.016	0.250
rs11196171	G	0.50	*TCF7L2*	0.91(0.83-0.98)	0.017	0.250
rs3087967	C	0.27	*C11orf53*	0.92(0.86-0.99)	0.017	0.250
rs10161980	G	0.35	*AL139383.1*	1.08(1.01-1.15)	0.019	0.250
rs174537	T	0.30	*MYRF*	1.08(1.01-1.16)	0.019	0.250
rs16959063	A	0.01	*FMN1*	0.72(0.53-0.97)	0.034	0.378
rs9929218	A	0.26	*CDH1*	0.93(0.86-0.99)	0.035	0.378
rs847208	A	0.33	*LINC01081*	0.93(0.88-1.00)	0.042	0.417
rs3217874	T	0.42	*CCND2*	1.07(1.00-1.14)	0.049	0.437

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs4759277	A	0.37	*LRP1*	1.07(1.00-1.14)	0.050	0.437
rs10951878	T	0.45	*AC004870.4*	0.94(0.88-1.00)	0.054	0.438
rs3184504	C	0.15	*SH2B3*	0.94(0.88-1.00)	0.058	0.438
rs73208120	G	0.05	*NOS1*	1.10(1.00-1.22)	0.061	0.438
rs9831861	G	0.43	*AC096887.1*	1.06(1.00-1.14)	0.065	0.442
rs4919687	A	0.19	*CYP17A1*	1.06(0.99-1.14)	0.076	0.491
rs7398375	G	0.32	*LRP1*	0.93(0.86-1.01)	0.079	0.491
rs4776316	G	0.22	*SMAD6*	0.94(0.87-1.01)	0.110	0.637
rs7593422	T	0.35	*SATB2*	0.95(0.89-1.01)	0.113	0.637
rs35107139	C	0.48	*BMP4*	1.05(0.98-1.13)	0.133	0.641
rs73376930	G	0.32	*GREM1*	1.06(0.98-1.14)	0.136	0.641
rs1570405	A	0.50	*Intergenic*	1.05(0.98-1.13)	0.137	0.641
rs2279290	G	0.11	*LRIG1*	0.94(0.87-1.02)	0.143	0.641
rs2696839	C	0.34	*Intergenic*	0.95(0.90-1.02)	0.143	0.641
rs6066825	G	0.49	*PREX1*	1.05(0.98-1.12)	0.149	0.646
rs2238126	G	0.23	*ETV6*	0.94(0.86-1.03)	0.167	0.702
rs1800469	G	0.37	*TMEM91*	1.05(0.98-1.13)	0.184	0.749
rs6983267	T	0.39	*CASC8*	1.04(0.98-1.11)	0.196	0.749
rs1391441	A	0.34	*TET2*	0.96(0.89-1.02)	0.200	0.749
rs11064437	T	0.16	*TPI1/RPL13P5*	0.48(0.15-1.51)	0.208	0.749
rs3824999	G	0.33	*POLD3*	0.96(0.90-1.02)	0.210	0.749
rs3131043	G	0.49	*HCG20*	1.04(0.98-1.11)	0.220	0.749
rs13020391	T	0.30	*PNKD*	1.04(0.97-1.11)	0.226	0.749
rs17836917	A	0.05	*ASIC2*	1.16(0.91-1.47)	0.226	0.749
rs10936599	T	0.27	*MYNN*	1.05(0.97-1.13)	0.230	0.749
rs983318	A	0.12	*LINC00511*	1.05(0.97-1.13)	0.238	0.754

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs647161	A	0.46	*C5orf66*	1.04(0.97-1.11)	0.251	0.772
rs35509282	A	0.23	*Intergenic*	0.94(0.86-1.04)	0.256	0.772
rs10152518	A	0.37	*Intergenic*	0.96(0.88-1.03)	0.261	0.772
rs12603526	C	0.05	*NXN*	1.13(0.91-1.40)	0.282	0.786
rs61776719	A	0.38	*FHL3*	1.04(0.97-1.10)	0.283	0.786
rs3801081	G	0.25	*TNS3*	0.96(0.90-1.03)	0.284	0.786
rs2186607	A	0.37	*TRPC6*	0.97(0.91-1.03)	0.300	0.800
rs961253	A	0.29	*Intergenic*	0.97(0.91-1.03)	0.315	0.800
rs6085661	T	0.30	*Intergenic*	0.97(0.91-1.03)	0.319	0.800
rs10849438	G	0.17	*Intergenic*	0.95(0.87-1.05)	0.324	0.800
rs78341008	C	0.02	*Intergenic*	0.94(0.83-1.06)	0.332	0.800
rs4313119	T	0.28	*Intergenic*	1.04(0.96-1.12)	0.336	0.800
rs1741640	C	0.33	*LAMA5*	1.04(0.96-1.12)	0.349	0.800
rs62404966	T	0.13	*BMP5*	0.97(0.90-1.04)	0.352	0.800
rs285245	T	0.12	*AC020911.2*	0.95(0.86-1.06)	0.360	0.800
rs12427600	C	0.20	*SMAD9*	1.03(0.96-1.11)	0.361	0.800
rs639933	A	0.31	*C5orf66*	0.97(0.91-1.04)	0.365	0.800
rs72647484	C	0.03	*Intergenic*	0.95(0.85-1.06)	0.372	0.800
rs10849432	T	0.17	*Intergenic*	0.95(0.86-1.06)	0.381	0.800
rs2516420	T	0.06	*HCP5*	0.95(0.83-1.07)	0.387	0.800
rs4944940	A	0.03	*CHRDL2*	0.93(0.78-1.10)	0.388	0.800
rs16892766	C	0.08	*Intergenic*	0.96(0.86-1.06)	0.397	0.806
rs11893063	A	0.30	*AC019330.1*	0.97(0.92-1.04)	0.417	0.828
rs1321310	C	0.28	*Intergenic*	0.97(0.90-1.04)	0.420	0.828
rs10411210	T	0.26	*RHPN2*	1.05(0.93-1.18)	0.435	0.834
rs45597035	G	0.20	*KLF5*	1.03(0.96-1.10)	0.436	0.834

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs17816465	A	0.12	*FMN1*	1.03(0.95-1.12)	0.456	0.847
rs56324967	C	0.39	*SMAD3*	0.97(0.91-1.05)	0.459	0.847
rs1412834	C	0.31	*CDKN2B-AS1*	0.98(0.92-1.04)	0.463	0.847
rs10774214	C	0.45	*CCND2-AS1*	0.98(0.91-1.04)	0.470	0.849
rs10980628	C	0.16	*LPAR1*	1.03(0.95-1.11)	0.480	0.850
rs73975588	C	0.07	*NXN*	0.97(0.88-1.07)	0.492	0.850
rs4811050	A	0.19	*Intergenic*	1.03(0.95-1.12)	0.501	0.850
rs11903757	C	0.12	*Intergenic*	0.97(0.89-1.06)	0.502	0.850
rs1078643	A	0.45	*TMEM238L*	1.03(0.95-1.11)	0.503	0.850
rs11610543	G	0.48	*Intergenic*	1.02(0.96-1.09)	0.515	0.858
rs2735940	G	0.47	*TERT*	1.02(0.96-1.09)	0.534	0.860
rs16878812	G	0.13	*FKBP5*	0.97(0.88-1.07)	0.539	0.860
rs186722897	T	0.10	*Intergenic*	0.96(0.83-1.10)	0.544	0.860
rs6933790	C	0.20	*TFEB*	1.03(0.94-1.13)	0.546	0.860
rs75686861	A	0.03	*HHIP*	1.03(0.93-1.14)	0.549	0.860
rs17035289	C	0.25	*Intergenic*	1.03(0.94-1.12)	0.557	0.862
rs12241008	C	0.19	*VTI1A*	0.97(0.88-1.08)	0.579	0.886
rs7495132	T	0.15	*CRTC3*	1.03(0.93-1.13)	0.602	0.887
rs4711689	A	0.23	*TFEB*	0.98(0.92-1.05)	0.605	0.887
rs12635946	T	0.29	*Intergenic*	1.02(0.95-1.09)	0.608	0.887
rs1035209	T	0.12	*Intergenic*	0.98(0.91-1.06)	0.639	0.887
rs6928864	A	0.29	*Intergenic*	0.97(0.86-1.10)	0.650	0.887
rs2295444	T	0.39	*PIGU*	1.01(0.95-1.08)	0.652	0.887
rs6091213	C	0.35	*Intergenic*	0.98(0.91-1.06)	0.657	0.887
rs73068325	T	0.16	*MZF1-AS1*	1.02(0.94-1.11)	0.665	0.887
rs994308	T	0.50	*Intergenic*	1.01(0.95-1.08)	0.666	0.887

Variant	**MA**	**MAF**	**Gene**	**HR(95%CI)**	**P(uncorrected)**	**Pfdr**
rs7708610	A	0.32	*Intergenic*	0.99(0.92-1.05)	0.673	0.887
rs11632715	A	0.45	*Intergenic*	0.99(0.93-1.05)	0.677	0.887
rs72013726	C	0.48	*Intergenic*	0.99(0.92-1.05)	0.677	0.887
rs77776598	C	0.03	*SLC6A18*	0.97(0.85-1.11)	0.678	0.887
rs12522693	A	0.11	*Intergenic*	0.98(0.89-1.08)	0.679	0.887
rs12979278	T	0.24	*MAMSTR*	1.01(0.95-1.09)	0.684	0.887
rs10795668	A	0.23	*RNA5SP299*	1.01(0.95-1.08)	0.689	0.887
rs704017	G	0.45	*ZMIZ1-AS1*	0.99(0.92-1.06)	0.714	0.911
rs1445011	C	0.13	*Intergenic*	0.99(0.92-1.06)	0.726	0.917
rs75954926	G	0.47	*AC144831.1*	0.99(0.92-1.07)	0.745	0.922
rs16969681	T	0.20	*SCG5*	0.98(0.88-1.10)	0.761	0.922
rs61336918	T	0.39	*Intergenic*	1.01(0.94-1.08)	0.765	0.922
rs2423279	C	0.36	*AL031679.1*	0.99(0.92-1.07)	0.771	0.922
rs72942485	A	0.05	*BOC*	0.96(0.73-1.27)	0.773	0.922
rs1370821	T	0.26	*Intergenic*	1.01(0.95-1.08)	0.775	0.922
rs448513	C	0.50	*TANC1*	0.99(0.93-1.06)	0.781	0.922
rs2179593	A	0.32	*TOX2*	0.99(0.92-1.06)	0.795	0.929
rs3787089	T	0.38	*RTEL1*	1.01(0.94-1.08)	0.818	0.929
rs2070699	T	0.36	*EDN1*	0.99(0.93-1.06)	0.822	0.929
rs1330889	C	0.13	*LINC00446*	1.01(0.92-1.11)	0.824	0.929
rs12672022	C	0.13	*TBRG4*	1.01(0.93-1.10)	0.826	0.929
rs4450168	C	0.08	*SBF2*	0.99(0.91-1.08)	0.829	0.929
rs9271770	A	0.29	*HLA-DQA1*	0.99(0.91-1.08)	0.859	0.954
rs6658977	T	0.21	*LINC02257*	0.99(0.93-1.06)	0.875	0.960
rs10049390	A	0.43	*SLCO2A1*	1.01(0.93-1.09)	0.888	0.960
rs11692435	A	0.02	*ACTR1B*	0.99(0.88-1.12)	0.888	0.960

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs899244	T	0.21	*AC009154.1*	0.99(0.92-1.07)	0.896	0.960
rs28488	T	0.29	*BMP2*	1.00(0.93-1.07)	0.913	0.960
rs1810502	T	0.46	*Intergenic*	1.00(0.94-1.07)	0.914	0.960
rs7993934	T	0.42	*COL4A2*	1.00(0.93-1.07)	0.917	0.960
rs10911251	C	0.37	*LAMC1*	1.00(0.94-1.07)	0.928	0.960
rs17094983	A	0.10	*LINC01500*	1.00(0.91-1.11)	0.935	0.960
rs4939827	C	0.35	*SMAD7*	1.00(0.94-1.06)	0.938	0.960
rs11169552	T	0.25	*ATF1*	1.00(0.93-1.07)	0.958	0.973
rs35360328	A	0.08	*Intergenic*	1.00(0.91-1.09)	0.966	0.973
rs3987	G	0.39	*LINC02264*	1.00(0.94-1.07)	0.989	0.989

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values corrected using the false positive rate approach.

In relation to CRC-specific survival, details of effect estimates of other variants on CRC-specific survival can be found in **Table 5-20**. Ten variants were identified with an uncorrected p<0.05 (**Table 5-20**). Among these 10 variants, the CRC-risk increasing alleles of seven variants were observed to be associated with inferior CRC-specific survival (rs12143541, rs2696839, rs3217810, rs34405347, rs4759277, rs6065668, rs7495132).

Table 5-20 Summary of associations between 128 CRC-risk variants and CRC-specific survival of CRC patients in the SOCCS study (N=5,675).

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs6065668	T	0.28	*Intergenic*	0.89(0.82-0.97)	0.011	0.880
rs10994860	T	0.16	*A1CF*	1.13(1.02-1.24)	0.018	0.836
rs12143541	G	0.07	*TTC22*	1.13(1.02-1.25)	0.023	0.832
rs847208	A	0.33	*LINC01081*	0.92(0.85-0.99)	0.027	0.836
rs4759277	A	0.37	*LRP1*	1.09(1.01-1.18)	0.028	0.836
rs7495132	T	0.15	*CRTC3*	1.13(1.01-1.26)	0.032	0.836

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs35509282	A	0.23	*Intergenic*	0.88(0.78-0.99)	0.033	0.926
rs34405347	G	0.17	*Intergenic*	0.85(0.73-0.99)	0.042	0.836
rs3217810	T	0.05	*CCND2*	1.13(1.00-1.27)	0.044	0.880
rs2696839	C	0.34	*Intergenic*	0.93(0.86-1.00)	0.048	0.832
rs2238126	G	0.23	*ETV6*	0.90(0.82-1.00)	0.053	0.836
rs7398375	G	0.32	*LRP1*	0.91(0.83-1.00)	0.056	0.836
rs3087967	C	0.27	*C11orf53*	0.93(0.86-1.00)	0.062	0.676
rs10951878	T	0.45	*AC004870.4*	0.93(0.87-1.01)	0.073	0.836
rs72647484	C	0.03	*Intergenic*	0.89(0.78-1.01)	0.079	0.603
rs4811050	A	0.19	*Intergenic*	1.09(0.99-1.20)	0.085	0.836
rs12427600	C	0.20	*SMAD9*	1.07(0.98-1.17)	0.123	0.912
rs4776316	G	0.22	*SMAD6*	0.93(0.85-1.02)	0.124	0.832
rs35107139	C	0.48	*BMP4*	1.07(0.98-1.16)	0.131	0.836
rs73376930	G	0.32	*GREM1*	1.07(0.98-1.17)	0.137	0.962
rs3217874	T	0.42	*CCND2*	1.06(0.98-1.14)	0.139	0.836
rs9537521	A	0.18	*AL139383.1*	1.06(0.98-1.15)	0.153	0.836
rs10161980	G	0.35	*AL139383.1*	1.06(0.98-1.14)	0.160	0.836
rs11196171	G	0.50	*TCF7L2*	0.93(0.85-1.03)	0.161	0.603
rs6066825	G	0.49	*PREX1*	1.06(0.98-1.15)	0.176	0.836
rs7593422	T	0.35	*SATB2*	0.95(0.88-1.02)	0.181	0.823
rs12979278	T	0.24	*MAMSTR*	1.06(0.97-1.15)	0.182	0.955
rs2279290	G	0.11	*LRIG1*	0.94(0.85-1.03)	0.200	0.836
rs2186607	A	0.37	*TRPC6*	0.95(0.89-1.03)	0.214	0.836
rs6658977	T	0.21	*LINC02257*	1.05(0.97-1.14)	0.215	0.933
rs1800469	G	0.37	*TMEM91*	1.06(0.97-1.15)	0.217	0.832
rs174537	T	0.30	*MYRF*	1.05(0.97-1.14)	0.218	0.832

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs17836917	A	0.05	*ASIC2*	1.19(0.90-1.59)	0.222	0.836
rs3131043	G	0.49	*HCG20*	1.05(0.97-1.13)	0.231	0.955
rs647161	A	0.46	*C5orf66*	1.05(0.97-1.14)	0.237	0.836
rs13020391	T	0.30	*PNKD*	1.05(0.97-1.14)	0.243	0.967
rs45597035	G	0.20	*KLF5*	1.05(0.97-1.14)	0.256	0.836
rs75686861	A	0.03	*HHIP*	1.07(0.95-1.21)	0.267	0.997
rs3824999	G	0.33	*POLD3*	0.96(0.89-1.03)	0.269	0.832
rs1570405	A	0.50	*Intergenic*	1.05(0.96-1.14)	0.272	0.836
rs16959063	A	0.01	*FMN1*	0.83(0.59-1.16)	0.273	0.836
rs3987	G	0.39	*LINC02264*	1.04(0.97-1.13)	0.280	0.832
rs10849438	G	0.17	*Intergenic*	0.94(0.84-1.05)	0.282	0.836
rs9929218	A	0.26	*CDH1*	0.96(0.88-1.04)	0.294	0.836
rs10152518	A	0.37	*Intergenic*	0.95(0.87-1.05)	0.320	0.836
rs11903757	C	0.12	*Intergenic*	0.95(0.86-1.05)	0.322	0.836
rs899244	T	0.21	*AC009154.1*	0.95(0.87-1.05)	0.322	0.832
rs72942485	A	0.05	*BOC*	1.17(0.85-1.61)	0.331	0.836
rs73208120	G	0.05	*NOS1*	1.06(0.94-1.20)	0.336	0.832
rs2516420	T	0.06	*HCP5*	0.93(0.80-1.08)	0.343	0.832
rs1412834	C	0.31	*CDKN2B-AS1*	0.97(0.89-1.04)	0.354	0.955
rs3801081	G	0.25	*TNS3*	0.96(0.89-1.04)	0.355	0.836
rs1035209	T	0.12	*Intergenic*	0.96(0.87-1.05)	0.359	0.961
rs1741640	C	0.33	*LAMA5*	1.04(0.95-1.15)	0.377	0.836
rs17094983	A	0.10	*LINC01500*	1.05(0.93-1.19)	0.409	0.832
rs6983267	T	0.39	*CASC8*	1.03(0.96-1.11)	0.409	0.603
rs10936599	T	0.27	*MYNN*	1.04(0.95-1.13)	0.411	0.832
rs35360328	A	0.08	*Intergenic*	1.05(0.94-1.16)	0.415	0.836

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs994308	T	0.50	*Intergenic*	1.03(0.96-1.11)	0.416	0.875
rs12635946	T	0.29	*Intergenic*	1.03(0.95-1.12)	0.419	0.836
rs186722897	T	0.10	*Intergenic*	0.93(0.79-1.10)	0.420	0.603
rs16878812	G	0.13	*FKBP5*	0.95(0.85-1.07)	0.433	0.955
rs10795668	A	0.23	*RNA5SP299*	0.97(0.89-1.05)	0.446	0.836
rs73068325	T	0.16	*MZF1-AS1*	1.04(0.94-1.14)	0.446	0.836
rs9831861	G	0.43	*AC096887.1*	1.03(0.95-1.11)	0.453	0.616
rs6085661	T	0.30	*Intergenic*	0.97(0.90-1.05)	0.456	0.832
rs961253	A	0.29	*Intergenic*	0.97(0.90-1.05)	0.475	0.836
rs11692435	A	0.02	*ACTR1B*	0.95(0.82-1.10)	0.476	0.603
rs17816465	A	0.12	*FMN1*	1.03(0.94-1.14)	0.483	0.823
rs4450168	C	0.08	*SBF2*	0.96(0.87-1.07)	0.489	0.603
rs285245	T	0.12	*AC020911.2*	0.96(0.85-1.08)	0.498	0.823
rs6933790	C	0.20	*TFEB*	1.04(0.93-1.16)	0.504	0.836
rs7993934	T	0.42	*COL4A2*	0.97(0.90-1.05)	0.504	0.603
rs10774214	C	0.45	*CCND2-AS1*	0.97(0.90-1.06)	0.505	0.836
rs12603526	C	0.05	*NXN*	0.91(0.68-1.21)	0.505	0.836
rs28488	T	0.29	*BMP2*	0.97(0.90-1.06)	0.522	0.832
rs12241008	C	0.19	*VTI1A*	0.96(0.85-1.09)	0.533	0.832
rs11064437	T	0.16	*TPI1/RPL13P5*	0.70(0.22-2.19)	0.535	0.997
rs4939827	C	0.35	*SMAD7*	1.02(0.95-1.11)	0.536	0.836
rs11893063	A	0.30	*AC019330.1*	0.98(0.91-1.05)	0.542	0.961
rs2179593	A	0.32	*TOX2*	0.97(0.90-1.06)	0.553	0.832
rs11610543	G	0.48	*Intergenic*	1.02(0.95-1.10)	0.557	0.997
rs4919687	A	0.19	*CYP17A1*	1.02(0.94-1.11)	0.561	0.603
rs17035289	C	0.25	*Intergenic*	0.97(0.87-1.08)	0.570	0.823

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs10980628	C	0.16	*LPAR1*	1.03(0.94-1.12)	0.577	0.692
rs1370821	T	0.26	*Intergenic*	0.98(0.91-1.06)	0.582	0.836
rs4944940	A	0.03	*CHRDL2*	0.94(0.77-1.16)	0.583	0.836
rs77776598	C	0.03	*SLC6A18*	0.96(0.82-1.12)	0.596	0.836
rs61336918	T	0.39	*Intergenic*	1.02(0.94-1.11)	0.604	0.836
rs639933	A	0.31	*C5orf66*	0.98(0.91-1.06)	0.612	0.603
rs3787089	T	0.38	*RTEL1*	1.02(0.94-1.11)	0.615	0.832
rs10411210	T	0.26	*RHPN2*	1.04(0.90-1.20)	0.617	0.836
rs6091213	C	0.35	*Intergenic*	0.98(0.90-1.07)	0.621	0.836
rs1321310	C	0.28	*Intergenic*	0.98(0.90-1.07)	0.639	0.836
rs16892766	C	0.08	*Intergenic*	1.03(0.91-1.16)	0.643	0.955
rs3184504	C	0.15	*SH2B3*	0.98(0.91-1.06)	0.644	0.871
rs56324967	C	0.39	*SMAD3*	0.98(0.90-1.07)	0.649	0.836
rs16969681	T	0.20	*SCG5*	1.03(0.91-1.17)	0.650	0.832
rs6928864	A	0.29	*Intergenic*	0.97(0.83-1.12)	0.655	0.832
rs10911251	C	0.37	*LAMC1*	1.02(0.94-1.10)	0.656	0.836
rs2295444	T	0.39	*PIGU*	1.02(0.94-1.10)	0.663	0.836
rs983318	A	0.12	*LINC00511*	1.02(0.93-1.11)	0.675	0.836
rs62404966	T	0.13	*BMP5*	0.98(0.90-1.07)	0.704	0.935
rs7708610	A	0.32	*Intergenic*	0.99(0.91-1.07)	0.710	0.880
rs2423279	C	0.36	*AL031679.1*	0.98(0.90-1.08)	0.720	0.683
rs10849432	T	0.17	*Intergenic*	1.02(0.90-1.16)	0.736	0.836
rs10049390	A	0.43	*SLCO2A1*	1.02(0.93-1.11)	0.741	0.836
rs72013726	C	0.48	*Intergenic*	1.01(0.94-1.09)	0.745	0.836
rs75954926	G	0.47	*AC144831.1*	0.99(0.90-1.08)	0.769	0.823
rs11169552	T	0.25	*ATF1*	1.01(0.93-1.10)	0.786	0.961

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs1078643	A	0.45	*TMEM238L*	0.99(0.90-1.08)	0.805	0.603
rs12672022	C	0.13	*TBRG4*	1.01(0.91-1.12)	0.818	0.603
rs704017	G	0.45	*ZMIZ1-AS1*	0.99(0.92-1.07)	0.827	0.832
rs78341008	C	0.02	*Intergenic*	0.99(0.85-1.14)	0.836	0.832
rs12522693	A	0.11	*Intergenic*	0.99(0.89-1.10)	0.864	0.900
rs2735940	G	0.47	*TERT*	1.01(0.93-1.09)	0.868	0.871
rs61776719	A	0.38	*FHL3*	1.01(0.93-1.09)	0.881	0.836
rs1330889	C	0.13	*LINC00446*	0.99(0.88-1.11)	0.883	0.937
rs1810502	T	0.46	*Intergenic*	0.99(0.92-1.07)	0.888	0.836
rs448513	C	0.50	*TANC1*	1.00(0.92-1.08)	0.907	0.603
rs73975588	C	0.07	*NXN*	1.01(0.90-1.13)	0.912	0.836
rs2070699	T	0.36	*EDN1*	1.00(0.92-1.08)	0.917	0.997
rs11632715	A	0.45	*Intergenic*	1.00(0.93-1.08)	0.925	0.832
rs1391441	A	0.34	*TET2*	1.00(0.92-1.09)	0.937	0.836
rs1445011	C	0.13	*Intergenic*	1.00(0.92-1.09)	0.981	0.836
rs4313119	T	0.28	*Intergenic*	1.00(0.91-1.10)	0.986	0.844
rs4711689	A	0.23	*TFEB*	1.00(0.93-1.08)	0.991	0.836
rs9271770	A	0.29	*HLA-DQA1*	1.00(0.91-1.10)	0.997	0.836

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values corrected using the false positive rate approach.

I also created a polygenic risk score (PRS) for each individual in the SOCCS cohort by adding up the number of CRC-risk increasing alleles of the 128 genetic variants. Here an assumption was made that each CRC-risk increasing allele carried the same detrimental prognostic effect on CRC survival. On average, each CRC patient in the cohort carries 34.9 (standard deviation=6.7) CRC-risk increasing alleles. As shown in **Figure 5-18**, the distribution of the PRS of CRC-risk variants is approximately normal. I then fitted a Cox regression model with age, sex and AJCC as covariates to investigate the association between the CRC-risk PRS and survival outcomes. The results suggested neither overall (HR=1.00, 95%CI=0.95-1.04, p=0.864) nor CRC-specific survival (HR=1.03, 95%CI=0.97-1.08, p=0.340) was significantly associated with the CRC-risk PRS.

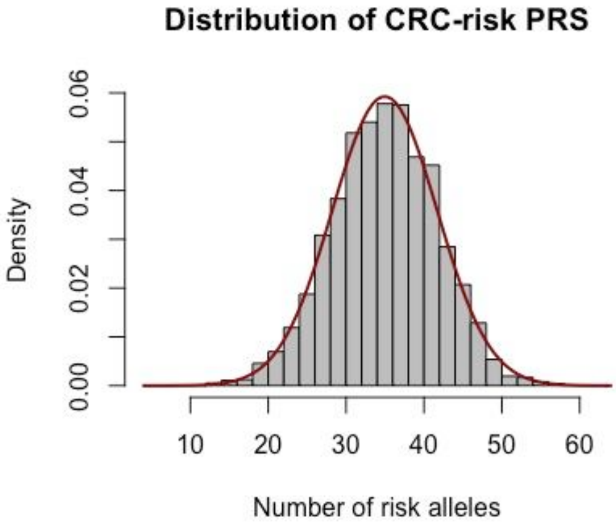


Figure 5-18 Distribution of polygenic risk score of variants associated with CRC risk in the SOCCS study

Stratified analysis

I investigated additive genetic effects of the 128 CRC-risk variants stratified by sex, AJCC stage and tumour site. Multiple testing was corrected separately using the FDR approach within each stratum. Stratified by sex, no significant associations were observed between any of the 128 variants and survival outcomes of CRC after FDR correction, although a number of suggestive associations with uncorrected p<0.05 were found. To be specific, I detected eight associations between genetic variants and overall survival in male patients and six variants were associated with CRC-specific survival (uncorrected p<0.05). In female patients, nine genetic CRC-risk variants were associated with overall survival and 12 variants associated with CRC-specific survival. I summarised the effect estimates along with the uncorrected and

FDR-corrected p-values in **Table 5-21**. Additional details regarding the remaining variants with uncorrected p>0.05 are presented in **Appendix Table 10**.

Table 5-21 Summary of associations (p<0.05) between CRC-risk variants and survival of CRC patients in the SOCCS study stratified by sex

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
Male (N=3,235)						
OS						
	rs12143541	G	0.07	1.17(1.04-1.31)	0.009	0.592
	rs34405347	G	0.17	0.81(0.69-0.96)	0.015	0.592
	rs11196171	G	0.5	0.88(0.79-0.98)	0.018	0.592
	rs961253	A	0.29	0.90(0.83-0.98)	0.021	0.592
	rs647161	A	0.46	1.11(1.01-1.21)	0.023	0.592
	rs4776316	G	0.22	0.90(0.82-1.00)	0.043	0.629
	rs17836917	A	0.05	1.35(1.01-1.80)	0.045	0.629
	rs9831861	G	0.43	1.09(1.00-1.19)	0.046	0.629
CSS						
	rs7495132	T	0.15	1.21(1.05-1.40)	0.008	0.791
	rs647161	A	0.46	1.13(1.02-1.26)	0.019	0.791
	rs7993934	T	0.42	0.89(0.80-0.99)	0.033	0.791
	rs10994860	T	0.16	1.14(1.00-1.30)	0.044	0.791
	rs35509282	A	0.23	0.85(0.73-1.00)	0.047	0.791
	rs4811050	A	0.19	1.13(1.00-1.28)	0.049	0.791
Female (N=2,440)						
OS						
	rs3087967	C	0.27	0.86(0.77-0.95)	0.005	0.417
	rs9537521	A	0.18	1.16(1.04-1.29)	0.007	0.417
	rs10161980	G	0.35	1.15(1.03-1.27)	0.010	0.417

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
	rs3217810	T	0.05	1.21(1.04-1.40)	0.016	0.459
	rs3131043	G	0.49	1.13(1.02-1.26)	0.018	0.459
	rs6065668	T	0.28	0.88(0.78-0.98)	0.024	0.519
	rs3217874	T	0.42	1.12(1.01-1.24)	0.036	0.569
	rs9929218	A	0.26	0.89(0.79-0.99)	0.038	0.569
	rs28488	T	0.29	0.89(0.80-1.00)	0.047	0.569
CSS						
	rs3131043	G	0.49	1.18(1.04-1.33)	0.008	0.397
	rs6065668	T	0.28	0.84(0.73-0.96)	0.011	0.397
	rs13020391	T	0.3	1.17(1.04-1.33)	0.012	0.397
	rs3087967	C	0.27	0.86(0.76-0.97)	0.018	0.397
	rs847208	A	0.33	0.86(0.77-0.98)	0.018	0.397
	rs10161980	G	0.35	1.16(1.02-1.31)	0.019	0.397
	rs28488	T	0.29	0.86(0.76-0.98)	0.021	0.397
	rs9537521	A	0.18	1.16(1.02-1.31)	0.026	0.423
	rs10980628	C	0.16	1.16(1.01-1.33)	0.037	0.519
	rs72647484	C	0.03	0.80(0.64-0.99)	0.041	0.519
	rs35107139	C	0.48	1.14(1.00-1.30)	0.044	0.519
	rs4759277	A	0.37	1.13(1.00-1.28)	0.048	0.520

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

As for stage-stratified analysis, no statistically significant associations survived FDR correction. Nevertheless, I observed 12 CRC-risk variants associated with overall survival with uncorrected p-values<0.05, and seven variants for CRC-specific survival for stage II/III. With respect to stage IV patients, seven CRC-risk variants were related to overall survival and six variants were observed to be associated with CRC-specific survival. Summarised results of associations with uncorrected p<0.05 in stage II/III and stage IV patients are presented in **Table 5-22**. A full list of results of the 128 CRC variants can be found in **Appendix Table 11**.

Table 5-22 Summary of associations (p<0.05) between CRC-risk variants and survival of CRC patients in the SOCCS study stratified by stage

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
Stage II/III (N=3,886)						
OS						
	rs9537521	A	0.18	1.13(1.04-1.23)	0.005	0.236
	rs10161980	G	0.35	1.13(1.03-1.22)	0.006	0.236
	rs6065668	T	0.28	0.88(0.80-0.96)	0.006	0.236
	rs847208	A	0.33	0.89(0.82-0.97)	0.008	0.236
	rs1800469	G	0.37	1.13(1.03-1.24)	0.009	0.236
	rs3087967	C	0.27	0.90(0.83-0.98)	0.021	0.452
	rs983318	A	0.12	1.11(1.01-1.22)	0.036	0.516
	rs1570405	A	0.50	1.09(1.00-1.19)	0.040	0.516
	rs10951878	T	0.45	0.92(0.85-1.00)	0.047	0.516
	rs3131043	G	0.49	1.09(1.00-1.19)	0.047	0.516
	rs11196171	G	0.50	0.90(0.81-1.00)	0.048	0.516
	rs17816465	A	0.12	1.11(1.00-1.22)	0.048	0.516
CSS						
	rs3087967	C	0.27	0.86(0.78-0.96)	0.005	0.520
	rs847208	A	0.33	0.87(0.79-0.97)	0.010	0.520
	rs6065668	T	0.28	0.86(0.77-0.97)	0.013	0.520
	rs1800469	G	0.37	1.15(1.02-1.29)	0.019	0.520
	rs1570405	A	0.50	1.14(1.02-1.27)	0.020	0.520

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
	rs2696839	C	0.34	0.90(0.81-0.99)	0.036	0.782
	rs4759277	A	0.37	1.11(1.00-1.23)	0.044	0.786
Stage IV (N=784)						
OS						
	rs10994860	T	0.16	1.22(1.04-1.43)	0.015	0.639
	rs72647484	C	0.03	0.78(0.63-0.96)	0.021	0.639
	rs3801081	G	0.25	0.87(0.77-0.99)	0.028	0.639
	rs73208120	G	0.05	1.23(1.01-1.50)	0.036	0.639
	rs12143541	G	0.07	1.18(1.01-1.37)	0.042	0.639
	rs2238126	G	0.23	0.85(0.72-1.00)	0.045	0.639
	rs847208	A	0.33	0.88(0.77-1.00)	0.045	0.639
CSS						
	rs72647484	C	0.03	0.76(0.61-0.95)	0.016	0.703
	rs10994860	T	0.16	1.21(1.03-1.43)	0.019	0.703
	rs3217810	T	0.05	1.23(1.01-1.50)	0.035	0.703
	rs3801081	G	0.25	0.88(0.77-0.99)	0.039	0.703
	rs12143541	G	0.07	1.18(1.01-1.38)	0.040	0.703
	rs3987	G	0.39	1.14(1.00-1.30)	0.042	0.703

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

I also conducted association analyses in colon and rectal cancer patients separately. A total of 3,392 patients diagnosed with colon cancer were included in analysis. Overall, I detected no statistically significant signals after FDR correction. Twenty-six unique genetic variants were associated with either overall or CRC-specific survival of colon cancer patients (uncorrected p<0.05); among them, 12 variants were related to both outcomes with the same direction of effect. Detailed results are presented in **Table 5-23.**

.

Table 5-23 Summary of associations (p<0.05) between CRC-risk variants and survival of colon cancer patients (N=3,392) in the SOCCS study

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
OS						
	rs10994860	T	0.16	1.16(1.05-1.29)	0.005	0.120
	rs2696839	C	0.34	0.89(0.82-0.96)	0.005	0.120
	rs34405347	G	0.17	0.77(0.65-0.92)	0.005	0.120
	rs6065668	T	0.28	0.87(0.79-0.96)	0.005	0.120
	rs11196171	G	0.5	0.86(0.77-0.96)	0.006	0.120
	rs174537	T	0.3	1.13(1.04-1.24)	0.006	0.120
	rs3087967	C	0.27	0.88(0.81-0.97)	0.006	0.120
	rs3217874	T	0.42	1.12(1.03-1.22)	0.008	0.125
	rs16959063	A	0.01	0.55(0.35-0.87)	0.010	0.125
	rs3801081	G	0.25	0.89(0.81-0.97)	0.011	0.125
	rs847208	A	0.33	0.89(0.82-0.97)	0.011	0.125
	rs3217810	T	0.05	1.18(1.04-1.35)	0.012	0.125
	rs35107139	C	0.48	1.12(1.02-1.23)	0.015	0.153
	rs9929218	A	0.26	0.90(0.82-0.99)	0.024	0.222
	rs12143541	G	0.07	1.14(1.02-1.28)	0.027	0.231
	rs9537521	A	0.18	1.10(1.01-1.21)	0.031	0.252
	rs4759277	A	0.37	1.10(1.01-1.19)	0.034	0.259
	rs4919687	A	0.19	1.10(1.00-1.21)	0.040	0.288
	rs3184504	C	0.15	0.92(0.84-1.00)	0.045	0.306
	rs11610543	G	0.48	1.09(1.00-1.18)	0.048	0.31
CSS						
	rs10849438	G	0.17	0.82(0.70-0.97)	0.018	0.301

Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
rs10994860	T	0.16	1.20(1.06-1.37)	0.004	0.225
rs11196171	G	0.50	0.88(0.77-1.00)	0.048	0.371
rs11893063	A	0.30	0.90(0.81-0.99)	0.030	0.348
rs12143541	G	0.07	1.15(1.00-1.31)	0.045	0.371
rs16959063	A	0.01	0.54(0.32-0.93)	0.025	0.328
rs2696839	C	0.34	0.87(0.79-0.96)	0.006	0.225
rs3087967	C	0.27	0.86(0.78-0.96)	0.006	0.225
rs3217810	T	0.05	1.20(1.03-1.41)	0.020	0.301
rs3217874	T	0.42	1.11(1.01-1.23)	0.034	0.348
rs35509282	A	0.23	0.84(0.72-0.99)	0.038	0.348
rs3801081	G	0.25	0.88(0.79-0.97)	0.015	0.301
rs4759277	A	0.37	1.15(1.04-1.27)	0.007	0.225
rs6065668	T	0.28	0.88(0.78-0.98)	0.021	0.301
rs6066825	G	0.49	1.12(1.01-1.24)	0.036	0.348
rs7495132	T	0.15	1.16(1.00-1.33)	0.048	0.371
rs847208	A	0.33	0.88(0.79-0.97)	0.011	0.285

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

Pertaining to analysis in rectal cancer patients (N=2,201), none of the 128 variants remained statistically significant after FDR correction. Five variants were identified with uncorrected p-values<0.05. In particular, the G allele of two variants (rs2238126 and rs2279290) exhibited potential protective effects on overall survival (rs2238126: HR=0.85, 95%CI=0.74-0.98, uncorrected p=0.026, Pfdr=0.868; rs2279290: HR=0.84, 95%CI=0.74-0.96, uncorrected p=0.008, Pfdr=0.834) as well as CRC-specific survival (rs2238126: HR=0.82, 95%CI=0.69-0.97, uncorrected p=0.020, Pfdr=0.994; rs2279290: HR=0.81, 95%CI=0.69-0.95, uncorrected p=0.009, Pfdr=0.994). In addition, the T allele of variant rs7593422 was associated with better overall survival of rectal cancer patients (HR=0.88, 95%CI=0.80-0.97, uncorrected p=0.013, Pfdr=0.834), and I also observed an association between variant rs7993934 (T allele) and improved CRC-specific survival (HR=0.87, 95%CI=0.77-0.99, uncorrected p=0.031, Pfdr=0.994). An additional variant (rs16969681, effect allele: T) was found to be associated with inferior CRC-specific survival (HR=1.22, 95%CI=1.02-1.47, uncorrected p=0.034, Pfdr=0.994). The full list of results of site-stratified analysis is presented in **Appendix Table 12**.

Sensitivity analysis

I examined associations between CRC-risk variants and survival outcomes under a recessive model as a sensitivity analysis. Three associations were statistically significant after FDR correction. In particular, the GG genotype of the variant rs10161980 was significantly associated with inferior overall survival (**Table 5-24**). A similar detrimental effect was observed for the AA genotype of the variant rs9537521 on overall survival. It is worth mentioning that the GG and AA genotypes of these two variants were previously linked with reduced CRC risk. With respect to CRC-specific survival, patients carrying the TT genotype of rs7495132 had significantly worse survival compared with ones with TC or CC genotypes. The TT genotype of this variant was associated with increased CRC risk in the previous GWAS. Other variants associated with CRC survival outcomes under recessive pattern at nominal significance (p<0.05) are summarised in **Table 5-24**. The full set of results of all the 128 variants under a recessive model is presented in **Appendix Table 13**.

Table 5-24 Summary of associations (p<0.05) between CRC-risk variants and survival outcomes of CRC patients in the SOCCS study under a recessive model (N=5,675)

	Variant	MG	MGF	HR(95%CI)	P(uncorrected)	Pfdr
OS						
	rs9537521	AA	0.03	1.25(1.11-1.41)	2.50E-04	0.022
	rs10161980	GG	0.12	1.24(1.10-1.39)	3.40E-04	0.022
	rs174537	TT	0.09	1.23(1.07-1.41)	0.003	0.091

	Variant	MG	MGF	HR(95%CI)	P(uncorrected)	Pfdr
	rs6066825	GG	0.24	1.22(1.07-1.40)	0.003	0.091
	rs73975588	CC	0.01	0.55(0.35-0.86)	0.009	0.235
	rs3087967	CC	0.07	0.89(0.82-0.98)	0.014	0.311
	rs3217810	TT	0.003	1.42(1.06-1.92)	0.021	0.380
	rs78341008	CC	0.001	1.79(1.08-2.98)	0.024	0.395
	rs35509282	AA	0.05	0.61(0.40-0.96)	0.030	0.437
	rs10951878	TT	0.20	0.89(0.80-0.99)	0.035	0.454
	rs35360328	AA	0.01	1.35(1.01-1.80)	0.041	0.467
	rs7495132	TT	0.02	1.40(1.01-1.93)	0.045	0.467
	rs16892766	CC	0.01	0.55(0.30-1.00)	0.049	0.467
CSS						
	rs7495132	TT	0.02	1.97(1.41-2.74)	6.10E-05	0.008
	rs6066825	GG	0.24	1.30(1.11-1.52)	0.001	0.063
	rs10161980	GG	0.12	1.22(1.06-1.40)	0.005	0.196
	rs9537521	AA	0.03	1.22(1.06-1.41)	0.006	0.201
	rs4811050	AA	0.04	1.38(1.08-1.78)	0.012	0.301
	rs35509282	AA	0.05	0.52(0.30-0.90)	0.019	0.414
	rs10951878	TT	0.20	0.87(0.76-0.99)	0.030	0.492
	rs3217810	TT	0.003	1.47(1.04-2.09)	0.030	0.492
	rs13020391	TT	0.09	1.17(1.01-1.35)	0.037	0.536

MG, minor genotype; MGF, minor genotype frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival.

I then conducted replication analysis of the two significant signals using the UK Biobank cohort (N=2,474) by fitting Cox regression models adjusting for age at diagnosis and sex. The GG genotype of the variant rs10161980 was not significantly associated with overall survival of CRC patients in the UK Biobank cohort (HR=1.21, 95%CI=0.99-1.46), p=0.057). However, I observed a significant association between the GG genotype of rs10161980 and CRC-specific survival in the UK Biobank (HR=1.26, 95%CI=1.01-1.56, p=0.040). As for the other variant rs7495132, I replicated the observed association between the TT genotype and CRC-specific survival in the UK Biobank (HR=1.69, 95%CI=1.03-2.79, p=0.038), although no significant effect was detected of this variant on overall survival (HR=1.39, 95%CI=0.86-2.25, p=0.179). I was unable to map the variant rs9537521 or a proxy in the UK Biobank cohort as this variant had not been arrayed by the UK Biobank and not included in the 1000 Genome Reference Panel.

5.3.4 Variants associated with survival outcomes of other cancers

Eligible variants

I obtained 18 studies with a total of 113 autosomal genetic variants associated at $p< 5x10^{-5}$ with other solid tumour survival from the initial search in the GWAS catalogue (Azad *et al*, 2016; Ghesquieres *et al*, 2015; Guo *et al*, 2015; Khan *et al*, 2018; Khan *et al*, 2015; Koster *et al*, 2018; Moore *et al*, 2017; Rafiq *et al*, 2014; Sato *et al*, 2011; Shu *et al*, 2012; Song *et al*, 2015; Szulkin *et al*, 2015; Tang *et al*, 2017; Tang *et al*, 2015; Wu *et al*, 2013; Wu *et al*, 2010; Yoon *et al*, 2014; Ziv *et al*, 2015). Among them, 31 variants with larger p-values were excluded as they were in linkage disequilibrium with other included variants ($r^2<0.2$). Finally, a total of 82 genetic variants were included for this study. These variants were reportedly associated with survival outcomes of eleven types of cancer—breast cancer (number of variants n=13), prostate cancer (n=2), non-small cell lung cancer (n=7), small cell lung cancer (n=1), ovarian cancer (n=10), diffuse large B-cell lymphoma (n=2), oesophageal squamous cell cancer (n=1), head and neck cancer (n=1), multiple myeloma (n=1), pancreatic cancer (n=35) and osteosarcoma (n=9). The RSIDs of the variants for each cancer type are summarised in **Table 5-25**. Among these included variants, seven variants (rs2059614, rs72773978, rs148760487, rs2314686, rs7701292, rs763780, rs1050631) were identified to be associated with cancer survival at the level of GWAS significance ($p<5x10^{-8}$). With regard to the reported survival outcomes, four variants were linked to disease-free survival of cancers; 66 variants were identified to be associated with overall survival and six variants were reported using cancer-specific survival as the primary outcome (**Table 5-25**). Notably, the time-to-recurrence was adopted as the primary outcome for one variant associated with non-small cell lung cancer survival. Ghesquieres et al. defined a new outcome of event-free survival—primary events included disease progression, relapse, re-treatment, or death from any cause—in their GWAS investigating survival outcomes of diffuse large B-cell lymphoma (Ghesquieres *et al*, 2015).

Table 5-25 Summary details of the included genetic variants previously associated survival outcomes of other cancers

Variant	locus	MA	MAF	Gene	Reported outcomes	Reference
rs10736390	1p32.3	A	0.46	*MROH7*	Pancreatic cancer(OS)	Tang,2017
rs1391315	1p34.2	G	0.1	*SMAP2*	Pancreatic cancer(OS)	Tang,2017
rs16861827	1p36.13	T	0.14	*IGSF21*	Pancreatic cancer(OS)	Tang,2017
rs1567532	2p12	T	0.14	*CTNNA2*	Pancreatic cancer(OS)	Tang,2017
rs113988120	2p13.3	A	0.01	*PAIP2B*	Pancreatic cancer(OS)	Tang,2017
rs12620038	2p21	G	0.45	*EPCAM-DT*	Pancreatic cancer(OS)	Tang,2017
rs114997855	2p23.1	A	0.03	*Intergenic*	Prostate cancer(CSS)	Szulkin,2015
rs148760487	2q24.3	G	0.01	*Intergenic*	Breast cancer(CSS)	Guo Q,2015
rs1656402	2q37.1	T	0.33	*EIF4E2*	Non-small cell lung cancer(OS)	Sato,2011
rs76010824	3p14.1	A	0.06	*SUCLG2*	Prostate cancer(CSS)	Szulkin,2015
rs4955138	3p22.3	G	0.31	*Intergenic*	Osteosarcoma(OS)	Koster,2018
rs361052	3p25.2	A	0.27	*IQSEC1*	Pancreatic cancer(OS)	Tang,2017
rs770996	3p26.2	T	0.35	*AC034195.1*	Pancreatic cancer(OS)	Tang,2017
rs4568126	3q13.32	A	0.33	*B4GALT4*	Pancreatic cancer(OS)	Tang,2017
rs295315	3q23	G	0.37	*Intergenic*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs6797464	3q26.2	A	0.12	*MECOM*	Osteosarcoma(OS)	Koster,2018
rs17248137	4q13.3	G	0.05	*Intergenic*	Osteosarcoma(OS)	Koster,2018
rs11733008	4q22.1	T	0.36	*Intergenic*	Non-small cell lung cancer(OS)	Tang,2015
rs10023113	4q26	G	0.17	*CAMK2D*	Non-small cell lung cancer(OS)	Tang,2015
rs17305086	4q34.3	T	0.11	*TENM3-AS1*	Non-small cell lung cancer(OS)	Tang,2015
rs1454694	4q34.3	C	0.17	*Intergenic*	Non-small cell lung cancer(recurrence)	Yoon,2014
rs421379	5q14.3	T	0.28	*Intergenic*	Breast cancer(OS)	Rafiq,2014
rs7701292	5q21.3	C	0.08	*Intergenic*	Breast cancer (ER+) (DFS)	Khan,2018
rs7712513	5q23.2	G	0.26	*Intergenic*	Diffuse large B-cell lymphoma(EFS)	Ghesquieres,2015
rs4285214	5q23.2	T	0.39	*ZNF608*	Pancreatic cancer(OS)	Tang,2017
rs763780	6p12.2	C	0.09	*IL17F*	Pancreatic cancer(OS)	Tang,2017
rs12209785	6p21.1	G	0.15	*RUNX2*	Pancreatic cancer(OS)	Tang,2017

Variant	locus	MA	MAF	Gene	Reported outcomes	Reference
rs4618572	6p25.3	T	0.1	*Intergenic*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs7765004	6q21	C	0.32	*Intergenic*	Diffuse large B-cell lymphoma(EFS)	Ghesquieres,2015
rs7777171	7p21.2	C	0.47	*AGMO*	Osteosarcoma(OS)	Koster,2018
rs2299187	7q21.11	T	0.07	*CACNA2D1*	Head and neck cancer(OS)	Azad,2016
rs2314686	8p21.2	A	0.13	*SLC25A37*	Breast cancer (ER+) (DFS)	Khan,2018
rs4382459	8q13.2	T	0.08	*PREX2*	Pancreatic cancer(OS)	Tang,2017
rs202280	8q21.13	G	0.04	*Intergenic*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs6986444	8q24.12	T	0.33	*SNTB1*	Osteosarcoma(OS)	Koster,2018
rs55933544	9p24.1	T	0.2	*GLDC*	Osteosarcoma(OS)	Koster,2018
rs823920	9q31.1	G	0.12	*Intergenic*	Pancreatic cancer(OS)	Tang,2017
rs10817611	9q32	C	0.23	*WHRN*	Pancreatic cancer(OS)	Tang,2017
rs10983614	9q33.1	C	0.28	*ASTN2*	Pancreatic cancer(OS)	Tang,2017
rs1414153	9q33.1	C	0.23	*Dec-01*	Pancreatic cancer(OS)	Tang,2017
rs1564271	10p12.1	A	0.23	*PDSS1*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs12358475	10p14	A	0.11	*Intergenic*	Breast cancer(OS)	Rafiq,2014
rs10825036	10q21.1	G	0.19	*Intergenic*	Breast cancer(DFS)	Song,2015
rs17465450	10q22.3	C	0.03	*LRMDA*	Osteosarcoma(OS)	Koster,2018
rs17693104	10q23.1	T	0.35	*SH2D4B*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs1408536	10q26.3	A	0.12	*Intergenic*	Pancreatic cancer(OS)	Tang,2017
rs10767646	11p14.1	T	0.33	*BDNF-AS*	Pancreatic cancer(OS)	Tang,2017
rs10835188	11p14.1	G	0.36	*LIN7C*	Pancreatic cancer(OS)	Tang,2017
rs4150579	11p15.1	A	0.33	*GTF2H1*	Pancreatic cancer(OS)	Tang,2017
rs10500780	11p15.3	A	0.14	*BTBD10*	Serous epithelial ovarian cancer(PFS)	Moore,2017
rs12362504	11p15.4	C	0.38	*SBF2*	Pancreatic cancer(OS)	Tang,2017
rs10899426	11q14.1	C	0.02	*Intergenic*	Serous epithelial ovarian cancer(PFS)	Moore,2017
rs1944782	11q21	G	0.21	*Intergenic*	Pancreatic cancer(OS)	Tang,2017
rs716274	11q22.3	G	0.44	*Intergenic*	Small cell lung cancer(OS)	Wu,2010
rs2059614	11q24.2	G	0.04	*PKNOX2*	Breast cancer(CSS)	Guo Q,2015
rs2900174	12p13.2	G	0.16	*PRB2*	Pancreatic cancer(OS)	Tang,2017
rs11062040	12p13.33	C	0.46	*DCP1B*	Pancreatic cancer(OS)	Tang,2017
rs17548007	12q23.2	T	0.04	*Intergenic*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs12146774	12q24.23	T	0.18	*AC084880.2*	Osteosarcoma(OS)	Koster,2018
rs1352757	13q21.31	A	0.43	*Intergenic*	Pancreatic cancer(OS)	Tang,2017
rs9593831	13q31.1	T	0.22	*Intergenic*	Pancreatic cancer(OS)	Tang,2017
rs9517906	13q32.3	A	0.42	*CLYBL*	Pancreatic cancer(OS)	Tang,2017
rs7149859	14q24.1	T	0.35	*PIGH*	Breast cancer(CSS)	Guo Q,2015

Variant	locus	MA	MAF	Gene	Reported outcomes	Reference
rs3784099	14q24.1	A	0.4	*RAD51B*	Breast cancer(OS)	Shu,2012
rs17124276	14q31.3	T	0.33	*KCNK10*	Pancreatic cancer(OS)	Tang,2017
rs11621975	14q32.31	G	0.08	*LINC02320*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs166870	15q25.1	T	0.24	*Intergenic*	Breast cancer(DFS)	Song,2015
rs12101726	15q26.2	C	0.24	*LINC01579*	Pancreatic cancer(OS)	Tang,2017
rs72773978	16p13.11	T	0.11	*FOPNL*	Multiple myeloma(OS)	Ziv,2015
rs4780973	16p13.2	T	0.3	*Intergenic*	Pancreatic cancer(OS)	Tang,2017
rs11639759	16p13.3	T	0.16	*RBFOX1*	Pancreatic cancer(OS)	Tang,2017
rs9934948	16q22.3	T	0.38	*ZFHX3*	Breast cancer(OS)	Shu,2012
rs1728400	16q24.1	A	0.33	*Intergenic*	Breast cancer(OS)	Rafiq,2014
rs3795244	17q11.2	T	0.05	*ZNF207*	Pancreatic cancer(OS)	Tang,2017
rs981621	18p11.21	G	0.45	*LDLRAD4*	Pancreatic cancer(OS)	Tang,2017
rs1050631	18q12.2	A	0.26	*SLC39A6*	Esophageal squamous cell cancer (CSS)	Wu,2013
rs8113308	19q13.41	C	0.27	*ZNF613*	Breast cancer(ER+)(OS)	Khan,2015
rs6662005	1q42.3	A	0.22	*ERO1B*	Pancreatic cancer(OS)	Tang,2017
rs2050203	20p11.21	T	0.16	*GAPDHP53*	Serous epithelial ovarian cancer(OS)	Moore,2017
rs1209950	21q22.2	T	0.18	*ETS2*	Non-small cell lung cancer(OS)	Sato,2011
rs9981861	21q22.2	C	0.33	*DSCAM*	Non-small cell lung cancer(OS)	Sato,2011
rs9332377	22q11.21	T	0.17	*COMT*	Osteosarcoma(OS)	Koster,2018

OS, overall survival; EFS, event-free survival; PFS, progression-free survival; CSS, cancer-specific survival; DFS, disease-free survival.

Statistical power

In this analysis I included 82 variants associated with survival outcomes of other cancers. An approximated Bonferroni corrected α level of 5x10^{-4} was also adopted for this study. Similar to the power estimation described in section 5.3.3, this study had a power of 81% and 60% for overall and CRC-specific survival in the SOCCS cohort in order to detect an effect of 1.25 for 79% (65/82) of the variants previously related to survival outcomes of other cancers. As the same set of parameters was used, the power curves for this study can also be read from **Figure 5-17**.

Main analysis

With respect to the 82 genetic variants previously linked with survival outcomes of other cancers, I obtained the hazard ratios along with their confidence intervals by fitting Cox regression models adjusted for age, sex and AJCC stage. After FDR correction, none of these 82 variants were significantly associated with overall or CRC-specific survival of patients in the SOCCS cohort. The detailed effect estimates of the 82 included variants can be found in **Table 5-26**. In terms of uncorrected p-values, I observed three genetic variants (rs1728400, rs17693104 and rs202280) associated with overall survival at p<0.05. As for CRC-specific survival, another set of three variants (rs17693104, rs6797464 and rs823920) were identified with p<0.05. Full lists with detailed results of the associations between the 82 variants and CRC survival in SOCCS are presented in the **Table 5-27**. Among these six variants, the G allele of the variant rs6797464 was reportedly associated with favourable overall survival of osteosarcoma (Koster *et al*, 2018); our study found concordant direction of effect of the G allele related to better CRC-specific survival of CRC patients. The G allele of rs823920 was originally identified to be associated with inferior overall survival of pancreatic cancer (Tang *et al*, 2017). Our result also found the detrimental effect of the G allele of this variant on CRC-specific survival. I was unable to compare the direction of effects of the remaining four variants with previous findings because the effect allele information was unavailable in the original GWASs (Moore *et al*, 2017; Rafiq *et al*, 2014). Among the 82 candidate variants, a total of 51(62%) variants were originally reported without providing information about the effect alleles. Therefore, I was unable to construct a PRS for each individual and did not investigate the combined effect of these 82 variants as a group on CRC survival outcomes.

Table 5-26 Summary of associations between 82 genetic variants previously with survival of other cancers and overall survival of CRC patients in the SOCCS study (N=5,675)

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs202280	G	0.04	*Intergenic*	1.14(1.02-1.26)	0.018	0.722
rs17693104	T	0.35	*SH2D4B*	0.93(0.87-0.99)	0.021	0.722
rs1728400	A	0.33	*Intergenic*	0.93(0.87-0.99)	0.026	0.722
rs2059614	G	0.04	*PKNOX2*	1.13(0.99-1.28)	0.060	0.793

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs10023113	G	0.17	*CAMK2D*	1.09(1.00-1.20)	0.062	0.793
rs17465450	C	0.03	*LRMDA*	0.88(0.77-1.01)	0.064	0.793
rs17124276	T	0.33	*KCNK10*	1.07(1.00-1.16)	0.066	0.793
rs1408536	A	0.12	*Intergenic*	1.16(0.98-1.37)	0.089	0.867
rs1567532	T	0.14	*CTNNA2*	1.06(0.99-1.14)	0.119	0.867
rs114997855	A	0.03	*Intergenic*	1.20(0.94-1.52)	0.137	0.867
rs7149859	T	0.35	*PIGH*	0.95(0.89-1.02)	0.161	0.867
rs6797464	A	0.12	*MECOM*	0.91(0.80-1.04)	0.163	0.867
rs166870	T	0.24	*Intergenic*	1.05(0.98-1.13)	0.169	0.867
rs2900174	G	0.16	*PRB2*	1.14(0.94-1.38)	0.172	0.867
rs12620038	G	0.45	*EPCAM-DT*	1.05(0.98-1.12)	0.174	0.867
rs295315	G	0.37	*Intergenic*	1.05(0.97-1.14)	0.191	0.867
rs1050631	A	0.26	*SLC39A6*	1.04(0.98-1.12)	0.207	0.867
rs9517906	A	0.42	*CLYBL*	0.96(0.90-1.02)	0.213	0.867
rs2314686	A	0.13	*SLC25A37*	0.91(0.78-1.06)	0.226	0.867
rs4382459	T	0.08	*PREX2*	0.94(0.84-1.04)	0.231	0.867
rs9332377	T	0.17	*COMT*	1.05(0.97-1.15)	0.244	0.867
rs7765004	C	0.32	*Intergenic*	1.04(0.97-1.11)	0.251	0.867
rs823920	G	0.12	*Intergenic*	1.05(0.96-1.15)	0.256	0.867
rs12209785	G	0.15	*RUNX2*	0.96(0.89-1.03)	0.275	0.867
rs1944782	G	0.21	*Intergenic*	0.96(0.90-1.03)	0.277	0.867
rs3795244	T	0.05	*ZNF207*	0.93(0.81-1.07)	0.306	0.867
rs1656402	T	0.33	*EIF4E2*	0.96(0.89-1.04)	0.307	0.867
rs17305086	T	0.11	*TENM3-AS1*	0.96(0.89-1.04)	0.326	0.867
rs76010824	A	0.06	*SUCLG2*	1.06(0.94-1.19)	0.334	0.867
rs72773978	T	0.11	*FOPNL*	1.06(0.94-1.20)	0.356	0.867

Variant	**MA**	**MAF**	**Gene**	**HR(95%CI)**	**P(uncorrected)**	**Pfdr**
rs11733008	T	0.36	*Intergenic*	1.03(0.97-1.10)	0.357	0.867
rs113988120	A	0.01	*PAIP2B*	0.90(0.71-1.14)	0.371	0.867
rs1352757	A	0.43	*Intergenic*	1.03(0.97-1.10)	0.375	0.867
rs1414153	C	0.23	*1-Dec*	1.04(0.96-1.12)	0.385	0.867
rs7712513	G	0.26	*Intergenic*	0.97(0.91-1.04)	0.388	0.867
rs4568126	A	0.33	*B4GALT4*	1.03(0.96-1.10)	0.390	0.867
rs11639759	T	0.16	*RBFOX1*	0.95(0.84-1.07)	0.405	0.867
rs10835188	G	0.36	*LIN7C*	1.03(0.96-1.11)	0.419	0.867
rs7701292	C	0.08	*Intergenic*	1.04(0.95-1.13)	0.430	0.867
rs1564271	A	0.23	*PDSS1*	1.03(0.96-1.10)	0.434	0.867
rs148760487	G	0.01	*Intergenic*	0.90(0.68-1.18)	0.435	0.867
rs2050203	T	0.16	*GAPDHP53*	0.97(0.88-1.06)	0.456	0.867
rs16861827	T	0.14	*IGSF21*	0.96(0.87-1.06)	0.473	0.867
rs10767646	T	0.33	*BDNF-AS*	1.03(0.95-1.11)	0.476	0.867
rs10500780	A	0.14	*BTBD10*	1.03(0.94-1.13)	0.497	0.867
rs4618572	T	0.1	*Intergenic*	0.97(0.89-1.06)	0.509	0.867
rs770996	T	0.35	*AC034195.1*	1.02(0.96-1.09)	0.511	0.867
rs2299187	T	0.07	*CACNA2D1*	1.11(0.81-1.53)	0.518	0.867
rs17248137	G	0.05	*Intergenic*	1.04(0.92-1.18)	0.523	0.867
rs1209950	T	0.18	*ETS2*	0.98(0.92-1.04)	0.526	0.867
rs716274	G	0.44	*Intergenic*	0.98(0.92-1.04)	0.526	0.867
rs9593831	T	0.22	*Intergenic*	0.97(0.88-1.07)	0.546	0.871
rs3784099	A	0.4	*RAD51B*	1.02(0.95-1.10)	0.556	0.871
rs4150579	A	0.33	*GTF2H1*	1.02(0.95-1.10)	0.567	0.871
rs6662005	A	0.22	*ERO1B*	0.97(0.86-1.09)	0.590	0.871
rs10817611	C	0.23	*WHRN*	0.98(0.90-1.07)	0.645	0.871

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs1391315	G	0.1	*SMAP2*	1.03(0.90-1.18)	0.646	0.871
rs17548007	T	0.04	*Intergenic*	1.03(0.92-1.14)	0.646	0.871
rs55933544	T	0.2	*GLDC*	0.98(0.91-1.06)	0.650	0.871
rs7777171	C	0.47	*AGMO*	0.98(0.92-1.06)	0.654	0.871
rs421379	T	0.28	*Intergenic*	0.97(0.84-1.12)	0.664	0.871
rs4955138	G	0.31	*Intergenic*	0.98(0.90-1.07)	0.666	0.871
rs361052	A	0.27	*IQSEC1*	1.02(0.94-1.10)	0.671	0.871
rs4285214	T	0.39	*ZNF608*	0.99(0.93-1.05)	0.674	0.871
rs1454694	C	0.17	*Intergenic*	1.02(0.94-1.09)	0.692	0.871
rs12362504	C	0.38	*SBF2*	0.99(0.92-1.06)	0.696	0.871
rs8113308	C	0.27	*ZNF613*	0.98(0.89-1.08)	0.699	0.871
rs981621	G	0.45	*LDLRAD4*	0.99(0.92-1.06)	0.705	0.871
rs763780	C	0.09	*IL17F*	0.98(0.84-1.13)	0.748	0.894
rs9934948	T	0.38	*ZFHX3*	1.01(0.93-1.11)	0.756	0.894
rs4780973	T	0.3	*Intergenic*	1.01(0.94-1.08)	0.804	0.925
rs6986444	T	0.33	*SNTB1*	1.01(0.91-1.13)	0.836	0.949
rs12101726	C	0.24	*LINC01579*	1.01(0.87-1.17)	0.886	0.982
rs11062040	C	0.46	*DCP1B*	1.00(0.93-1.06)	0.900	0.982
rs9981861	C	0.33	*DSCAM*	1.00(0.94-1.07)	0.900	0.982
rs11621975	G	0.08	*LINC02320*	0.99(0.89-1.11)	0.921	0.992
rs10983614	C	0.28	*ASTN2*	1.00(0.94-1.06)	0.946	0.997
rs10825036	G	0.19	*Intergenic*	1.00(0.93-1.07)	0.965	0.997
rs12146774	T	0.18	*AC084880.2*	1.00(0.91-1.10)	0.973	0.997
rs10899426	C	0.02	*Intergenic*	1.00(0.85-1.17)	0.976	0.997
rs12358475	A	0.11	*Intergenic*	1.00(0.93-1.07)	0.992	0.997
rs10736390	A	0.46	*MROH7*	1.00(0.93-1.07)	0.997	0.997

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values corrected using the false positive rate approach.

Table 5-27 Summary of associations between 82 genetic variants previously with survival of other cancers and CRC-specific survival of CRC patients in the SOCCS study (N=5,675)

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs6797464	A	0.12	*MECOM*	0.85(0.73-0.98)	0.030	0.594
rs17693104	T	0.35	*SH2D4B*	0.92(0.85-0.99)	0.031	0.594
rs823920	G	0.12	*Intergenic*	1.11(1.00-1.23)	0.042	0.594
rs202280	G	0.04	*Intergenic*	1.13(1.00-1.29)	0.051	0.594
rs17465450	C	0.03	*LRMDA*	0.85(0.72-1.00)	0.053	0.594
rs1728400	A	0.33	*Intergenic*	0.93(0.86-1.00)	0.060	0.594
rs17124276	T	0.33	*KCNK10*	1.09(1.00-1.19)	0.060	0.594
rs10023113	G	0.17	*CAMK2D*	1.11(0.99-1.23)	0.063	0.594
rs1408536	A	0.12	*Intergenic*	1.22(0.99-1.50)	0.064	0.594
rs1050631	A	0.26	*SLC39A6*	1.07(0.99-1.16)	0.096	0.631
rs76010824	A	0.06	*SUCLG2*	1.11(0.98-1.27)	0.110	0.631
rs7712513	G	0.26	*Intergenic*	0.94(0.86-1.01)	0.110	0.631
rs2299187	T	0.07	*CACNA2D1*	1.32(0.93-1.87)	0.117	0.631
rs10736390	A	0.46	*MROH7*	1.07(0.98-1.15)	0.118	0.631
rs7765004	C	0.32	*Intergenic*	1.06(0.98-1.15)	0.131	0.631
rs12209785	G	0.15	*RUNX2*	0.94(0.86-1.02)	0.136	0.631
rs1567532	T	0.14	*CTNNA2*	1.07(0.98-1.17)	0.142	0.631
rs1656402	T	0.33	*EIF4E2*	0.94(0.86-1.02)	0.157	0.631
rs2314686	A	0.13	*SLC25A37*	0.87(0.72-1.06)	0.160	0.631
rs2900174	G	0.16	*PRB2*	1.17(0.94-1.47)	0.162	0.631
rs12620038	G	0.45	*EPCAM-DT*	1.06(0.98-1.15)	0.165	0.631
rs1352757	A	0.43	*Intergenic*	1.05(0.98-1.13)	0.191	0.697
rs1414153	C	0.23	*01-Dec*	1.06(0.97-1.17)	0.202	0.704
rs9517906	A	0.42	*CLYBL*	0.95(0.88-1.03)	0.210	0.704

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs4780973	T	0.3	*Intergenic*	1.05(0.97-1.14)	0.238	0.720
rs114997855	A	0.03	*Intergenic*	1.19(0.89-1.58)	0.239	0.720
rs763780	C	0.09	*IL17F*	0.90(0.75-1.07)	0.240	0.720
rs4568126	A	0.33	*B4GALT4*	1.05(0.97-1.13)	0.256	0.740
rs17248137	G	0.05	*Intergenic*	1.08(0.94-1.25)	0.284	0.794
rs4618572	T	0.1	*Intergenic*	0.95(0.85-1.06)	0.337	0.884
rs716274	G	0.44	*Intergenic*	0.96(0.90-1.04)	0.349	0.884
rs9332377	T	0.17	*COMT*	1.05(0.95-1.17)	0.360	0.884
rs1944782	G	0.21	*Intergenic*	0.96(0.89-1.04)	0.371	0.884
rs72773978	T	0.11	*FOPNL*	0.93(0.79-1.09)	0.375	0.884
rs1209950	T	0.18	*ETS2*	0.97(0.90-1.04)	0.390	0.884
rs7149859	T	0.35	*PIGH*	0.97(0.89-1.04)	0.392	0.884
rs11733008	T	0.36	*Intergenic*	1.03(0.96-1.12)	0.412	0.884
rs4150579	A	0.33	*GTF2H1*	1.04(0.95-1.13)	0.419	0.884
rs1454694	C	0.17	*Intergenic*	1.04(0.95-1.13)	0.440	0.884
rs9593831	T	0.22	*Intergenic*	0.96(0.86-1.07)	0.453	0.884
rs2050203	T	0.16	*GAPDHP53*	0.96(0.86-1.07)	0.473	0.884
rs12362504	C	0.38	*SBF2*	0.97(0.89-1.06)	0.491	0.884
rs981621	G	0.45	*LDLRAD4*	0.97(0.89-1.06)	0.496	0.884
rs2059614	G	0.04	*PKNOX2*	1.06(0.90-1.23)	0.501	0.884
rs1564271	A	0.23	*PDSS1*	1.03(0.95-1.12)	0.513	0.884
rs295315	G	0.37	*Intergenic*	1.03(0.94-1.13)	0.514	0.884
rs10817611	C	0.23	*WHRN*	1.03(0.93-1.14)	0.516	0.884
rs4285214	T	0.39	*ZNF608*	0.98(0.91-1.05)	0.549	0.922
rs9981861	C	0.33	*DSCAM*	0.98(0.91-1.06)	0.579	0.954
rs3795244	T	0.05	*ZNF207*	0.96(0.81-1.13)	0.597	0.964
rs55933544	T	0.2	*GLDC*	0.98(0.89-1.07)	0.627	0.980

Variant	**MA**	**MAF**	**Gene**	**HR(95%CI)**	**P(uncorrected)**	**Pfdr**
rs12146774	T	0.18	*AC084880.2*	1.03(0.92-1.15)	0.639	0.980
rs1391315	G	0.1	*SMAP2*	1.04(0.89-1.21)	0.653	0.980
rs16861827	T	0.14	*IGSF21*	0.98(0.87-1.10)	0.673	0.980
rs10500780	A	0.14	*BTBD10*	1.02(0.92-1.14)	0.675	0.980
rs4955138	G	0.31	*Intergenic*	1.02(0.92-1.13)	0.677	0.980
rs11639759	T	0.16	*RBFOX1*	1.03(0.89-1.19)	0.696	0.982
rs8113308	C	0.27	*ZNF613*	1.02(0.91-1.14)	0.724	0.982
rs148760487	G	0.01	*Intergenic*	0.95(0.69-1.30)	0.756	0.982
rs7701292	C	0.08	*Intergenic*	1.02(0.92-1.13)	0.764	0.982
rs17548007	T	0.04	*Intergenic*	1.02(0.90-1.16)	0.784	0.982
rs12358475	A	0.11	*Intergenic*	1.01(0.93-1.10)	0.792	0.982
rs4382459	T	0.08	*PREX2*	0.98(0.87-1.11)	0.807	0.982
rs10899426	C	0.02	*Intergenic*	1.02(0.85-1.23)	0.815	0.982
rs166870	T	0.24	*Intergenic*	1.01(0.93-1.10)	0.822	0.982
rs9934948	T	0.38	*ZFHX3*	0.99(0.89-1.10)	0.837	0.982
rs770996	T	0.35	*AC034195.1*	0.99(0.92-1.07)	0.866	0.982
rs10983614	C	0.28	*ASTN2*	1.01(0.93-1.09)	0.868	0.982
rs361052	A	0.27	*IQSEC1*	1.01(0.91-1.11)	0.873	0.982
rs11062040	C	0.46	*DCP1B*	0.99(0.92-1.07)	0.876	0.982
rs11621975	G	0.08	*LINC02320*	1.01(0.89-1.14)	0.877	0.982
rs421379	T	0.28	*Intergenic*	0.99(0.83-1.17)	0.882	0.982
rs17305086	T	0.11	*TENM3-AS1*	0.99(0.90-1.09)	0.899	0.982
rs10767646	T	0.33	*BDNF-AS*	1.00(0.92-1.10)	0.917	0.982
rs113988120	A	0.01	*PAIP2B*	0.99(0.75-1.30)	0.925	0.982
rs3784099	A	0.4	*RAD51B*	1.00(0.91-1.09)	0.933	0.982
rs10835188	G	0.36	*LIN7C*	1.00(0.91-1.09)	0.942	0.982
rs6986444	T	0.33	*SNTB1*	1.00(0.88-1.13)	0.945	0.982

Variant	MA	MAF	Gene	HR(95%CI)	P(uncorrected)	Pfdr
rs7777171	C	0.47	*AGMO*	1.00(0.92-1.08)	0.956	0.982
rs12101726	C	0.24	*LINC01579*	1.00(0.85-1.19)	0.959	0.982
rs6662005	A	0.22	*ERO1B*	1.00(0.87-1.15)	0.987	0.997
rs10825036	G	0.19	*Intergenic*	1.00(0.92-1.09)	0.997	0.997

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values corrected using the false positive rate approach.

Stratified analysis

I conducted survival analysis stratified by sex, stage and tumour site investigating the effects on CRC survival of the 82 genetic variants previously linked with survival outcomes of other cancers. Overall, no statistically significant associations that survived correction for multiple testing were identified in any strata of patients. With respect to sex-stratified analysis, five variants were found to be associated with overall survival in male CRC patients (N=3,235); eight variants were related to CRC-specific survival in the presence of nominal statistical significance (p<0.05) (**Table 5-28**). Amongst female CRC patients in SOCCS, we found two variants (rs10500780 and rs12362504) associated with both overall and CRC-specific survival. Genetic variant rs202280 was associated with overall survival of female patients, whereas the variant rs10736390 was correlated with CRC-specific survival (**Table 5-28**). Detailed results of all the 82 variants are presented in **Appendix Table 14**.

Table 5-28 Summary of associations (p<0.05) stratified by sex between CRC survival in the SOCCS cohort and genetic variants associated with survival outcomes of other cancers

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
Male (N=3,235)						
OS						
	rs2059614	G	0.04	1.25(1.06-1.48)	0.008	0.522
	rs10023113	G	0.17	1.15(1.03-1.30)	0.017	0.522
	rs12209785	G	0.15	0.89(0.81-0.98)	0.019	0.522

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
	rs1408536	A	0.12	1.24(1.00-1.54)	0.048	0.755
	rs1728400	A	0.33	0.92(0.85-1.00)	0.049	0.755
CSS						
	rs823920	G	0.12	1.17(1.03-1.33)	0.016	0.480
	rs12209785	G	0.15	0.88(0.79-0.98)	0.022	0.480
	rs7712513	G	0.26	0.89(0.80-0.99)	0.026	0.480
	rs10023113	G	0.17	1.17(1.02-1.35)	0.028	0.480
	rs17693104	T	0.35	0.90(0.82-0.99)	0.034	0.480
	rs72773978	T	0.11	0.79(0.63-0.99)	0.043	0.480
	rs1050631	A	0.26	1.11(1.00-1.23)	0.048	0.480
	rs1408536	A	0.12	1.30(1.00-1.69)	0.050	0.480
Female (N=2,440)						
OS						
	rs202280	G	0.04	1.28(1.09-1.51)	0.003	0.234
	rs12362504	C	0.38	0.87(0.77-0.98)	0.018	0.665
	rs10500780	A	0.14	1.15(1.01-1.32)	0.042	0.665
CSS						
	rs10736390	A	0.46	1.20(1.06-1.36)	0.004	0.317
	rs12362504	C	0.38	0.84(0.73-0.97)	0.016	0.506
	rs10500780	A	0.14	1.21(1.03-1.42)	0.018	0.506

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

I identified a total of 22 associations at nominal significance (p<0.05) between 15 unique genetic variants and survival outcomes of CRC in stage II/III or stage IV CRC patients in SOCCS, although none of them retained statistical significance after FDR correction. The summarised effect estimates of these 22 associations are presented in **Table 5-29** and all non-significant results (p>0.05) are summarised in **Appendix Table 15**.

Table 5-29 Summary of associations (p<0.05) stratified by stage between CRC survival in the SOCCS cohort and genetic variants associated with survival outcomes of other cancers

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
Stage II/III (N=3,886)						
OS						
	rs1728400	A	0.33	0.90(0.83-0.98)	0.014	0.543
	rs9517906	A	0.42	0.90(0.83-0.98)	0.014	0.543
	rs1408536	A	0.12	1.28(1.03-1.60)	0.026	0.543
	rs4382459	T	0.08	0.85(0.73-0.98)	0.027	0.543
	rs10835188	G	0.36	1.11(1.00-1.22)	0.041	0.543
	rs2900174	G	0.16	1.27(1.01-1.60)	0.043	0.543
	rs17305086	T	0.11	0.90(0.80-1.00)	0.045	0.543
CSS						
	rs1408536	A	0.12	1.59(1.17-2.15)	0.003	0.160
	rs9517906	A	0.42	0.86(0.77-0.95)	0.004	0.160
	rs1728400	A	0.33	0.88(0.80-0.98)	0.016	0.461
	rs2900174	G	0.16	1.33(1.01-1.77)	0.045	0.679
	rs361052	A	0.27	1.14(1.00-1.31)	0.047	0.679
Stage IV (N=784)						
OS						
	rs1050631	A	0.26	1.17(1.04-1.32)	0.010	0.468
	rs17693104	T	0.35	0.86(0.76-0.97)	0.012	0.468
	rs1414153	C	0.23	1.20(1.03-1.40)	0.022	0.468
	rs1454694	C	0.17	1.18(1.02-1.36)	0.022	0.468
	rs1567532	T	0.14	1.16(1.01-1.33)	0.038	0.579
CSS						

Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
rs1050631	A	0.26	1.18(1.04-1.33)	0.008	0.633
rs17693104	T	0.35	0.86(0.77-0.98)	0.019	0.633
rs1454694	C	0.17	1.18(1.02-1.36)	0.031	0.633
rs202280	G	0.04	1.24(1.02-1.52)	0.035	0.633
rs4150579	A	0.33	1.16(1.00-1.33)	0.042	0.633

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

Regarding tumour site, I detected no significant associations after FDR correction in either colon or rectal cancer patients. Seven associations at nominal significance (p<0.05) including six unique genetic variants were found in colon cancer patients (details in **Table 5-30**). For rectal cancer patients, however, I did not identify any variants associated with overall survival with uncorrected p<0.05. Three variants were correlated with CRC-specific survival of rectal cancer patients in SOCCS in the presence of nominal significance (p<0.05). I summarise the detailed results of the 82 variants in relation to survival outcomes stratified by tumour site in **Appendix Table 16**.

Table 5-30 Summary of associations (p<0.05) stratified by tumour site between CRC survival in the SOCCS cohort and genetic variants associated with survival outcomes of other cancers

Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
Colon (N=3,392)					
OS					
rs2059614	G	0.04	1.24(1.05-1.46)	0.009	0.669
rs17124276	T	0.33	1.13(1.02-1.24)	0.019	0.669
rs17693104	T	0.35	0.91(0.84-0.99)	0.026	0.669
rs1567532	T	0.14	1.11(1.01-1.22)	0.032	0.669
CSS					

	Variant	MA	MAF	HR(95%CI)	P(uncorrected)	Pfdr
	rs823920	G	0.12	1.18(1.04-1.35)	0.010	0.775
	rs17124276	T	0.33	1.15(1.02-1.29)	0.022	0.775
	rs8113308	C	0.27	1.15(1.00-1.32)	0.050	0.775
Rectal (N=2,201)						
CSS						
	rs6797464	A	0.12	0.76(0.60-0.96)	0.024	0.996
	rs10023113	G	0.17	1.21(1.02-1.43)	0.033	0.996
	rs17465450	C	0.03	0.76(0.58-1.00)	0.047	0.996

MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival

Sensitivity analysis

I analysed associations between the 82 genetic variants previously linked with survival outcomes of other cancers and CRC survival under a recessive model in the SOCCS study. Colorectal cancer patients carrying the TT genotype of the variant rs2299187 had significantly favourable overall (HR=0.54, 95%CI=0.42-0.70, uncorrected p=2.6x10^{-6}, Pfdr=1.8x10^{-4}) and CRC-specific survival (HR=0.52, 95%CI=0.39-0.69, uncorrected=6.1x10^{-6}, Pfdr=2.6x10^{-4}) compared with ones with AA or AT genotypes. This effect showed concordant direction with the original report where the TT genotype was linked with improved overall survival of head and neck cancer (Azad *et al*, 2016). I also observed three variants associated with CRC survival at nominal significance (p<0.05) although they did not survive FDR correction (**Table 5-31**). Additional details on the results of other remaining candidate variants can be found in **Appendix Table 17**.

Table 5-31 Summary of associations (p<0.05) between survival outcomes of CRC patients in the SOCCS study and genetic variants previously linked with survival outcomes of other cancers under a recessive model (N=5,675)

	Variant	MG	MGF	HR(95%CI)	P(uncorrected)	Pfdr
OS						

	Variant	MG	MGF	HR(95%CI)	P(uncorrected)	Pfdr
	rs2299187	TT	0.01	0.54(0.42-0.70)	2.6E-06	1.8E-04
	rs72773978	TT	0.02	0.29(0.09-0.89)	0.03	0.84
CSS						
	rs2299187	TT	0.01	0.52(0.39-0.69)	6.1E-06	2.6E-04
	rs17124276	TT	0.02	1.32(1.05-1.65)	0.015	0.424
	rs6797464	GG	0.02	0.84(0.72-0.99)	0.036	0.555
	rs10736390	GG	0.21	1.16(1.00-1.35)	0.048	0.555

MG, minor genotype; MGF, minor genotype frequency; HR, hazard ratio; CI, confidence interval; Pfdr, p-values adjusted using the false positive rate approach. OS, overall survival, CSS, CRC-specific survival.

I failed to replicate the association between the variant rs2299187 and CRC survival in the UK Biobank due to the fact that this variant had a low minor allele frequency (T allele: 0.02) and there were no homozygous individuals (TT) present in the UK Biobank cohort.

5.4 Genome-wide association analysis

5.4.1 Statistical power

Following similar procedures as in previous sections, I estimated the statistical power of the GWA study. The α level of GWAS significance ($5x10^{-8}$) was used in combination with other metrics including the sample size (N=5,675), the minor allele frequency (0.01 to 0.50), proportions of events (34% for overall survival and 24% for CRC-specific survival) and various effect sizes (HR from 1.2 to 2.0). According to the method proposed by Owzar et al.(Owzar *et al*, 2012), the GWA study had a power of 75% to detect an effect of 1.30 (HR) on overall survival for a genetic variant with a minor allele frequency of 0.15. For CRC-specific survival, however, a statistical power of 76% was expected to observe an effect of 1.30 for a variant with a minor allele frequency of 0.25. I plotted power curves for variants of varied minor allele frequencies in relation to a range of effect sizes in **Figure 5-19**. As suggested by the power curves, the GWA study had limited statistical power (<30%) to identify small or moderate survival effects (HR<1.3) for genetic variants with low minor allele frequencies (<0.05) especially on CRC-specific survival. Notably, there could be slight overestimation of the statistical power (2-4%) as here the power was estimated assuming a Cox regression approach whereas the main GWAS analysis was conducted using a Martingale residual-based approach (details in page 104).

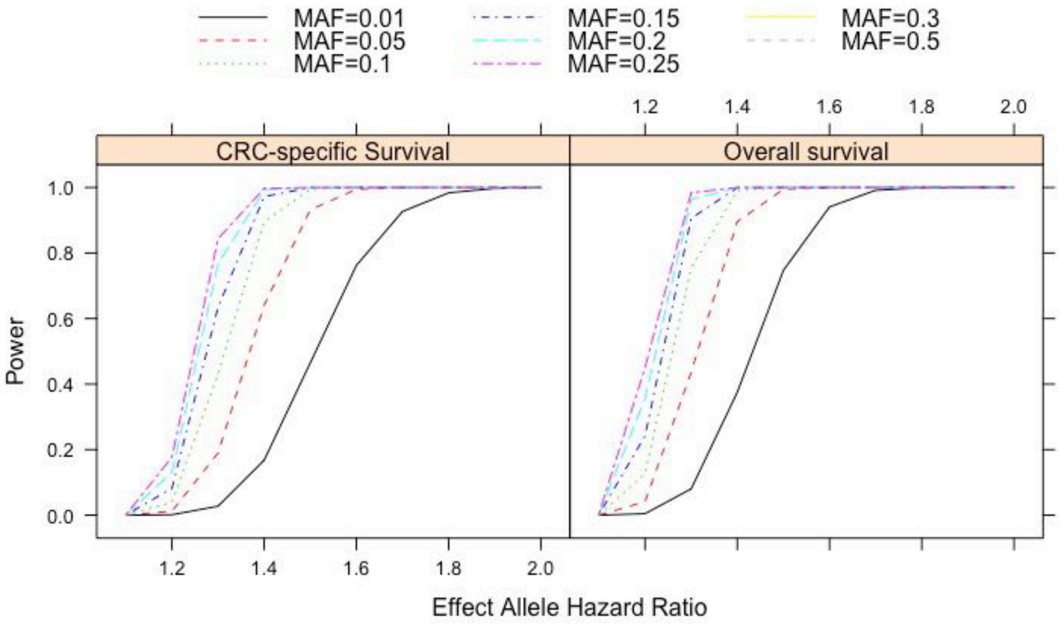


Figure 5-19 Power curves for the genome-wide association study in the SOCCS cohort

5.4.2 Main results

As described in Chapter 4, I included a total of 8,328,632 autosomal genetic variants and investigated their associations with survival outcomes in 5,675 CRC patients of the SOCCS cohort using the Martingale residual-based approach. Similar to previous candidate association studies, I adjusted for covariates including age at CRC diagnosis, sex and AJCC stage to estimate the effect of a certain variant on overall and CRC-specific survival.

For overall survival, I identified 415,724 (4.99%) at p<0.05. Associations between these variants (p<0.05) and the survival outcome (measured by the Martingale residual) are presented in the Manhattan plot (**Figure 5-20**). As shown in the plot, no genetic variants were identified to be associated with overall survival with GWAS significance ($p<5x10^{-8}$). The strongest signal was for rs143664541 in chromosome 6 (Martingale residual coefficient =0.769, SE=0.154, $p=5.89x10^{-7}$).

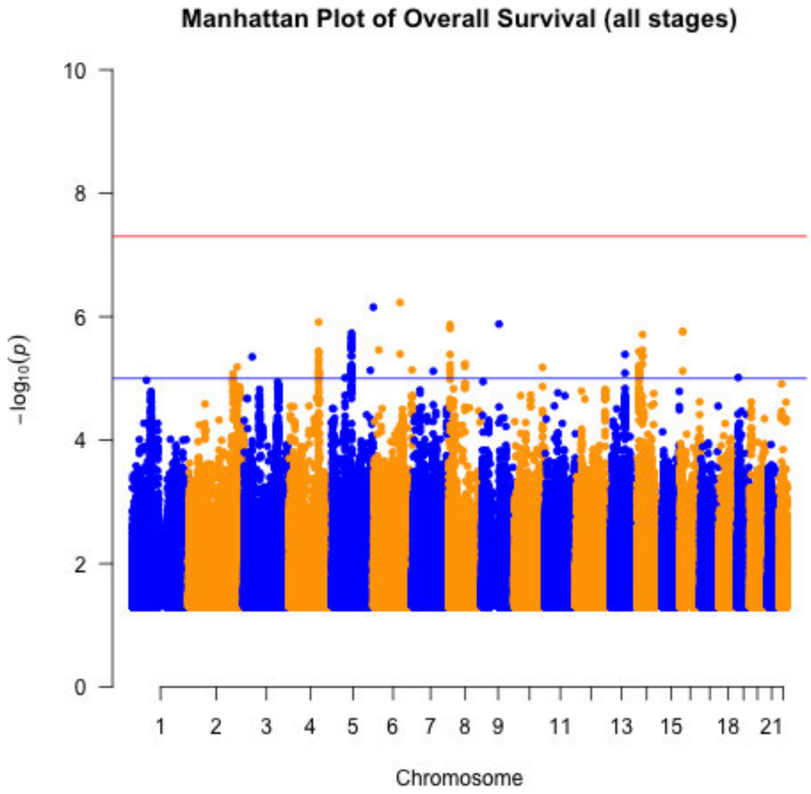


Figure 5-20 Manhattan plot of GWAS results on overall survival of all CRC patients in the SOCCS cohort (blue line: $p=10^{-5}$, red line=$5x10^{-8}$)

I also plotted all the eight million observed p-values from the GWA analysis against a theoretical null distribution in a QQ plot (**Figure 5-21**). The QQ plot indicated no systematic inflation of statistical significance in the GWAS results on overall survival, which was underpinned by an inflation factor (lambda) of 0.9991 (<1.1).

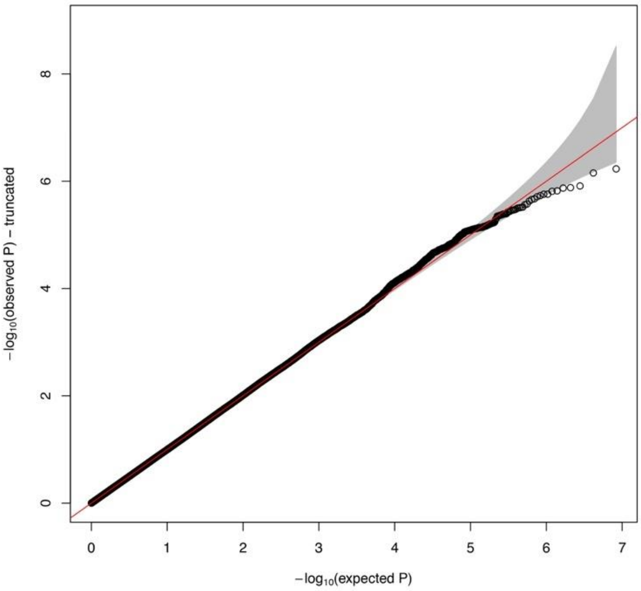


Figure 5-21 QQ plot of GWAS results on overall survival of all CRC patients in the SOCCS cohort

By implementing a less stringent p-value threshold ($p<5x10^{-6}$), I found a total of 38 correlated genetic variants associated with overall survival. Pairs of genetic variants in linkage disequilibrium (LD) were investigated by calculating the r^2 and for pairs with $r^2>0.2$, I kept the variant with the smaller p-values in the association with survival outcomes in SOCCS. After controlling for LD, I obtained 10 independent genetic variants with p-values less than $5x10^{-6}$; basic characteristics along with regression coefficients of these variants are summarised in **Table 5-32**.

Table 5-32 Genetic variants identified from the genome-wide association analysis associated with overall survival of CRC patients in the SOCCS cohort ($p<5x10^{-6}$) using the Martingale-residual based approach (N=5,675)

Variant	Chr	MA	MAF	Beta	SE	P
rs143664541	6	A	0.014	0.769	0.154	5.89E-07
rs6869766	5	A	0.139	0.248	0.050	7.00E-07
rs185673294	4	G	0.126	-0.256	0.053	1.23E-06
rs75809467	9	T	0.034	0.490	0.101	1.32E-06
rs4484717	8	C	0.429	-0.164	0.034	1.34E-06

Variant	Chr	MA	MAF	Beta	SE	P
rs138959556	16	T	0.037	-0.461	0.096	1.74E-06
rs4441183	14	A	0.128	0.239	0.050	1.95E-06
rs34858830	14	A	0.332	0.166	0.036	3.68E-06
rs141093197	13	T	0.485	0.158	0.034	4.09E-06
rs60676294	3	G	0.139	0.237	0.052	4.46E-06

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; Beta, regression coefficients; SE, standard error.

I then re-fitted Cox regression models to obtain exact effect estimates and p-values of these 10 genetic variants. The summarised results of Cox regression models are presented in **Table 5-33**. Notably, I identified one variant rs143664541 in chromosome 6 that reached GWAS statistical significance (HR=1.92, 95%CI=1.52-2.42, p=4.24x10^{-8}). Given the low minor allele frequency (0.014), I plotted Kaplan-Meier estimates of this variant in **Figure 5-22** under a dominant genetic model (GA+AA vs GG).

Table 5-33 Summary of effect estimates from Cox models of GWAS-identified variants associated with overall survival in the SOCCS cohort (p<5x10^{-6})

Variant	Chr	MA	MAF	HR(95%CI)	P
rs143664541	6	A	0.014	1.92(1.52-2.42)	4.24E-08
rs75809467	9	T	0.034	1.58(1.33-1.87)	1.55E-07
rs4441183	14	A	0.128	1.26(1.15-1.38)	5.26E-07
rs6869766	5	A	0.139	1.30(1.17-1.44)	6.02E-07
rs4484717	8	C	0.429	0.85(0.80-0.91)	7.81E-07
rs185673294	4	G	0.126	0.77(0.69-0.86)	1.43E-06
rs60676294	3	G	0.139	1.25(1.14-1.37)	1.82E-06
rs141093197	13	T	0.485	1.17(1.10-1.25)	2.34E-06
rs34858830	14	A	0.332	1.18(1.10-1.26)	3.26E-06

Variant	Chr	MA	MAF	HR(95%CI)	P
rs138959556	16	T	0.037	0.61(0.50-0.76)	4.85E-06

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval.

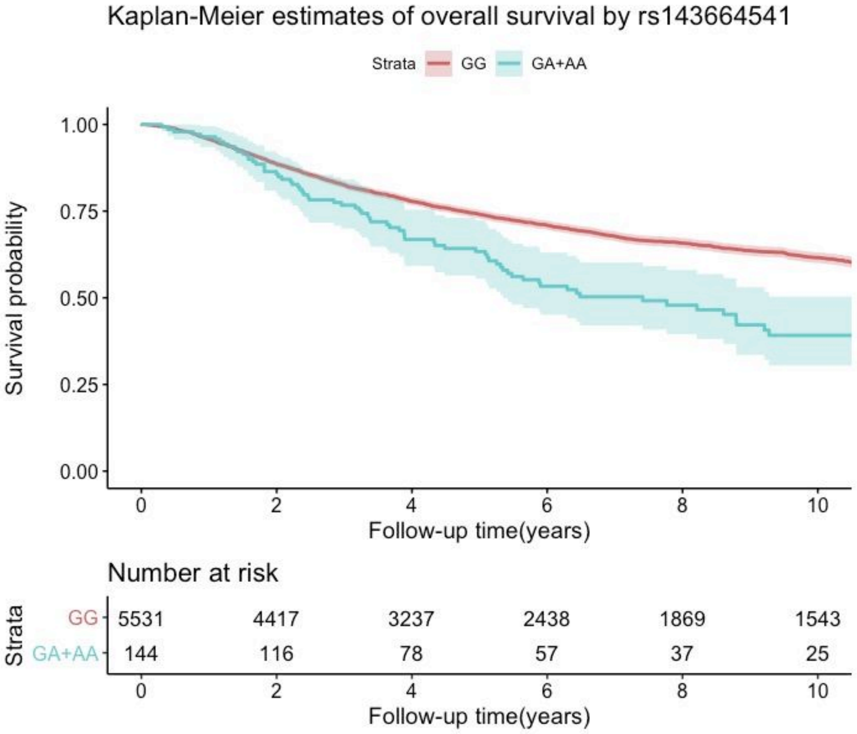


Figure 5-22 Kaplan-Meier estimates of overall survival in the SOCCS stratified by the genetic variant rs143664541

With respect to the GWAS on CRC-specific survival, in total, I found 414,706 (4.98%) genetic variants associated with CRC-specific survival in the SOCCS cohort. The Manhattan plot displaying all the 414,706 signals throughout the genome is presented in **Figure 5-23**. As indicated by the plot, the strongest signal was detected at the same variant rs143664541 in chromosome 6 (Martingale residual coefficient=1.04, SE=0.182, p=9.10x10^{-9}).

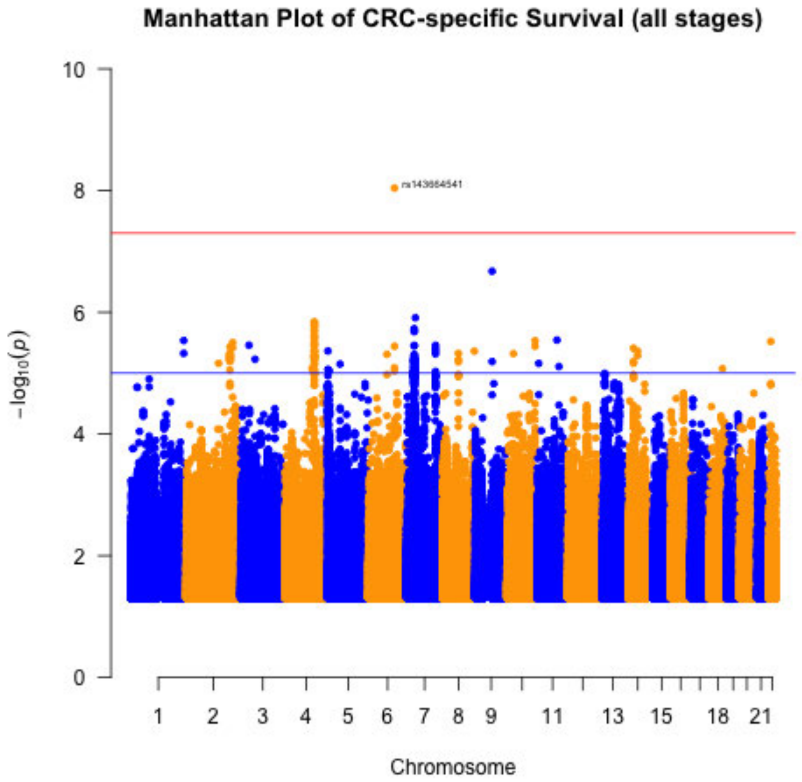


Figure 5-23 Manhattan plot of GWAS results on CRC-specific survival of all CRC patients in the SOCCS cohort (blue line: p=10^{-5}, red line: $5x10^{-8}$)

Similar to the results of overall survival, the QQ plot of CRC-specific survival (presented in **Figure 5-24**) aggregating p-values from all the eight million variants showed no systematic inflation of statistical significance. I also observed an inflation factor (lambda) of 1.004 (<1.1) verifying the absence of systematic inflation.

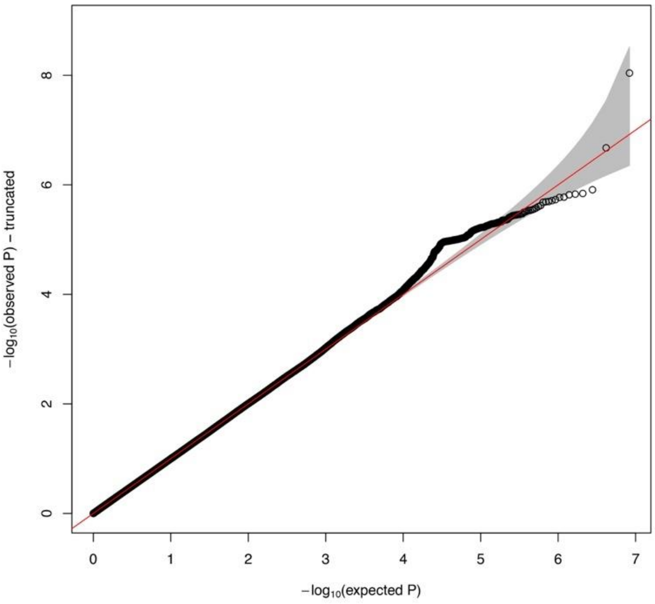


Figure 5-24 QQ plot of GWAS results on CRC-specific survival of all CRC patients in the SOCCS cohort

I then extracted a total of 52 genetic variants associated with CRC-specific survival with p<5x10^{-6}. After addressing LD among these variants, 19 independent variants were retained (presented in **Table 5-34**). In addition to rs143664541, I identified another variant rs75809467 in chromosome 9 in association with CRC-specific survival at p<5x10^{-7} (Martingale residual coefficient=0.621, standard error=0.119, p=2.11x10^{-7})

Table 5-34 Genetic variants identified from the genome-wide association analysis associated with CRC-specific survival of CRC patients in the SOCCS cohort (p<5x10^{-6}) using the Martingale-residual based approach

Variant	Chr	MA	MAF	Beta	SE	P
rs143664541	6	A	0.014	1.045	0.182	9.10E-09
rs75809467	9	T	0.034	0.621	0.119	2.11E-07
rs75796335	7	C	0.111	0.317	0.065	1.23E-06
rs12648214	4	C	0.157	-0.263	0.055	1.44E-06
rs117363837	7	G	0.213	0.234	0.049	1.86E-06
rs76941929	11	A	0.022	0.645	0.138	2.86E-06

rs72767774	1	A	0.053	0.423	0.09	2.92E-06
rs17105163	10	A	0.029	0.55	0.118	2.93E-06
rs79014181	22	A	0.06	0.41	0.088	3.01E-06
rs1207145	2	G	0.017	-0.731	0.157	3.15E-06
rs6799780	3	G	0.107	0.266	0.057	3.48E-06
rs12113115	7	T	0.139	0.265	0.057	3.49E-06
rs4441183	14	A	0.128	0.274	0.059	3.89E-06
rs7724103	5	G	0.034	0.516	0.112	4.32E-06
rs17808098	14	A	0.474	-0.181	0.039	4.36E-06
rs185274835	8	C	0.011	0.914	0.199	4.37E-06
rs111411038	8	G	0.019	0.684	0.149	4.74E-06
rs72803621	10	G	0.044	0.454	0.099	4.78E-06
rs147529871	6	T	0.02	-0.682	0.149	4.92E-06

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; Beta, regression coefficients; SE, standard error.

The effect estimates on CRC-specific survival of these 19 variants were re-estimated in Cox regression models **(Table 5-35**). I observed two variants (rs143664541 and rs75809467) that reached GWAS significance in association with CRC-specific survival. In particular, the A allele of the variant rs143664541 conferred significantly higher hazard of CRC-related death (HR=1.92, 95%CI=1.52-2.42, p=4.24x10^{-8}). The other variant rs75809467, with the T allele as the risk allele, was significantly associated with inferior CRC-specific survival (HR=1.81, 95%CI=1.48-2.20, p=7.07x10^{-9}). Survival curves of these two variants are plotted in **Figure 5-25** and **Figure 5-26** with the variants coded under a dominant genetic model.

Table 5-35 Summary of effect estimates from Cox models of GWAS-identified variants associated with CRC-specific survival in the SOCCS cohort (p<5x10^{-6})

Variant	**Chr**	**EA**	**MAF**	**HR(95%CI)**	**P**
rs143664541	6	A	0.014	2.17(1.69-2.78)	1.14E-09
rs75809467	9	T	0.034	1.80(1.48-2.20)	7.07E-09
rs76941929	11	A	0.022	1.84(1.47-2.30)	8.37E-08
rs185274835	8	C	0.011	2.18(1.63-2.90)	1.17E-07
rs75796335	7	C	0.111	1.36(1.21-1.53)	2.45E-07
rs17105163	10	A	0.029	1.64(1.36-1.98)	2.86E-07

Variant	Chr	EA	MAF	HR(95%CI)	P
rs7724103	5	G	0.034	1.61(1.34-1.94)	5.59E-07
rs117363837	7	G	0.213	1.26(1.15-1.38)	6.70E-07
rs72803621	10	G	0.044	1.53(1.29-1.81)	8.65E-07
rs79014181	22	A	0.06	1.45(1.25-1.68)	9.53E-07
rs72767774	1	A	0.053	1.45(1.25-1.69)	1.01E-06
rs12113115	7	T	0.139	1.29(1.16-1.43)	1.28E-06
rs6799780	3	G	0.107	1.28(1.16-1.42)	1.30E-06
rs111411038	8	G	0.019	1.76(1.40-2.22)	1.32E-06
rs4441183	14	A	0.128	1.29(1.16-1.43)	1.89E-06
rs12648214	4	C	0.157	0.77(0.69-0.86)	2.22E-06
rs17808098	14	A	0.474	0.84(0.78-0.90)	2.98E-06
rs1207145	2	G	0.017	0.37(0.24-0.58)	1.63E-05
rs147529871	6	T	0.02	0.44(0.30-0.64)	2.19E-05

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval.

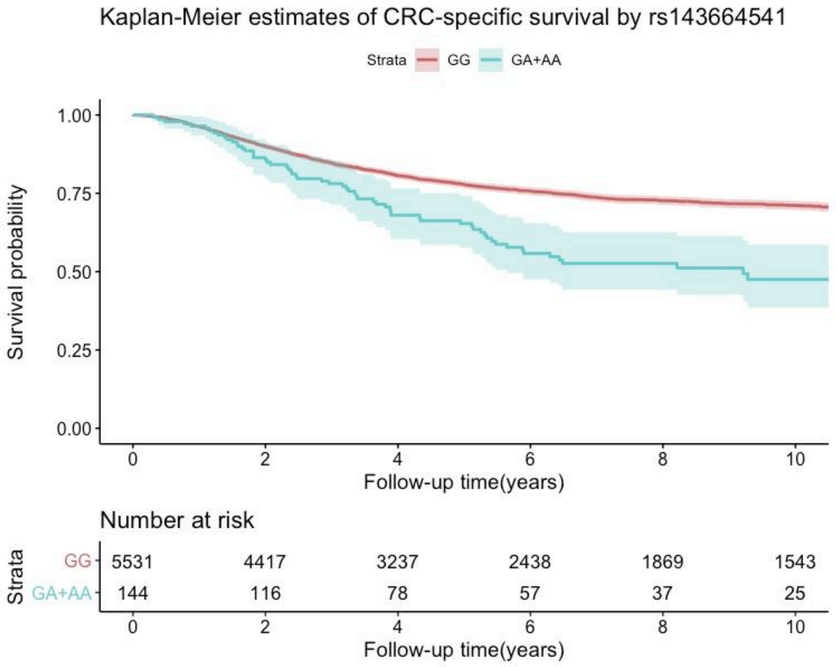


Figure 5-25 Kaplan-Meier estimates of CRC-specific survival in the SOCCS stratified by the genetic variant rs143664541

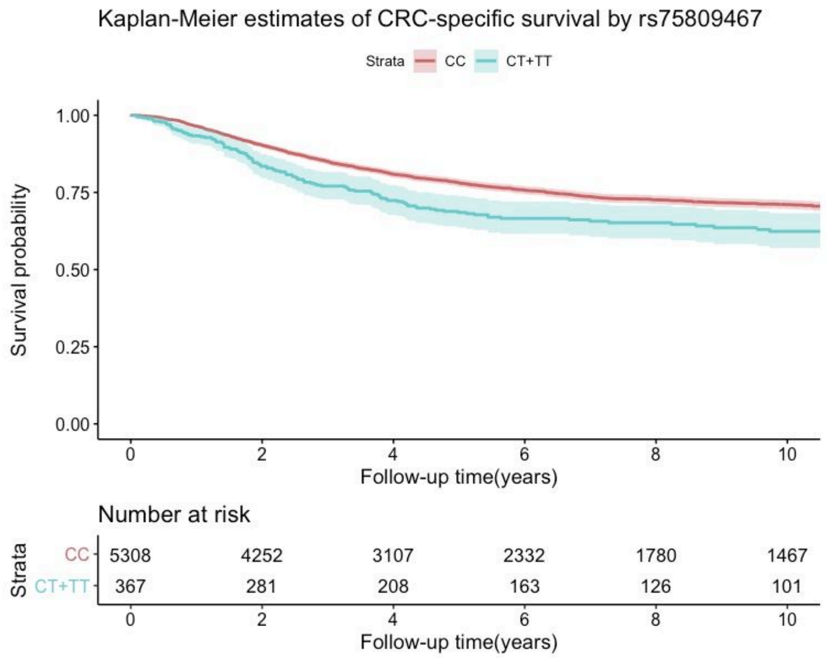


Figure 5-26 Kaplan-Meier estimates of CRC-specific survival in the SOCCS stratified by the genetic variant rs75809467

To sum up, I identified two genetic variants (rs143664541 and rs75809467) associated with survival outcomes at GWAS significance ($p<5x10^{-8}$) of CRC patients in the SOCCS study. I created locus-zoom plots to provide visualisation of the genomic region near these two variants. Both of these two variants locate in non-coding regions of the genome. **Figure 5-27** (rs143664541) and **Figure 5-28** (rs75809467) annotate flanking genes as well as the LD structure surrounding the two variants. As shown by the **Figure 5-27**, the variant rs143664541 locates in an intergenic region near the *FRK* gene in chromosome 6, whereas rs75809467 locates between the *GDA* and the *C9orf57* gene in chromosome 9 (**Figure 5-28**). The plots also suggest lack of variants in strong linkage disequilibrium (r^2>0.8) with these two variants.

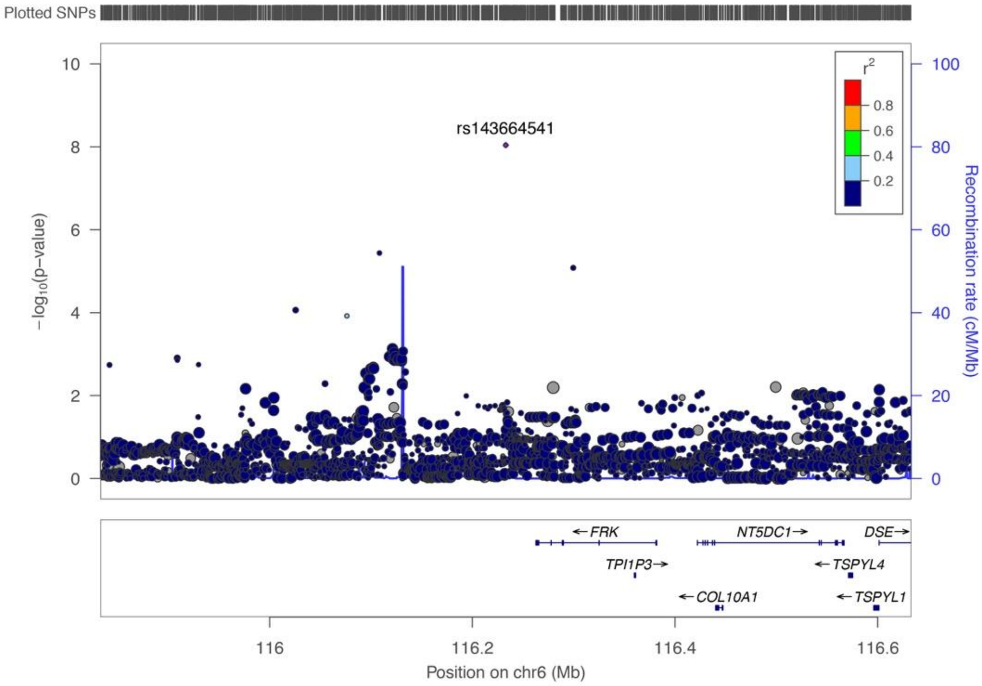


Figure 5-27 Locus-zoom plot for variant rs143664541

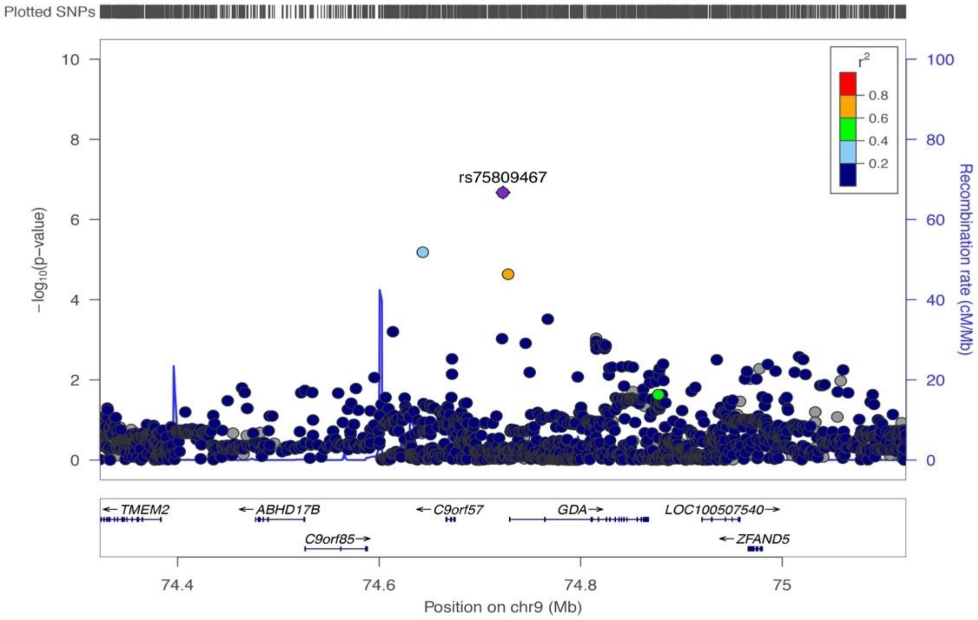


Figure 5-28 Locus-zoom plot for variant rs75809467

5.4.3 Replication of discoveries

The discoveries of two genetic variants (rs143664541 and rs75809467) were validated in the UK Biobank cohort and datasets from three published clinical trials including the VICTOR, the SCOT and the QUASAR2 trial. As described in preceding sections, the UK Biobank study cohort include 2,474 incident CRC patients. With respect to the trial datasets, in collaboration with Dr. Claire Palles (Institute of Cancer and Genomic Sciences, the University of Birmingham), we were able to collect 4,768 stage II/III patients from a pooled dataset of the VICTOR, SCOT and QUASAR2 trials for rs143664541. Eventually, the association of rs143664541 with overall survival was validated in an aggregate sample of 7,242 CRC patients. I obtained the final effect estimates by combining hazard ratios along with their standard errors from these datasets in a fixed-effect meta-analysis. Overall, I did not observe a significant association (HR=1.32, 95%CI=0.90-1.93, p=0.152). The forest plot of the meta-analysis is presented in **Figure 5-29**. It is worth noting that although no significant association was detected, the direction of effects observed in the three replication datasets was concordant with the direction from the SOCCS study—The A allele was related to worse survival outcomes. No significant heterogeneity was detected among these three datasets (I^2=0%, P_{het}=0.61).

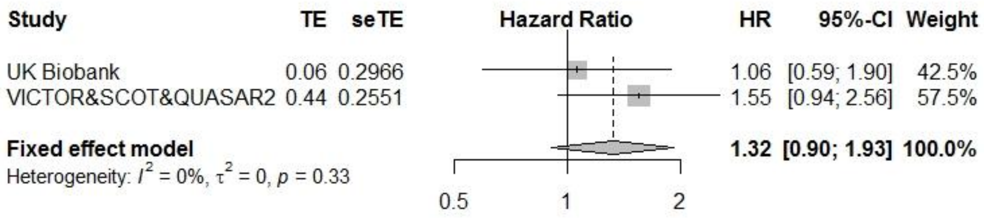


Figure 5-29 Forest plot of meta-analysis of replication datasets on the variant rs143664541

The outcome of CRC-specific survival was available only in the UK Biobank study. Similarly, I identified a point estimate of effect with concordant direction yet insignificant association between the variant rs143664541 and CRC-specific survival (HR=1.32, 95%CI=0.73-2.40, p=0.361).

Pertaining to the other variant rs75809467, we managed to include a total of 4,771 stage II/III CRC patients from the three trials along with the 2,474 patients from the UK Biobank to conduct the replication analysis. Meta-analysis found no significant effect of this variant on overall survival (HR=0.90, 95%CI=0.72-1.14, p=0.394). The point estimates of effect of this variant observed from replication datasets were inconsistent with estimates from the SOCCS. I present the forest plot of the fixed-effect meta-analysis in **Figure 5-30**. This variant was not significantly associated with CRC-specific survival in the UK Biobank cohort either (HR=0.91, 95% CI=0.66-1.25, p=0.557). No significant heterogeneity was detected among these three datasets (I^2=0%, P_{het}=0.59).

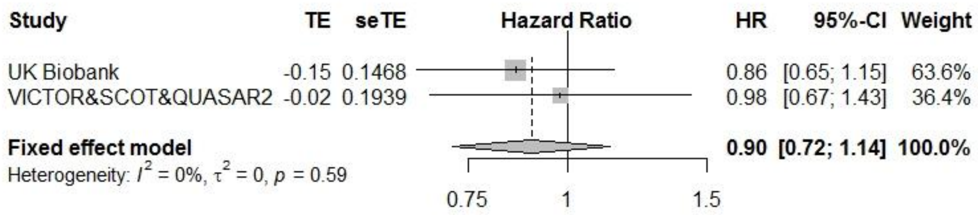


Figure 5-30 Forest plot of meta-analysis of replication datasets on the variant rs75809467

5.4.4 Gene and Gene-set based enrichment analysis

All included genetic variants were mapped to 18,420 protein coding genes by the FUMA platform (Watanabe *et al*, 2017). The mapped genes were further grouped into 15,480 gene-sets curated in the MSigDB database (Liberzon *et al*, 2015). According to the number of genes and gene-sets tested in this part of analysis, a Bonferroni corrected α level was applied to evaluate the statistical significance of the results (gene based analysis: α=2.71x10^{-6}; gene-set based analysis: α=3.23x10^{-6}).

Overall survival

Gene based analysis

For the outcome of overall survival, I did not observe any significant signals among the mapped 18,420 genes ($p<2.71\times10^{-6}$). The Manhattan plot aggregating all these test results is presented in **Figure 5-31**.

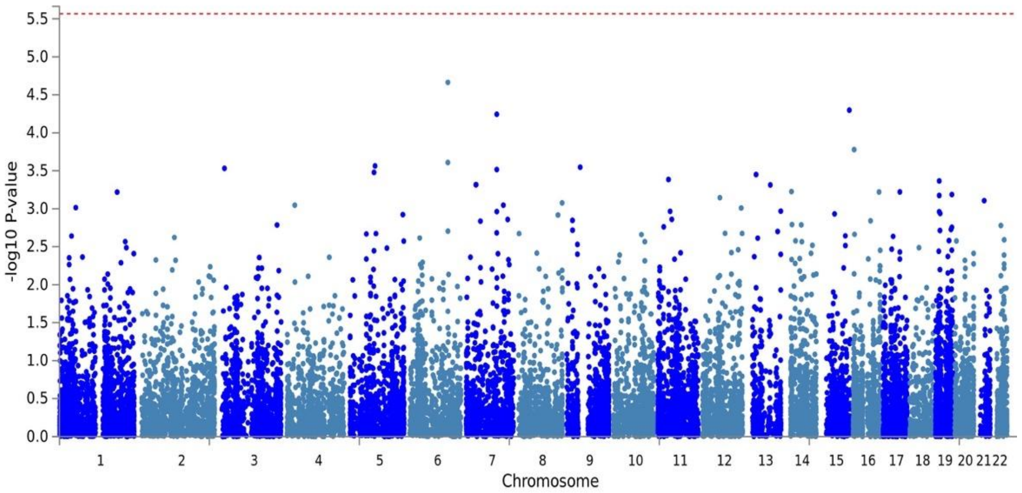


Figure 5-31 Manhattan plot of genome-wide gene based analysis on overall survival of CRC patients in the SOCCS cohort (red line: $p=2.71\times10^{-6}$)

Similar to the preceding variant-based GWA analysis, I plotted all the p-values obtained from all the gene based tests against the theoretical null distribution to examine systematic excess of statistical significance. The QQ plot is shown in **Figure 5-32**. Based on these p-values, I observed a genome-wide inflation factor of 1.017, indicating the absence of systematic inflation of the observed results ($\lambda<1.1$).

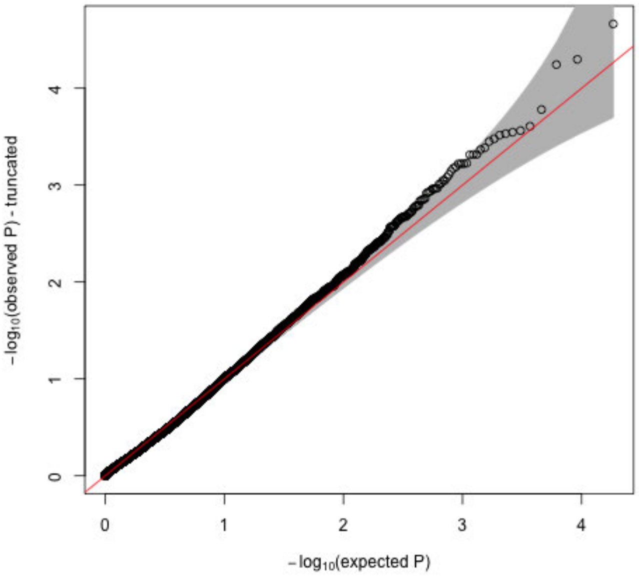


Figure 5-32 QQ plot of genome-wide gene based analysis on overall survival of CRC patients in the SOCCS cohort

Although no significant signals were detected after correcting for multiple testing, I observed one gene (*ECHDC1*) with the strongest signal associated with overall survival (p=1.58x10^{-5}). This gene locates in the chromosome 6 and encodes the protein Ethylmalonyl-CoA decarboxylase 1. A total of 122 genetic variants in this gene were included in the gene based analysis (the detailed list of variants can be found at URL4-23).

Gene-set based analysis

In relation to the gene-set based analysis, after Bonferroni correction, I identified significant enrichment of gene signals in one set of genes involved in the biosynthetic process of galactolipid (MSigDB ID: go_galactolipid_biosynthetic_process) associated with overall survival of CRC patients (p=2.09x10^{-6}). This set included six genes whose basic characteristics as well as test statistics obtained from the gene based analysis are presented in

Table 5-36.

Table 5-36 Genes involved in the biosynthetic process of galactolipid

Gene	Gene ID	Chr	No. variants	P*
GAL3ST1	ENSG00000128242	22	55	0.002
B3GALT1	ENSG00000172318	2	89	0.031
FA2H	ENSG00000103089	16	206	0.058
B4GALT3	ENSG00000158850	1	8	0.101
B3GALT2	ENSG00000162630	1	14	0.149
UGT8	ENSG00000174607	4	137	0.289

*P-values of gene based tests for each gene
Chr, chromosome

CRC-specific survival

Gene based analysis

The Manhattan plot of genome-wide gene based analysis on CRC-specific survival presented in **Figure 5-33**. As shown by the plot, I detected one statistically significant association after Bonferroni correction between the *CCDC135* gene and CRC-specific survival of patients in SOCCS. This gene locates in the chromosome 16 and encodes the Coiled-coil domain-containing protein 135. A total of 76 genetic variants in this gene were included to test the potential overall effect.

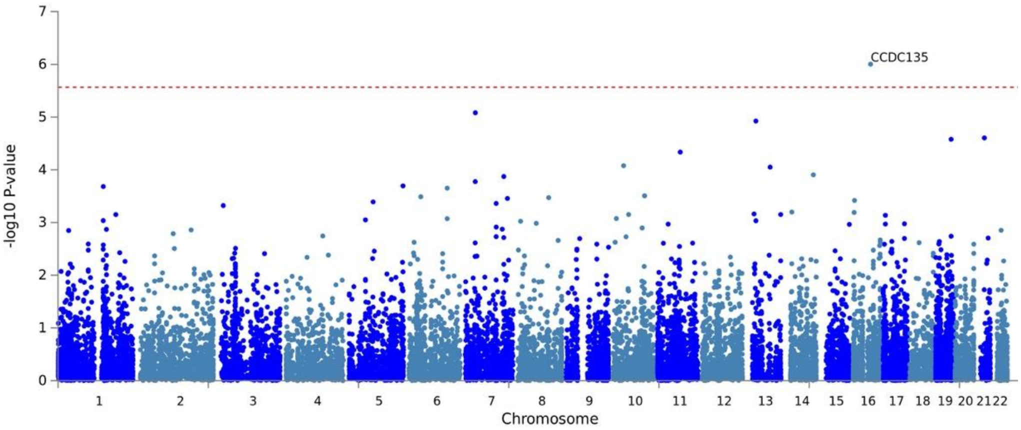


Figure 5-33 Manhattan plot of genome-wide gene based analysis on CRC-specific survival of CRC patients in the SOCCS cohort

Regarding the distribution of statistical significance, the QQ plot (Figure 5.34) provided moderate evidence of the presence of enriched signals. I obtained an inflation factor of 1.042 for the results of gene based analysis on CRC-specific survival.

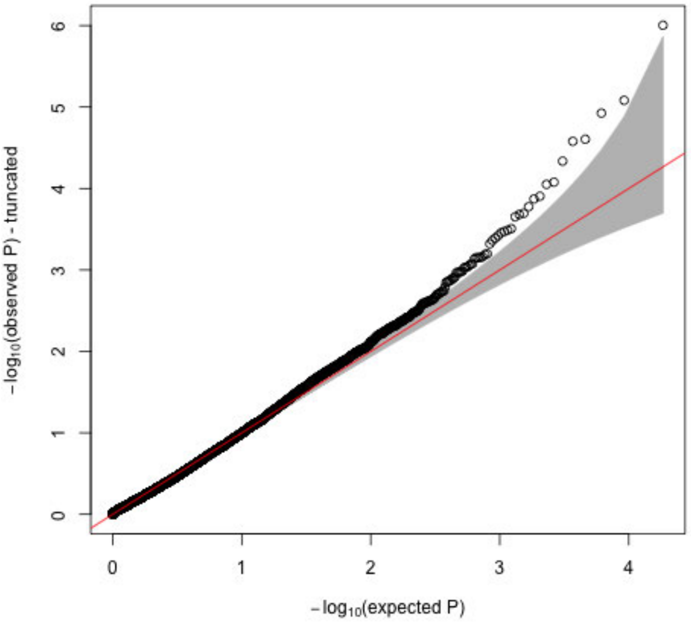


Figure 5-34 QQ plot of genome-wide gene-set based analysis on overall survival of CRC patients in the SOCCS cohort

When implementing a relatively lenient significance threshold, six genes were associated with CRC-specific survival with p<5x10^{-5}. Their basic characteristics along with the test statistic are presented in **Table 5-37**.

Table 5-37 Summary of genes associated with CRC-specific survival identified from the genome-wide gene based analysis (p<5x10^{-5})

Gene	Gene ID	Chr	No. variants	P
CCDC135	ENSG00000159625	16	76	9.92E-07
BBS9	ENSG00000122507	7	1387	8.28E-06
HSPH1	ENSG00000120694	13	22	1.19E-05
ADAMTS5	ENSG00000154736	21	143	2.49E-05
SIGLECL1	ENSG00000179213	19	64	2.64E-05
UVRAG	ENSG00000198382	11	516	4.62E-05

Chr, chromosome

Gene-set based analysis

With respect to gene-set based analysis on CRC-specific survival, I identified a statistically significant enrichment of signals in the set of genes associated with up-regulating the differentiation of adipocyte (MSigDB ID: urs_adipocyte_differentiation_up) (p=2.52x10^{-7}). This gene set includes 65 mapped genes and additional information about these genes can be found in **Table 5-38**.

Table 5-38 Genes involved in the up-regulating the differentiation of adipocyte

Gene	Gene ID	CHR	No. variants	P*
PLEK	ENSG00000115956	2	177	0.012
LYPLA1	ENSG00000120992	8	103	0.019

Gene	Gene ID	CHR	No. variants	P*
MASP1	ENSG00000127241	3	220	0.020
COL7A1	ENSG00000114270	3	22	0.022
LIPC	ENSG00000166035	15	472	0.022
DGAT1	ENSG00000185000	8	12	0.025
PCDH7	ENSG00000169851	4	1095	0.067
ECM2	ENSG00000106823	9	101	0.069
ADIPOQ	ENSG00000181092	3	48	0.079
PLIN2	ENSG00000147872	9	115	0.082
CHST1	ENSG00000175264	11	43	0.087
ALDH1A2	ENSG00000128918	15	1832	0.110
LRP8	ENSG00000157193	1	181	0.113
FXYD1	ENSG00000266964	19	4	0.118
SLC24A2	ENSG00000155886	9	760	0.130
IGFBP2	ENSG00000115457	2	23	0.152
C1orf61	ENSG00000125462	1	24	0.153
TNFAIP2	ENSG00000185215	14	27	0.188
GPD1	ENSG00000167588	12	5	0.189
VTN	ENSG00000109072	17	24	0.201
DPT	ENSG00000143196	1	147	0.203
PFKFB3	ENSG00000170525	10	312	0.231
AGT	ENSG00000135744	1	51	0.236

Gene	Gene ID	CHR	No. variants	P*
SKAP1	ENSG00000141293	17	697	0.238
PTPN21	ENSG00000070778	14	196	0.249
CAP2	ENSG00000112186	6	400	0.272
ATP2B2	ENSG00000157087	3	1349	0.278
ALDH6A1	ENSG00000119711	14	92	0.284
PTPRS	ENSG00000105426	19	645	0.294
LBP	ENSG00000129988	20	110	0.302
FABP4	ENSG00000170323	8	18	0.327
MTUS1	ENSG00000129422	8	836	0.377
STAT5B	ENSG00000173757	17	70	0.379
PPARG	ENSG00000132170	3	362	0.415
MMP7	ENSG00000137673	11	28	0.440
ATP8A2	ENSG00000132932	13	1925	0.497
ABCE1	ENSG00000164163	4	19	0.506
CIR1	ENSG00000138433	2	94	0.507
FABP5	ENSG00000164687	8	1	0.515
KCNH2	ENSG00000055118	7	74	0.524
CTSG	ENSG00000100448	14	7	0.524
DPF2	ENSG00000133884	11	23	0.541
AMT	ENSG00000145020	3	7	0.541
C8orf59	ENSG00000176731	8	10	0.559

Gene	Gene ID	CHR	No. variants	P*
APOB	ENSG00000084674	2	61	0.577
ACOX3	ENSG00000087008	4	307	0.607
ADORA2B	ENSG00000170425	17	49	0.613
HSD11B2	ENSG00000176387	16	10	0.613
E2F1	ENSG00000101412	20	13	0.621
FABP7	ENSG00000164434	6	11	0.623
GLUL	ENSG00000135821	1	26	0.639
MTMR12	ENSG00000150712	5	242	0.661
PLCD1	ENSG00000187091	3	73	0.669
MAP4K3	ENSG00000011566	2	427	0.670
INSR	ENSG00000171105	19	631	0.746
PTPRZ1	ENSG00000106278	7	443	0.769
ACSL1	ENSG00000151726	4	211	0.771
TFCP2	ENSG00000135457	12	219	0.795
CRYAB	ENSG00000109846	11	12	0.809
SMARCB1	ENSG00000099956	22	262	0.817
USP8	ENSG00000138592	15	259	0.838
LPL	ENSG00000175445	8	277	0.840
RXRA	ENSG00000186350	9	364	0.868
DGKG	ENSG00000058866	3	497	0.884
APLNR	ENSG00000134817	11	9	0.958

*P-values of gene based tests for each gene
Chr, chromosome

5.4.5 Stage-stratified GWAS

GWAS in stage II/III patients

Based on the same procedure described in aforementioned GWASs, I also performed GWAS in patients diagnosed with locally advanced (stage II/III) CRC. A total of 3,886 CRC patients in the SOCCS study with 8,328,632 autosomal genetic variants were included in this analysis. Among all included variants, I detected 402,488 (4.83%) variants associated with overall survival and 407,825 (4.90%) variants associated with CRC-specific survival respectively using the Martingale-residual based approach. The Manhattan plots displaying all variants with p<0.05 from the two GWASs on overall and CRC specific survival are presented in **Figure 5-35**.

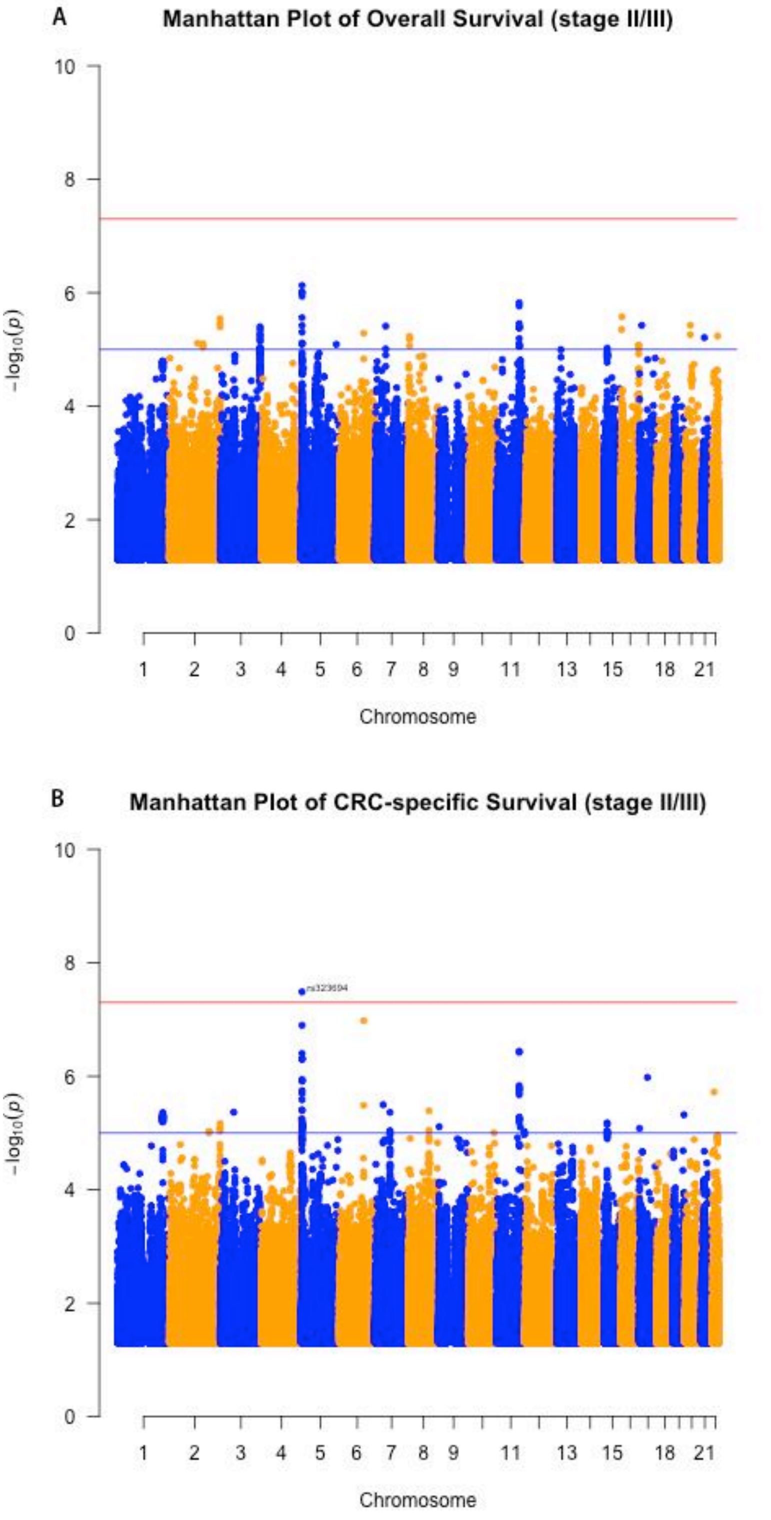


Figure 5-35 Manhattan plot of GWAS results on survival outcomes of stage II/III CRC patients in the SOCCS cohort (A for overall survival and B for CRC-specific survival; blue line: p=10^{-5}, red line: $5x10^{-8}$)

As depicted by the Manhattan plots, I identified one genetic locus in chromosome 5 that showed aggregated signals of correlated variants associated with overall and CRC-specific survival in stage II/III CRC patients. The top hit in this locus reached GWAS significance in the association with CRC-specific survival (rs323694: Martingale residual coefficient=0.287, SD=0.052, p=3.25x10^{-8})

With respect to the distribution of all genetic variants, I created QQ plots for the two outcomes and present them in **Figure 5-36**.

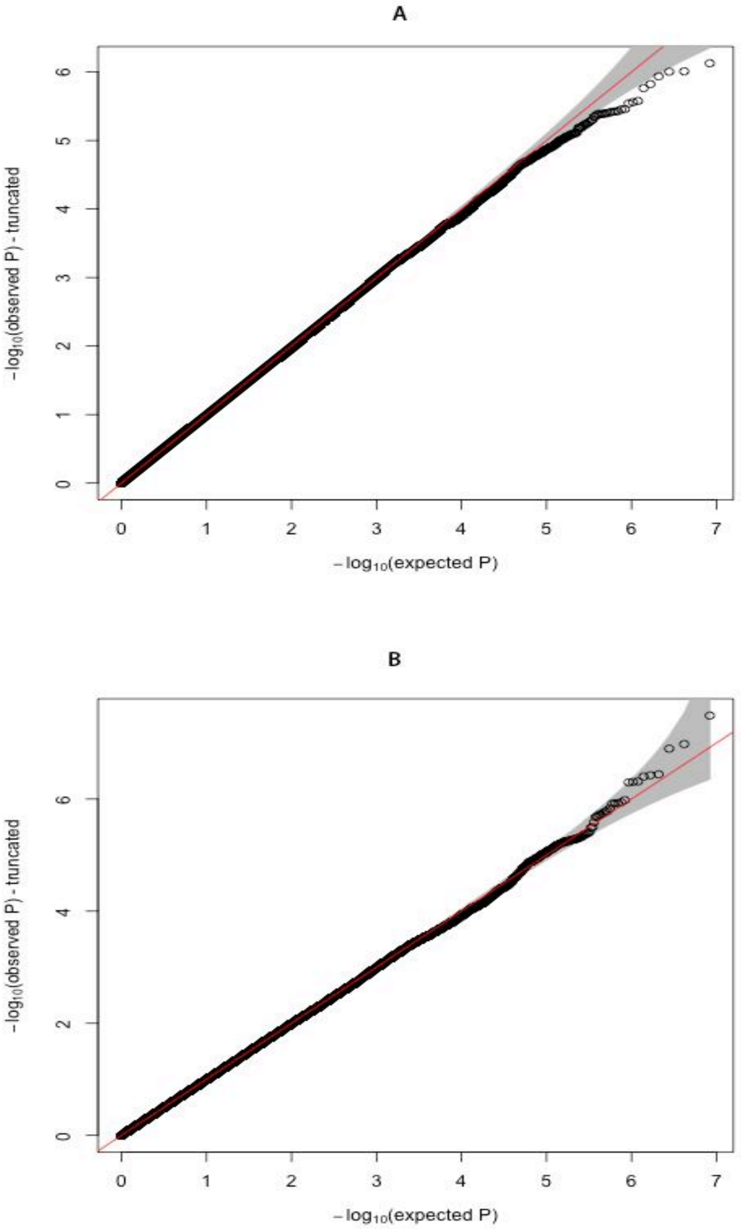


Figure 5-36 QQ plot of GWAS results on survival outcomes of stage II/III CRC patients in the SOCCS cohort (A for overall survival and B for CRC-specific survival)

The QQ plots indicate good concordance between the observed and theoretical distribution of p-values, and the inflation factors also suggest no systematic inflation of the GWAS results (overall survival: λ=0.988; CRC-specific survival: λ= 0.992).

I extracted genetic variants in association with overall or CRC-specific survival outcomes at $p<5x10^{-6}$. Initially, the GWAS on overall survival identified 24 variants prior to controlling for LD. Thirty-six correlated variants were selected for CRC-specific survival. After excluding variants in LD (r^2>0.2), I present summarised effect estimates obtained from the residual-based approach in **Table 5-39**.

Table 5-39 Genetic variants identified from the genome-wide association analysis associated with survival outcomes of stage II/III CRC patients in the SOCCS cohort ($p<5x10^{-6}$) using the Martingale-residual based approach

	Variant	Chr	MA	MAF	Beta	SE	P
OS							
	rs568921	5	C	0.446	0.202	0.041	9.77E-07
	rs7191260	16	G	0.077	-0.368	0.078	2.64E-06
	rs62184746	2	G	0.491	-0.211	0.045	2.87E-06
	rs4444042	11	G	0.043	-0.485	0.104	3.50E-06
	rs144559033	20	T	0.011	0.97	0.209	3.73E-06
	rs117392919	17	A	0.026	0.672	0.145	3.77E-06
	rs6962371	7	C	0.046	-0.484	0.105	3.88E-06
	rs6806922	3	C	0.162	-0.262	0.057	3.98E-06
	rs233176	16	T	0.308	0.213	0.046	4.45E-06
CSS							
	rs323694	5	G	0.424	0.287	0.052	3.25E-08
	rs143664541	6	A	0.014	1.258	0.236	1.05E-07
	rs201806734	11	T(Indel)	0.04	0.674	0.132	3.62E-07
	rs72832931	17	C	0.022	0.912	0.186	1.05E-06
	rs79014181	22	A	0.059	0.547	0.115	1.89E-06
	rs75796335	7	C	0.113	0.395	0.085	3.17E-06
	rs55698139	8	G	0.402	-0.244	0.053	4.07E-06
	rs6445392	3	C	0.4	-0.241	0.052	4.31E-06
	rs66494751	7	G	0.258	0.273	0.059	4.33E-06

rs10735490	1	A	0.18	-0.31	0.067	4.37E-06
rs7253608	19	T	0.306	0.274	0.06	4.78E-06

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; Beta, regression coefficients; Se, standard error; Indel, insertion-deletion variant; OS, overall survival; CRC-specific survival

Then, I re-estimated the effects of these above genetic variants on overall and CRC-specific survival in the context of Cox regression models. Notably, the variant rs323694 in chromosome 5 retained GWAS significance with an effect on CRC-specific survival (HR=1.33, 95% CI=1.20-1.47, p=2.92x10^{-8}). Detailed results of other variants are presented in **Table 5-40**.

Table 5-40 Summary of effect estimates from Cox models of GWAS-identified variants associated with survival outcomes of stage II/III patients in the SOCCS cohort (p<5x10-6)

	Variant	Chr	MA	MAF	HR(95%CI)	P
OS						
	rs144559033	20	T	0.011	2.20(1.62-3.00)	6.16E-07
	rs568921	5	C	0.446	1.22(1.13-1.33)	1.06E-06
	rs62184746	2	G	0.491	0.81(0.74-0.88)	3.03E-06
	rs117392919	17	A	0.026	1.72(1.37-2.17)	3.80E-06
	rs233176	16	T	0.308	1.23(1.12-1.34)	5.17E-06
	rs6806922	3	C	0.162	0.76(0.67-0.86)	7.28E-06
	rs7191260	16	G	0.077	0.67(0.56-0.80)	7.49E-06
	rs6962371	7	C	0.046	0.57(0.44-0.73)	1.40E-05
	rs4444042	11	G	0.043	0.81(0.67-0.98)	3.07E-02
CSS						
	rs323694	5	G	0.424	1.33(1.20-1.47)	2.92E-08
	rs143664541	6	A	0.014	2.42(1.75-3.34)	9.90E-08
	rs72832931	17	C	0.022	2.09(1.57-2.78)	3.48E-07

Variant	Chr	MA	MAF	HR(95%CI)	P
rs79014181	22	A	0.059	1.59(1.32-1.91)	1.20E-06
rs75796335	7	C	0.113	1.44(1.24-1.67)	1.88E-06
rs10735490	1	A	0.18	0.75(0.67-0.85)	3.90E-06
rs7253608	19	T	0.306	1.30(1.16-1.46)	4.16E-06
rs55698139	8	G	0.402	0.78(0.70-0.87)	4.37E-06
rs6445392	3	C	0.4	0.78(0.70-0.87)	4.53E-06
rs66494751	7	G	0.258	1.29(1.16-1.44)	4.57E-06
rs201806734	11	T(Indel)	0.04	1.33(1.05-1.67)	1.57E-02

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; OS, overall survival; CSS, CRC-specific survival

The GWAS-identified variant rs323694 is an intergenic variant located between the *IRX2* and the *LOC100506858* gene. I created a locus-zoom plot to show the LD structure and annotated genes near the variant in **Figure 5-37**. There are a number of flanking variants in moderate to strong LD (r^2>0.6) with the top variant.

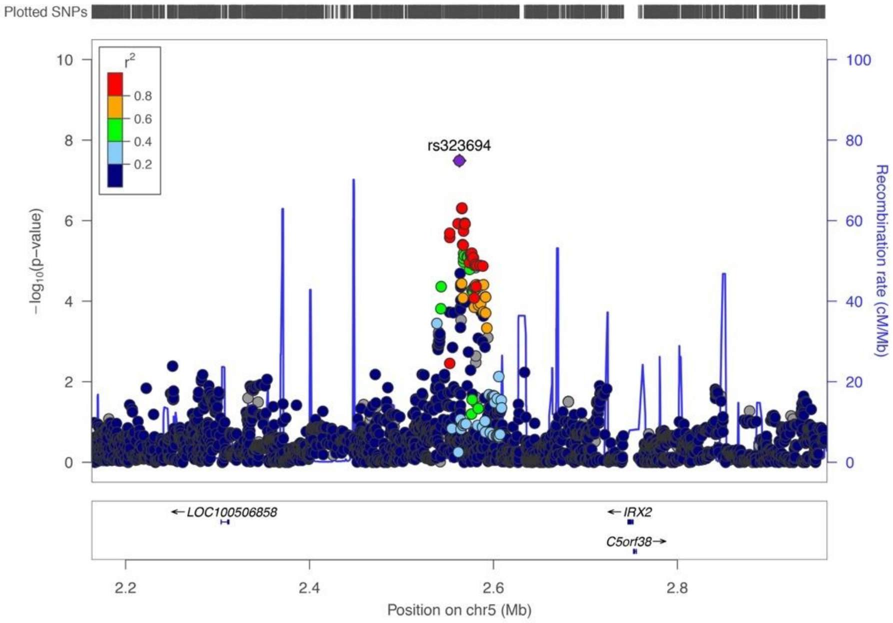


Figure 5-37 Locus-zoom plot for variant rs323694

Given that the minor allele (G) of the variant rs323694 is relatively common (MAF=0.42 in SOCCS), I plotted the Kaplan-Meier survival estimates of carriers of three genotypes separately (**Figure 5-38**). The survival curves in the figure indicate a possible recessive genetic effect of this variant on CRC-specific survival. I investigated the association between this variant and CRC-specific survival under a recessive model, but failed to detect an effect at GWAS significance (HR=1.56, 95%CI=1.33-1.84, p=1.07x10^{-7}).

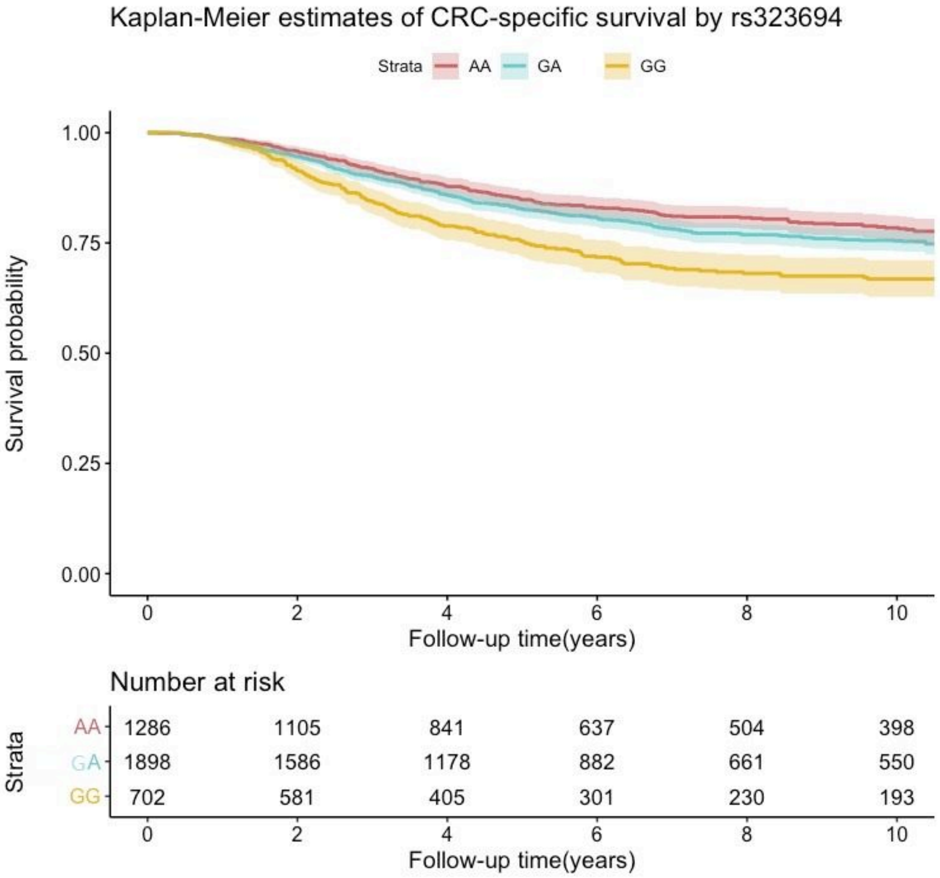


Figure 5-38 Kaplan-Meier estimates of CRC-specific survival in stage II/III CRC patients in SOCCS stratified by the genetic variant rs323694

GWAS in stage IV patients

I investigated effects of eight million autosomal genetic variants on survival outcomes of 784 stage IV patients. Using the Martingale residual-based approach, a total of 410,207(4.92%) and 413,833 (4.97%) genetic variants were associated with overall and CRC-specific survival of stage IV CRC patients ($p<0.05$) respectively. I presented these associations based on their statistical significance in Manhattan plots (**Figure 5-39**. No significant signals with $p<5x10^{-8}$ were identified in these two GWASs.

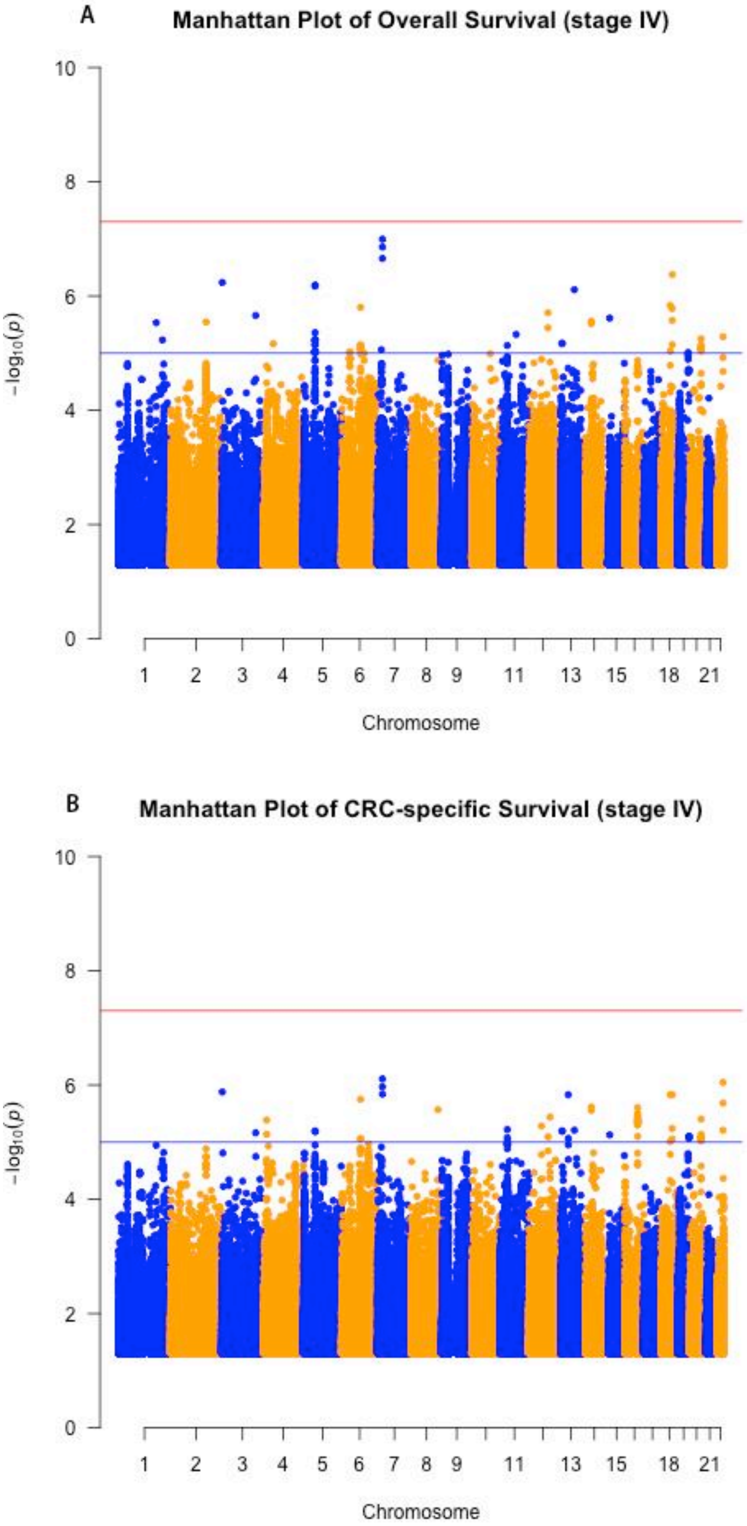


Figure 5-39 Manhattan plot of GWAS results on survival outcomes of stage IV CRC patients in the SOCCS cohort (A for overall survival and B for CRC-specific survival; blue line: $p=10^{-5}$, red line: $5x10^{-8}$)

Regarding the overall distribution of all associations of the eight million genetic variants, I estimated the inflation factors of the two GWASs on overall and CRC-specific survival of stage IV patients. A lambda value of 0.998 was observed for the GWAS on overall survival and 0.997 for CRC-specific survival. The QQ plots, presented in **Figure 5-40**, also suggested the absence of systematic inflation of statistical significance.

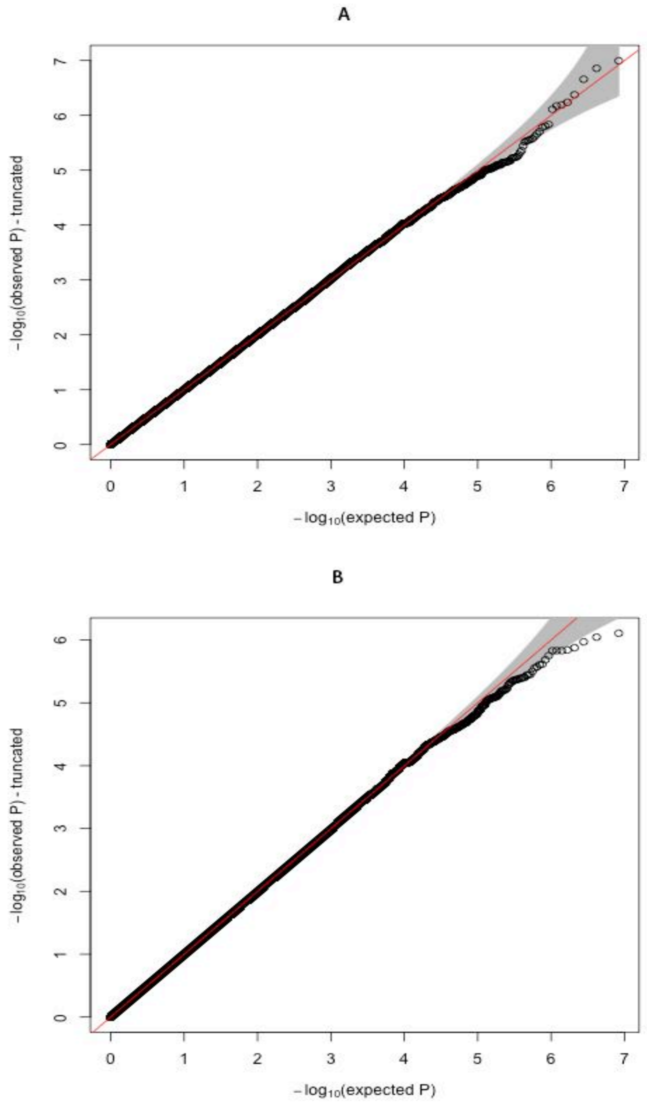


Figure 5-40 QQ plot of GWAS results on survival outcomes of stage IV CRC patients in the SOCCS cohort (A for overall survival and B for CRC-specific survival)

I also screened for associations between genetic variants and survival outcomes with $p<5x10^{-6}$. In total, the GWASs identified variants for overall survival and for CRC-specific survival. After controlling for LD, 13 independent variants (r^2<0.2) remained in association with overall survival and 11 variants were associated with CRC-specific survival. The original regression coefficients along with standard errors of these variants are summarised in **Table 5-41**.

Table 5-41 Genetic variants identified from the genome-wide association analysis associated with survival outcomes of stage IV CRC patients in the SOCCS cohort (p<5x10^{-6}) using the Martingale-residual based approach

	Variant	Chr	MA	MAF	Beta	SE	P
OS							
	rs76804061	7	C	0.019	-1.168	0.217	1.01E-07
	rs13061928	3	G	0.177	-0.39	0.077	5.79E-07
	rs78621268	5	C	0.111	0.494	0.098	6.45E-07
	rs117432600	13	A	0.011	-1.547	0.311	7.76E-07
	rs142605534	6	G	0.306	-0.312	0.064	1.57E-06
	rs111935424	18	C	0.014	-1.306	0.271	1.66E-06
	rs190425907	12	T	0.009	-1.664	0.347	1.97E-06
	rs4856722	3	C	0.014	-1.324	0.278	2.18E-06
	rs145764000	15	A	0.021	-0.986	0.208	2.44E-06
	rs79151344	2	G	0.072	-0.556	0.118	2.84E-06
	rs186078581	1	C	0.014	-1.233	0.262	2.93E-06
	rs144281883	14	A	0.104	-0.486	0.103	3.05E-06
	rs140179875	11	C	0.016	-1.26	0.273	4.70E-06
CSS							
	rs183168900	7	C	0.020	-1.128	0.227	7.81E-07
	rs35582295	22	G	0.029	-0.996	0.201	8.99E-07
	rs13061928	3	G	0.177	-0.386	0.079	1.32E-06
	rs192825132	13	A	0.011	-1.562	0.322	1.48E-06

	Variant	Chr	MA	MAF	Beta	SE	P
	rs142605534	6	G	0.306	-0.317	0.066	1.77E-06
	rs117505118	8	A	0.010	-1.569	0.332	2.70E-06
	rs144281883	14	A	0.104	-0.499	0.106	2.78E-06
	rs12928306	16	A	0.176	-0.373	0.079	3.15E-06
	rs75731204	12	T	0.364	-0.312	0.067	3.63E-06
	rs13045825	20	G	0.020	-1.07	0.23	3.94E-06
	rs74588306	4	G	0.036	-0.766	0.165	4.12E-06

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; Beta, regression coefficients; Se, standard error; Indel, insertion-deletion variant; OS, overall survival; CRC-specific survival

I then re-estimated the hazard ratios of these genetic variants in Cox regression models. However, none of the variants reached GWAS significance; the detailed results of effect estimates obtained from Cox models are presented in the following **Table 5-42**.

Table 5-42 Summary of effect estimates from Cox models of GWAS-identified variants associated with survival outcomes of stage IV patients in the SOCCS cohort (p<5x10^{-6})

	Variant	Chr	MA	MAF	HR(95%CI)	P
OS						
	rs78621268	5	C	0.111	1.59(1.34-1.89)	1.31E-07
	rs142605534	6	G	0.306	0.72(0.63-0.83)	2.24E-06
	rs13061928	3	G	0.177	0.67(0.57-0.79)	3.09E-06
	rs144281883	14	A	0.104	0.62(0.50-0.78)	2.88E-05
	rs79151344	2	G	0.072	0.59(0.45-0.76)	5.05E-05
	rs76804061	7	C	0.019	0.31(0.17-0.57)	1.41E-04
	rs186078581	1	C	0.014	0.22(0.10-0.51)	4.55E-04
	rs111935424	18	C	0.014	0.24(0.11-0.54)	4.68E-04
	rs117432600	13	A	0.011	0.20(0.08-0.50)	5.82E-04
	rs4856722	3	C	0.014	0.29(0.14-0.58)	5.83E-04
	rs145764000	15	A	0.021	0.41(0.24-0.68)	5.96E-04

	rs190425907	12	T	0.009	0.21(0.08-0.57)	2.15E-03
	rs140179875	11	C	0.016	0.73(0.43-1.26)	2.57E-01
CSS						
	rs142605534	6	G	0.306	0.72(0.62-0.82)	2.65E-06
	rs75731204	12	T	0.364	0.72(0.63-0.83)	3.95E-06
	rs13061928	3	G	0.177	0.67(0.57-0.80)	6.21E-06
	rs12928306	16	A	0.176	0.68(0.57-0.81)	1.09E-05
	rs144281883	14	A	0.104	0.61(0.48-0.77)	2.62E-05
	rs35582295	22	G	0.029	0.32(0.19-0.56)	4.57E-05
	rs74588306	4	G	0.036	0.46(0.31-0.69)	1.63E-04
	rs183168900	7	C	0.02	0.34(0.19-0.60)	2.68E-04
	rs13045825	20	G	0.02	0.34(0.19-0.62)	3.54E-04
	rs192825132	13	A	0.011	0.12(0.04-0.44)	1.15E-03
	rs117505118	8	A	0.01	0.16(0.05-0.50)	1.56E-03

Chr, chromosome; MA, minor allele; MAF, minor allele frequency; HR, hazard ratio; CI, confidence interval; OS, overall survival; CSS, CRC-specific survival

5.5 Summary

In this chapter, I presented my results grouped in four analytic sections. Firstly, in the descriptive analysis, survival outcomes of prevalent cases in the UK Biobank were found to be significantly better than incident cases, therefore the prevalent cases were excluded from subsequent analysis due to potential selection bias. The first part of the thesis is a replication study of 43 genetic variants reportedly associated with CRC survival. None of these previously identified associations were successfully replicated after adjusting for multiple testing in the SOCCS cohort. In the second part, a genetic predictor combining these 43 known genetic variants was developed in the UK Biobank, and it showed weak predictive value on 5-year survival outcomes of patients in the UK Biobank. The external validation found no meaningful predictive performance of the developed genetic predictor in the SOCCS cohort. In addition, the genetic predictor did not appear to add extra predictive value to other non-genetic variables in SOCCS. In the third part, 128 genetic variants associated with CRC risk and 82 variants previously linked with survival outcomes of other cancers were tested in terms of their potential effects on CRC survival. None of these variants showed significant effect on either overall or CRC-specific survival of CRC patients in the SOCCS under an additive genetic model. A possible recessive genetic effect on survival outcomes was identified for two CRC-risk variants (rs10161980 and rs7495132) in both the SOCCS and UK Biobank cohorts. In the last part of the thesis which features a genome-wide association study on CRC survival, the variant rs143664541 was identified to be significantly associated with both overall and CRC-specific survival; another variant rs75809467 showed a significant effect on CRC-specific survival in the SOCCS cohort. However, neither of them remained significant in the replication analysis based on meta-analyses combining the UK Biobank cohort and three clinical trial datasets. By conducting gene and gene-set based analysis using the SOCCS cohort, I observed potential enrichment of genetic signals in the *CCDC135* gene for overall survival, and in two gene sets involved in biosynthetic process of galactolipid (overall survival) and up-regulating the differentiation of adipocyte (CRC-specific survival) respectively.

Chapter 6 Discussion

6.1 Introduction

This chapter is divided into two parts: discussion of the methodological perspectives and interpretation of the main findings. Firstly, strengths and limitations of the study design, data sources and statistical analysis will be discussed. In the second part, main results derived from Chapters 3 and 5 will be summarised, and the interpretation of these results will be discussed in relation to the strength of evidence and potential biological implications. Genetic variants, genes and gene sets identified from this thesis that may be associated with survival outcomes of CRC will be discussed separately. Finally, conclusions and recommendations for future research will be summarised.

6.2 Methodological perspectives

6.2.1 Study design

Systematic literature review and meta-analysis

The main strengths and limitations of the published systematic literature review have been presented in the discussion section of Chapter 3 (page 66). In this section, main points in relation to the study design of the systematic review will be summarised and aspects that have not been extensively discussed in the published paper will also be described.

Study selection

I searched for published literature using the MEDLINE and Embase databases. The literature search was restricted to articles in English and, therefore, prediction models published in other languages were not reviewed. With respect to the inclusion and exclusion criteria, only multivariable prediction models with no less than two predictors were included. This resulted in the exclusion of studies investigating the predictive value of a single factor. In order to distinguish prediction models from simple association studies, only published papers that reported quantitative measures of model performance were deemed as eligible in this review. Moreover, I also excluded studies in which specific prediction rules were not reported or the prediction tools such as an online calculator were unavailable because a prediction model could not be independently validated and updated unless the prediction rules were explicitly presented.

Critical appraisal

In this review I used the CHARMS checklist to appraise the methodological quality of included prediction models (Moons *et al*, 2014). This checklist evaluates each model for six domains: participants, predictors, outcome, events per variable (EPV), cohort attrition and data analysis (see details in **Appendix Table 1**). This checklist is not specifically designed for studies on time-to-event outcomes and, therefore, survival related biases such as ‘health volunteer’ bias (further explained in the section ‘Selection bias’, page 236) cannot be assessed using this checklist. Given that critical appraisal of prediction models is based on personal judgment and can be inevitably subjective, a second reviewer, Dr Xue Li from the Usher Institute, conducted a parallel evaluation in a random sample of 25% of included models. Three models (9%) were identified with discrepant assignments of risk of bias, but agreement was reached after discussion. Based on the CHARMS checklist the same research team proposed an updated tool named ‘PROBAST’ to appraise prediction models (Moons *et al*, 2019). The new tool integrates two previous domains (EPV and cohort attrition) into the data analysis domain and thus only four domains are used to assess each model. More importantly, the new tool suggests assigning an overall risk of bias to each model. In particular, only models with low risk of bias for all domains can be assigned with low risk, which will not change the main finding of the systematic review where only two models reported by Rees at al. were classified as overall low risk of bias(Rees *et al*, 2008b).

Meta-analysis

I performed a random-effects meta-analysis to combine metrics evaluating model performance considering potential heterogeneity across multiple studies under varied clinical settings. A strength of this review is that the meta-analysis was conducted using only external validation studies and the initial model development study was excluded to avoid overestimation on the model performance. In this study, I was only able to conduct meta-analysis for reported C statistics based on data availability. Currently, there have been multiple methods to estimate the C statistic for time-to-event data (Blanche *et al*, 2013). For example, it can be calculated as the area under the receiver operating characteristic curve (ROC) at a specific observation point. In addition, it can be derived from the definition—the proportion of randomly selected pairs of individuals in which an individual with longer observed survival time shows higher predicted survival probability. Other methods such as time-dependent ROC can also be applied to calculate the C statistic (Blanche *et al*, 2013). However, detailed methods used to derive the C statistics were mostly not reported among included prediction models. Hence, I was unable to harmonise the reported C statistics and the pooled estimates could be inaccurate. Another limitation for the meta-analyses of prediction models is that quantitative methods to detect potential risk of bias have not been well developed. In a previously published umbrella review, we appraised the meta-analysis by applying a range of metrics including heterogeneity, small study effects, excess significance, prediction interval and credibility ceilings (He *et al*, 2018a). However, currently, only the prediction interval, which indicates possible range for the performance of a future validation study, has been recommended when conducting meta-analysis of prediction models (Debray *et al*, 2017). In the systematic review of this thesis, I calculated the 95% prediction intervals of C statistics when more than three validation studies were available. None of the models were identified with prediction intervals excluding the null (0.50), indicating the presence of substantial amount of uncertainty regarding the discriminative performance of future validation studies. Whether other metrics can be used to evaluate evidence strength of prediction models, still remain unclear. Future efforts should explore possible integration of multiple quantitative methods to appraise meta-analyses of prediction models.

Data sources for genetic association studies

Study cohorts

In this thesis, multiple CRC patient cohorts were employed. Of them, the SOCCS study is used as the main discovery cohort of genetic association studies. SOCCS is a prospectively collected population-based cohort that recruits patients from Scotland after their CRC diagnosis and then follows them up. The fact that they are already diagnosed with CRC may influence levels of specific biomarkers and the way they respond to questions. However, this should not affect the genotypic, demographic (age and sex) and pathological (AJCC stage, tumour site and tumour grade) variables included in the analysis. The other study cohort used in this thesis is comprised of incident CRC cases from the UK Biobank, which is also a prospectively collected population-based cohort in UK. As opposed to SOCCS, participants developed CRC after they entered the study. Prospective cohorts are considered as a preferable study design compared to retrospective cohorts, which are prone to more potential sources of biases, especially in the setting of investigating prognostic effects.

Selection bias

Notably, the prevalent CRC cases from the UK Biobank were excluded from analysis. As shown in Figure 5.1 and 5.2, prevalent CRC cases showed significantly better survival outcomes compared to incident cases from the UK Biobank as well as the SOCCS study. This is known as ‘left truncation’ or ‘survival bias’ in survival analysis. Left truncation happens when participants conditioning upon them not having experienced the event of interest are included. In this case, prevalent CRC cases from the UK Biobank were only included given they were alive at recruitment. This causes exclusion of prevalent cases who had inferior survival outcomes (died before recruitment) and therefore can introduce bias especially when comparing the results to general population. The SOCCS study is not severely affected by survival bias because all CRC patients were enrolled upon diagnosis, and similar survival estimates were found for SOCCS patients compared with incident cases in UK Biobank (Figure 5.1 and 5.2). There have been statistical models developed to accommodate left-truncated data. For example, the PROC PHREG module (URL6-1) can provide adjusted survival estimates accounting for the truncation time—in this case, the time from CRC diagnosis to enrolment of the prevalent cases. However, the methodology of dealing with left truncation in predictive modelling and large-scale analysis such as GWAS has not been well developed. Therefore, prevalent UK Biobank cases were excluded prior to analysis so as to derive unbiased estimates.

After excluding prevalent cases in UK Biobank, a 5-year overall survival rate of 75% and 74% was observed for the UK Biobank and SOCCS study respectively. These survival rates are higher than the 5-year survival rates provided by Scottish Cancer Registry (61%) (URL1-2), CRUK (60%) (URL1-3) and NCI (65%) (NCI, 2018) under the same follow-up scheme (from CRC diagnosis to death). By investigating the stage-specific 5-year survival estimates, I observed comparable survival rates for stage I (93%) and stage II (83%) patients in SOCCS compared with estimates provided by CRUK (95%-100% for stage I and 80%-90% for stage II patients). However, higher 5-year overall survival rates were found for stage III (72%) and stage IV (27%) patients in SOCCS than the registry data from CRUK (65% for stage III and 5%-10% for stage IV patients). It should be noted that survival rates provided by cancer registry are relative rates which take into account the expected death rates of the general population with similar age but without the disease. This makes the relative survival rates slightly higher than the actual overall survival, which does not change the fact that survival rates estimated from the SOCCS and UK Biobank cohorts are higher than registry data, especially for patients with stage III and IV CRC. These observed discrepancies may be explained by the ‘healthy volunteer’ selection bias—individuals with worse health status are less likely to participate in such cohort studies (Fry *et al*, 2017). This bias can also be reflected by the observed percentage of stage IV patients in SOCCS (13.8%) which is evidently lower than the population percentage (20%) provided by CRUK (URL1-3).

In addition to the SOCCS and UK Biobank cohorts, three published clinical trials were used in this thesis as part of the validation analysis. As described in section 4.2.3, the three trials included stage II and III CRC patients with varied chemotherapy strategies. Although, a total of 4,768 patients were provided by these trials, these datasets may not be fully representative of the CRC patient population. Firstly, these trials targeted stage II/III CRC patients, which only represent a subpopulation of all CRC patients compared to the SOCCS and UK Biobank cohorts. Secondly, included CRC patients were highly selected based on predefined inclusion criteria, rendering these datasets prone to selection bias. For instance, all the trials included CRC patients with performance status scores of 0 or 1 (Iveson *et al*, 2018; Kerr *et al*, 2016; Midgley *et al*, 2010), meaning that sick patients who needed extra assistance (scores of 2, 3 and 4) were excluded. Thus, these datasets could be severely influenced by the 'healthy volunteer' bias. Another issue is that patients allocated to the treatment arm received exploratory treatment strategies, which could also lead to varied survival rates compared to the patients with standard treatment. Therefore, findings derived from these three datasets merit further validation in future large population-based cohorts.

Although the observed survival rates of the study cohorts included in the thesis may not be fully representative of the CRC patient population, the relative effect estimates (HRs) of genetic and non-genetic factors should still be generalizable (Fry *et al*, 2017). However, in terms of prediction models reported in the thesis, one should be vigilant given the presence of the aforementioned selection bias. As introduced in Chapter 5 (page 141), baseline survival rates and relative effect estimates for candidate predictors are both required to obtain the predictive survival probability for a given individual. Although the relative effect estimates reported in Chapter 5 can be directly applied to make predictions, the baseline survival rates could be biased due to the study cohorts being unrepresentative. Therefore, one should update the baseline survival rates according to the target population before applying these prediction models.

Main variables

Genetic variables---Standard genotyping arrays along with rigorous quality control and imputation processes were used for both cohorts (details presented in page 70 and 79, Chapter 4). The genotype data of these two cohorts have been included in multiple published GWASs (Dunlop *et al*, 2012; Law *et al*, 2019). In this thesis, germline genetic variations were tagged by single nucleotide polymorphisms (SNPs) detected by SNP arrays. The main strengths of SNP arrays include their high resolution to assay millions of SNPs throughout the genome in one experiment and lower cost compared to other high throughput technologies such as sequencing. However, there are also limitations for SNPs arrays. For example, these arrays are unable to efficiently recognise chromosomal anomalies such as balanced translocations (with no missing or extra chromosomal materials) and inversions which are also possible germline genetic markers for disease outcomes (Mao *et al*, 2007).

Analytical tools have been developed to identify some chromosomal changes such as aneuploidies (abnormal chromosome copy number) (Ting *et al*, 2006). However, genetic data prepared for GWASs were used in this thesis; therefore, other germline genetic markers such as chromosomal changes were not employed. Future exploration may consider investigating germline markers other than SNPs in relation to their prognostic effects on CRC survival. Genotype imputation was performed for all the study cohorts used in this thesis in order to expand the coverage and improve statistical power of GWASs. In particular, the SHAPEIT-IMPUTE2 method (technical details are presented in page 73, Chapter 4) was adopted to impute the un-genotyped loci. Roshyara et al. compared performance of different imputation methods in a cohort of 2,500 individuals (Roshyara *et al*, 2016). Their results showed that the SHAPEIT-IMPUTE2 method might overestimate the certainty of genotype distributions, but this effect diminished as the sample size increased (Roshyara *et al*, 2016). Given the relatively large sample size for both the SOCCS (~20K) and UK Biobank (~500K) cohorts used for imputation, this effect should not be problematic for the imputed genotype data of this thesis. As genotypes are estimated based on the LD structure rather than measured genotypes, imputation error is therefore inevitable. Imputation error leads to increased uncertainty and causes loss of statistical power for an imputed SNP compared to the same SNP if genotyped (Huang *et al*, 2009). However, previous evidence found that imputation error rates are generally low (2%-6%) for widely used methods including the SHAPEIT-IMPUTE2 approach (Pei *et al*, 2008).

Non-genetic variables---The SOCCS cohort included age at diagnosis, sex, AJCC stage, tumour grade and tumour site, whereas only age at diagnosis and sex were available in the UK Biobank cohort. Demographic variables such as age and sex can be generally considered to be accurately measured. Tumour site and grade for patients from SOCCS were extracted from pathology reports and therefore, measurement error should also be rare.

As the strongest predictor of survival outcomes, the AJCC stage was assigned according to pathology reports for patients who underwent surgical resection, whereas patients without surgery were staged based on imaging prior to treatment. As described in page 114 Chapter 5, ten stage 0 rectal cancer patients were identified and excluded from analysis as they might have received neoadjuvant radiotherapy (applied before surgery) which could shrink the tumour and result in a downgraded stage in the pathology report. However, the remaining rectal cancer patients (N=2,201) in SOCCS were not manually checked to identify candidates with shrunken tumour after neoadjuvant radiotherapy. This led to inclusion of rectal cancer patients with possible underestimated stage, which could potentially bias the analysis. In addition, the AJCC stage is in fact a combination of three different measures (T, N and M). There has been evidence showing that modelling the TNM measures separately may improve prediction performance compared to the numeric AJCC stage (Li *et al*, 2014). Given that the focus of this thesis is to examine the potential added predictive value by germline genetic variants, the numeric AJCC stage was, therefore, directly used to predict CRC survival.

There have been other factors associated with CRC survival. For example, the Canadian Cancer Society recommends lympho-vascular invasion, histological type of CRC, CEA, MSI, bowel obstruction or perforation and somatic mutations including *KRAS* and *BRAF* mutations as prognostic indicators for CRC (URL1-9). These factors are not collected in SOCCS and have not been released in UK Biobank. Future efforts are expected to aggregate patient cohorts with more of these prognostic factors recorded to improve prediction of CRC survival. Another important factor that can significantly influence survival outcomes is the treatment that patients received. Adding details of the treatment into the model can not only improve predictive performance, but also can assist accurately estimating effects of genetic variants. There are ongoing efforts in our group in linking the SOCCS data to external databases maintained by oncologists to obtain treatment data for patients in SOCCS. Treatment effects will be properly modelled in future investigations.

Genetic association studies

Variant selection for candidate association studies

The first step of the thesis was a systematic literature review aiming to summarise all published prediction models used to predict CRC survival and to identify potential germline genetic markers employed as predictors. No germline genetic variants were used in the identified published prediction models. Given that only prediction models with quantitative assessment of model performance were included in the review, genetic variants investigated in genetic association studies were not reviewed in this step. In the second step, I validated previously reported variants that were associated with CRC survival outcomes using the SOCCS cohort. Variants were retrieved from the GWAS catalogue. By searching the GWAS catalogue, the eligible variants were restricted to ones identified by published GWASs. Therefore, genetic variants reported by candidate association studies were not considered. These variants were often selected based on established or putative biological mechanisms hence they may truly influence CRC prognosis. However, candidate association studies are hypothesis-driven and previously published studies could be subjected to multiple sources of biases (Ioannidis *et al*, 2001). For example, it has been reported that published candidate genetic association studies showed prominent excess of statistical significance which could possibly be attributed to publication bias (Kavvoura *et al*, 2008). Moreover, approximately 90% of previous candidate association studies did not control for multiple testing in a random sample of studies from 2001 to 2003, indicating pervasive false positive findings (Yesupriya *et al*, 2008). A field synopsis can systematically summarise and critically appraise published candidate association studies. A latest field synopsis on CRC risk, conducted by our group, identified 18 variants with high credibility, 72% of which were validated by subsequent GWASs (Montazeri *et al*, 2019). Hence, a field synopsis on survival outcomes is needed as more candidate association studies are published before further validation analysis can be planned. As for the variants retrieved from the GWAS catalogue, a relatively lenient threshold ($p<10^{-5}$) was used to screen for variants, which potentially raised the probability of type I error. In order to control for LD among eligible variants, an $r^2<0.2$ was used to exclude variants in LD. This could potentially cause exclusion of variants with larger effects on CRC survival in SOCCS. The aforementioned caveats about variant selection also apply to the other two candidate association studies in this thesis based on two hypotheses. The first hypothesis is that genetic variants known to be associated with CRC risk have subsequent effect on CRC progression and therefore serve as a potential prognostic indicator. The second hypothesis is that there exists a shared genetic basis of tumour progression across multiple cancer types, and variants associated with survival outcomes of other cancers also influence CRC survival. The rationale of these two hypotheses is described in Chapter 2. Due to a small number of variants being tested at a time, hypothesis-driven association studies are advantageous over GWAS in terms of statistical power when a relatively small sample size is available. In addition to the variants selected in this thesis, there are other candidate variants possibly associated with CRC survival. For example, genetic variants may impact survival outcomes of CRC through modifying patients' response to therapeutic agents such as 5-fluorouracil which has been widely used in chemotherapy for CRC patients. Previously published GWASs have identified a number of genetic variants to be associated with response to 5-fluorouracil (Chung *et al*, 2013; Low *et al*, 2013). These variants could be tested in a separate hypothesis-driven association study to further examine their impact on survival outcomes of CRC.

GWAS

Under no prior hypothesis, GWAS has been widely considered as a more comprehensive and unbiased approach compared with candidate association studies (Eberle *et al*, 2007). It facilitates identification of novel genetic risk loci that have not been linked to the outcome of interest, revealing possible new biological mechanisms involved in the disease. One of the main strengths of GWAS is the high reproducibility of its findings. Marigorta et al. reported that an estimate of 40% to 94% of GWAS-identified variants have been successfully replicated by later GWAS meta-analyses (Marigorta *et al*, 2018). In addition, GWAS can also shed light on the overall genetic architecture of a given trait by displaying the distribution of statistical significance of genetic signals throughout the genome. Despite clear advantages of GWAS, there are also limitations for this study design. For example, GWAS is conducted in a limited number of populations—some ethic groups are therefore not represented. Although germline genetic variations are generally not influenced by environmental factors, population stratification can confound the observed genetic associations through varying genetic allele frequencies (Hellwege *et al*, 2017). However, results of the PCA analysis indicated that population stratification was unlikely to have major impact on studies in this thesis (see Chapter 4, page 71). In addition, GWASs are penalised by multiple testing with millions of genetic variants being examined. A GWAS based on limited sample size can result in elevated rates of type II error (false negative findings), especially for variants of low MAF with small effects on the outcome. As shown in Chapter 5 (page 198), the GWAS based on 5,675 CRC cases in SOCCS is underpowered to detect small to moderate genetic effects (HR<1.4) for variants with MAF<0.05. This limitation can be overcome by future collaborative efforts aggregating multiple CRC cohorts to achieve improved statistical power. Another strategy is to reduce the number of tests; this includes approaches adopted in this thesis such as aforementioned candidate association analyses and gene or gene-set based analysis. An additional limitation for GWASs based on SNP arrays is that effects of rare genetic variants cannot be effectively detected (Tam *et al*, 2019). Although a huge number of un-genotyped variants are recovered through imputation, such rare variants are hard to be accurately imputed. Ongoing efforts of whole genome or whole exome sequencing are expected to unravel possible rare variants that influence CRC survival.

Validation analysis

In this thesis, a two-stage study design was employed for both candidate association studies and the GWAS. To be specific, the SOCCS study was used as the primary discovery dataset, whereas the UK Biobank cohort was used for replication of findings from candidate association studies and meta-analysis of UK Biobank and three clinical trials were used to validate GWAS-identified variants. Replication in an independent dataset is crucial in genetic association studies in terms of avoiding false positive findings. In addition to statistical significance by chance, possible non-random biases in the discovery set can also be detected by independent replication. For example, if the SNP array consistently generates incorrect genotypes for a specific variant, this cannot be corrected by accumulating more samples. However, an independent dataset genotyped by a different array can effectively detect these errors. Biased measurement may also occur at the outcome level. In this case, follow-up of CRC patients for each cohort was conducted by different teams, and therefore spurious findings driven by possible biased survival outcomes are unlikely to be validated in an independent cohort. Although combining all available cohorts into a discovery set would markedly increase the sample size, the two-stage approach was adopted in this thesis given that diverse clinical settings and different structure of available covariates could introduce substantial amount of heterogeneity into meta-analysis.

Study outcomes

Overall survival and CRC-specific survival were employed as study outcomes in this thesis. The definitions along with the strengths and limitations of different endpoint such as overall survival and disease-free survival are introduced in Chapter 1 (page 29). Currently, overall survival is still widely accepted as the gold standard endpoint in prognostic studies of oncology (Sargent *et al*, 2005). It is worth mentioning that the survival time for patients in SOCCS was defined as the time span from the date of definitive treatment (surgery or chemo/radiotherapy) to the date of death. Whilst the UK Biobank cohort used the date of diagnosis obtained from the Cancer Registry as the starting point as the date of definitive treatment was unavailable, and the endpoint (date of death) was retrieved from National Death Registry. Cancer Registry in UK defines the date of cancer diagnosis following the algorithm provided by the European Network of Cancer Registries (see details in the previous publication (Tyczynski *et al*, 2003)). The date of CRC diagnosis contains a mixture of following event dates in a declining order of priority: date of first histological confirmation, date documented in pathology report and date of first admission to hospital due to CRC. Given the fact that the date of CRC diagnosis should be no later than the start of treatment, the survival rate estimates would have been lower if the date of definitive treatment was used as the starting point. As discussed in the previous section (page 244), this should not bias the relative effect estimates (HR) if the time interval between diagnosis and treatment is independent from the genetic variant (Fry *et al*, 2017). However, future efforts are still expected, if possible, to re-calibrate the prediction model derived from the UK Biobank after harmonising the starting point of follow-up. As with the SOCCS study, the other three trials used in this thesis defined the follow-up time as from the date of randomisation (or the date treatment started) to the date of death. Date of definitive treatment is widely used in clinical settings as the starting point to evaluate the effect of certain treatment. The length of the delay is highly variable depending on the status of the disease and different medical systems over the world. Using the date of definitive treatment instead of diagnosis as the starting point will result in lower absolute survival estimates. There has been evidence showing that this effect of treatment delay on survival outcomes could be attenuated as the observation time period increases, especially when the follow-up time is longer than 3 years from the date of diagnosis (Roder *et al*, 2019). Hence, treatment delay should be less problematic given the relatively long median follow-up time (⩾3 years) for all the study cohorts used in the thesis.

It should be noted that assigning the specific cause of death can be subjective, although a parallel independent evaluation was performed by a colorectal surgeon and a low percentage of discrepancies was found between the two reviewers (1%). For example, a patient who died from myocardial infarction would be considered as a non-CRC related death based on our criteria (page 76, Chapter 4). However, the myocardial infarction could possibly be induced by chemotherapy, which could not be reflected in the death certificate. Different criteria of assigning cause of death can lead to varied observed survival estimates for CRC-specific survival. The primary cause of death for participants from UK Biobank was assigned by the UK Biobank and was retrieved directly for analysis.

Summary of strengths & limitations for study design

Strengths:

Systematic literature review:

*A comprehensive systematic literature and meta-analysis was conducted to identify potential genetic variants that had been used in published prediction models, and to quantitively evaluate performance of published models.

*Meta-analyses of C statistics reported in published prediction models were only conducted using external validation studies, which avoided inflated estimation of model performance.

Genetic association studies:

*The main study cohorts including the SOCCS and UK Biobank cohorts are population-based and prospectively collected.

*Rigorous quality control was conducted for genotyping and imputation of the genotype data for the SOCCS and UK Biobank cohorts.

*Both hypothesis-driven (candidate association studies) and hypothesis-free (GWAS) approaches were adopted to identify genetic variants associated with CRC survival.

*A two-stage study design was adopted which consists of a discovery set and independent validation sets for genetic variants identified by the candidate association studies and the GWAS.

Limitations:

Systematic literature review:

*The systematic literature review only included published prediction models, which reported quantitative measures of model performance. Therefore, predictors used by studies that did not report model performance might have been missed.

*Meta-analyses on metrics other than C statistics were not conducted due to data availability reported by published models.

Genetic association studies:

*Exclusion of prevalent CRC cases from the UK Biobank cohorts caused loss of statistical power.

*All the datasets including the SOCCS, UK Biobank and three published clinical trials were subjected to ‘healthy volunteer’ bias.

*The AJCC tumour stage for part of rectal cancer patients from SOCCS was possibly underestimated due to neoadjuvant therapy.

*Clinical and pathological variables were unavailable in the UK Biobank cohorts.

*Key prognostic factors such as somatic mutations and treatment were not collected for the SOCCS and UK Biobank cohorts.

*Analysis was based on imputed GWAS data assayed by SNP chips, thus germline variations such as chromosomal anomalies were not investigated.

*A comprehensive field synopsis to identify genetic variants reported by published candidate genetic association studies was not conducted.

* Candidate genetic variants other than the three groups of variants included in the thesis were not investigated.

* For validation analysis, only the UK Biobank cohort was used to validate variants identified from candidate association studies, whereas for GWAS, CRC-specific survival was not investigated due to data availability.

*Definitions of survival outcomes including the starting and end points were not harmonised, which led to varied observed survival estimates. However, relative effects (HRs) derived from this thesis should be generalizable.

* Due to limited sample size of included cohorts, analysis in this thesis is underpowered to detect small prognostic effects on survival outcomes, especially for variants with low minor allele frequency.

*Assignment of cause of death can be subjective, which could impact observed estimates of CRC-specific survival.

*For predictive modelling, due to limited data availability, prediction models including clinic-pathological variables were not externally validated.

6.2.2 Statistical analysis

Survival analysis

All analyses were performed using the Cox regression model which is the most widely used model to analyse time-to-event outcomes. The key assumption for this model is that the effect of a certain factor remains constant with time, namely the proportional hazard assumption. Although the Cox model has been widely adopted in various studies including clinical trials, prediction models and large-scale omics analysis such as GWAS, the proportional hazard assumption has seldom been examined in published literature(Guyot *et al*, 2011). Thus far, there have been different methods proposed to check this assumption, among which graphic examination is the most straightforward method. If the assumption holds, survival curves stratified by different values of a certain factor should be generally parallel to one another over the observation time window. Kaplan-Meier curves listed in Chapter 5 present generally parallel curves for non-genetic factors including the AJCC stage (**Figure 5-5**), tumour grade (**Figure 5-6**) and genetic variants identified from the GWAS of this thesis (**Figure 5-25** Kaplan-Meier estimates of CRC-specific survival in the SOCCS stratified by the genetic variant rs143664541, **Figure 5-26** for rs75809467 and **Figure 5-38** for rs323694), indicating no visible departure from the proportional hazard assumption. Besides graphical examination, statistical tests have also been proposed to detect potential violated proportional hazards assumption. One of the most commonly used tests is based on the Schoenfeld residual, which is defined as the difference between the covariate value for a certain individual and the weighted average of the covariate for all individuals under risk. The test examines the correlation between the scaled Schoenfeld residual and observation time (Schoenfeld, 1980). It is unrealistic to examine millions of genetic variants included in the GWAS, thus I performed the test using the SOCCS cohort for genetic variants identified in this thesis (rs143664541, rs75809467, rs323694 and rs7495132), and identified no evidence of significant violation to the proportional hazards assumption for these four variants ($p>0.05$). However, this test has been questioned due to possible spurious findings based on arbitrarily scaled time variable (Park & Hendry, 2015), and the results of the test should be interpreted with caution.

In order to improve the computational efficiency for the GWAS, I used a Martingale residual-based approach to estimate effects of eight million genetic variants on CRC survival. Technical details of this approach are described in Chapter 4 (page 104). Transformed from the standard Cox model this method, therefore, also relies on the same proportional hazard assumption. The main strength of this method is that the Cox model only needs to be fitted once including all the non-genetic covariates and the Martingale residuals generated from the fitted Cox model can be used as a new 'phenotype' and be tested using linear regression models which lead to remarkably increased computational efficiency. Reynisson et al. conducted simulation studies to compare different methods for GWAS using time-to-event data and found that the Martingale residual-based method showed slightly lower statistical power compared to standard Cox regression (Reynisson, 2018). However, they also identified a lower false discovery rate (type I error rate) for this method than standard Cox regression (Reynisson, 2018). Therefore, the Martingale residual-based approach adopted in the GWAS of this thesis is expected to provide more conservative results. In order to identify possible false negative findings from the Martingale residual-based approach, I also re-estimated the effects of variants with $p<5x10^{-6}$ using standard Cox regression.

Notably, there are other models available that do not necessarily rely on the proportional hazards assumption, such as the parametric Weibull model (Carroll, 2003). As opposed to the semiparametric Cox model, parametric methods pre-specify the survival function and are dependent on a distributional assumption, which is hard to test. Some argue that hazard ratios estimated from the Cox model may still be useful if the proportional hazard assumption is violated (Boyd *et al*, 2012), although loss of statistical power could be expected (Syed *et al*, 2016). The hazard ratios under this circumstance could to some extend be interpreted as 'time-averaged' effects. Future investigations may be considered to model the genetic effect using other methods such as the Weibull model.

With respect to the multivariable Cox model used to estimate genetic effects, I included age at diagnosis, sex and AJCC stage as covariates using the SOCCS cohort, and for UK Biobank, only age and sex were added in the model as stage was unavailable. Inclusion or exclusion of specific covariates has long been a dilemma. Mefford et al. summarised different situations of covariates and provided recommendations (Mefford & Witte, 2012). Common demographic factors such as age and sex can be deemed to be independent from autosomal genetic determinants and they are preferable to be included if they are associated with the outcome under study as increased statistical power is expected after adding them into the model (Mefford & Witte, 2012). This also applies for time-to-event outcomes according to previous simulation studies (Karrison & Kocherginsky, 2018). The situation for inclusion of the AJCC stage as a covariate is less easily perceived. As mentioned previously, AJCC stage is one of the strongest predictors of survival outcomes of CRC. Therefore, including AJCC stage in the model would increase statistical power to detect genetic effect provided that the stage is independent from the genetic variant. However, the relationships between genetic variants and stage at diagnosis remain largely unknown. It is possible that a genetic variant that determines tumour invasiveness results in more advanced stage of CRC at diagnosis, and then more advanced stage leads to worse survival outcome. In this case, the stage works as an effect mediator between the genetic variant and the outcome. Inclusion of an effect mediator could cause over-adjustment and lead to possibly diluted estimates of the genetic effects (Schisterman *et al*, 2009). I checked the genetic variants identified from this thesis (rs143664541, rs75809467, rs323694 and rs7495132) and found that none of them was significantly correlated with AJCC stage ($p<0.05$) in SOCCS. Thus, the effect estimates of these variants are relatively unlikely to have been affected by over-adjustment.

Patients who died from non-CRC related causes were considered as censored when analysing CRC-specific survival in this thesis. In fact, cases who died from non-CRC related causes are different from those who were alive until the end of observation since the occurrence of non-CRC related death precluded the patient from CRC-related death. This is also known as ‘competing risks’. Previous research has demonstrated that the Kaplan-Meier method tend to provide higher estimated event rates than the true cause-specific event rates in the presence of competing risks (Andersen *et al*, 2012; Feakins *et al*, 2018). Statistical methods, such as the Fine-Gray model (Fine & Gray, 1999) and the Lunn-McNeil model (Lunn & McNeil, 1995), have been proposed to address the competing risks by modelling the competing events separately. However, adding these models will drastically increase computational burden, and statistical tools that integrate these models in the setting of large-scale omics analysis as well as predictive modelling have not been well developed. Thus, competing risks were not modelled in the analysis. Previous evidence showed that the presence of competing risks can influence the observed survival rates, but it tends not to bias the relative effect estimates (HRs) derived from the Cox model (Feakins *et al*, 2018). In addition, the Cox model is also advantageous in terms of interpretation of the cause-specific hazard ratios which directly reflect how the variable of interest, such as genetic variants, are linked to survival rates, whilst interpretation of relative effect estimates derived from competing risk models is not straightforward (Andersen *et al*, 2012).

Statistical power and multiple testing

The statistical power for the genetic association studies was calculated based on the method provided by Owzar et al. (Owzar *et al*, 2012). Parameters which need to be specified to estimate power using this approach include: α level, sample size, proportion of events, assumed genetic model and effect size. The main strength of this method is that it does not rely on the contiguous alternative assumption which applies to most existing power estimation methods (Hsieh & Lavori, 2000). Under this assumption, the effect size should converge to zero at a constant rate. Owzar et al. argued that methods developed under this assumption may only be accurate when the expected effect size is reasonably small, and, therefore, these methods may not be suitable for designing GWAS (Owzar *et al*, 2012). Given the limited sample size of the SOCCS and UK Biobank cohorts, relatively large genetic effects could be detected after correcting for multiple testing. Thus, the method proposed by Owzar et al. was employed to estimate the power in this thesis. However, some caveats still need to be noted for the power estimation. Firstly, this method was developed based on the Cox model; hence the limitations of the Cox model which have been discussed also apply to the power estimation. Secondly, inclusion of covariates could alter the statistical power, but the covariate structure is not accounted for by this method. Thirdly, as described in Chapter 4, I conducted the GWAS using a Martingale residual based approach which has been shown to have slightly lower statistical power than the standard Cox model (Reynisson, 2018). Fourthly, the statistical power was estimated under the additive genetic model, which was used as the main analysis of this thesis, and hence the estimated power should not be applied to the sensitivity analysis under the recessive genetic model. In summary, the estimated statistical power should be only considered as an approximation rather than an accurate value.

In order to derive the statistical power after controlling for multiple testing, the α level was corrected by applying the Bonferroni correction. This is a rather stringent approach, which aims to control the familywise type I error—that is the probability of at least one test being false positive. As a trade-off, it leads to elevated type II error (false negative rate). Therefore, I also adopted another less stringent approach, namely the FDR correction, to evaluate the statistical significance of the results from the validation study of variants previously linked with CRC survival and the other two hypothesis-driven association studies. In fact, variants that remained significant after FDR correction also survived the Bonferroni correction. It is worth mentioning that in these genetic association studies, I only included independent variants by controlling the r^2. Under this circumstance, the FDR approach could be advantageous because it is less likely to exclude true associations compared to the Bonferroni correction, although both approaches produced the same results in this thesis. However, in the setting of GWAS where there exists intrinsic LD throughout the genome, the FDR approach could suffer from loss of statistical power by not accounting for the LD structure (Kaler & Purcell, 2019). Previous efforts have been made to estimate the burden of multiple testing for GWAS in the presence of LD, and the results demonstrated a million independent tests in Europeans (Pe'er *et al*, 2008). Since then a Bonferroni corrected threshold ($p< 5x10^{-8}$) has become the standard of evaluating statistical significance in GWAS. This threshold was therefore used in this thesis to report discoveries. Recently, more flexible methods of determining the significance threshold have been proposed using variant-based heritability for different traits of interest with an attempt to further reduce false negative rates (Kaler & Purcell, 2019). However, these methods have not been widely used and should be further validated.

Predictive modelling

The Cox regression model was used to obtain effect estimates for each included predictor. Strengths and limitations of the Cox model have been discussed in previous sections. Recently, other methods have gradually been applied in published literature to develop prognostic prediction models. For example, some studies showed that the Weibull model might outperform the Cox model in some aspects such as goodness-of-fit (Baghestani *et al*, 2015). Given that the main focus of this section of the thesis is to test the potential added predictive value of genetic variants, the performance of models developed using other methods such as the Weibull model was not evaluated. Once genetic variants that can robustly predict CRC survival are identified, efforts should set out to explore methods such as the parametric Weibull model or more advanced approaches like neural networks to develop models with better performance.

As for model performance, Harrell's C statistic was employed as the main metric to evaluate the discriminative performance of developed models. One advantage of the C statistic is that it has a natural interpretation—the proportion of randomly selected pairs with correct predictions, and the statistic is well defined for continuous, binary and censored outcomes. Currently it has become one of the most widely used measure of model performance. However, the interpretation of C statistic becomes less straightforward when comparing the performance of different models. The incremental value of C statistic is not additive and has no direct interpretation (McKeigue, 2019; Pencina *et al*, 2012). Moreover, it has been shown that the incremental value of C statistic is highly dependent on the performance of the baseline model (McKeigue, 2019). Although a U statistic-based test can be employedto assist inference on the incremental predictive value by examining if one model is more concordant than the other, it is still hard to interpret and quantify the magnitude of the potential gain of prediction performance after adding a new predictor. Novel metrics such as the expected information for discrimination have been proposed recently as potential alternatives of the C statistic (McKeigue, 2019). However, how these metrics can be applied in censored data remains unclear. Nonetheless, there might be limited implications in terms of the interpretation of the results of this thesis, as the C statistic remained approximately unchanged after including the genetic variants.

With respect to model calibration, the Hosmer-Lemeshow test along with visual examination of the calibration plot were used. One general caveat for these two methods is that they evaluate the model calibration at a specific time point and survival outcomes are dichotomised. I chose the 5th year since diagnosis as the cut-off time to evaluate model calibration, as the 5-year survival outcomes are generally used in clinical practice to assess prognosis of cancer patients. However, it should be noted that model calibration may vary at different observation times. The Hosmer-Lemshow test examines the hypothesis that the predicted and observed number of events are the same across all risk groups (details in Chapter 4 page 100). Although widely adopted in published literature, this test has been criticised as categorisation of risk groups is largely arbitrary and the test tends to generate unstable results (Bertolini *et al*, 2000). Moreover, this test is also sensitive to sample size. Kramer et al. conducted a simulation study and found that as the sample size reached 50,000, the Hosmer-Lemshow test would almost certainly provide significant findings ($p<0.05$) for simulated models with only slight departure from perfect fit (Kramer & Zimmerman, 2007). In this case, a significant finding from the Hosmer-Lemshow test does not necessarily mean the model is poorly calibrated. Therefore, other measures are needed to assist evaluating the model calibration. I used calibration plots as a visual presentation of the overall agreement of the observed and predicted survival rates. Recently, other metrics such as the calibration slope and the observed/expected ratio have been proposed (Crowson *et al*, 2016). These metrics should be investigated exhaustively once a compelling genetic predictor has been identified to be integrated into future models.

Missing data

Missing data are prevalent in medical research. Common reasons for this problem include, but are not limited to: participants did not report relevant information; investigators failed to collect it; or the information was simply unavailable. Regression analysis built in most statistical platforms is performed only in observations with complete information, also known as complete case analysis (CCA). Missing data mainly take three different patterns—missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Robins & Wang, 2000). The MCAR means that the probability of a missing value is independent from any other patients' characteristic and is only determined by chance. Under this circumstance, unbiased results will be obtained from CCA as the completed observations can be seen as a random sample from the original population. Missing at random (MAR) is based on the assumption that missing values are solely dependent on observed information, which makes it possible to re-construct the incomplete records with certain amount of uncertainty using methods such as multiple imputation. As for the third form, the MNAR means that missing data are dependent on unobserved information. Currently, there has been no general approaches to address MNAR. In this thesis, missing data are present primarily in the SOCCS study. With respect to genotype data of SOCCS, approximately 10% of the participants were excluded during quality control. Nonetheless, distribution of principle components of genotyped individuals is tightly clustered with the UK population **(Figure 4-1**). Therefore, it is less likely that exclusion of these individuals systematically biased the genotype distribution of the cohort. As for non-genetic variables, the proportion of missing values for these variables are: age at diagnosis (1%), date of definitive treatment (0.7%), AJCC stage (6.5%), tumour grade (14.8%) and tumour site (1.2%). Given that only a small number of variables were included, it could be unsafe to assume the missing patterns of these variables are solely dependent on observed information, and it is highly possible that other uncollected variables can impact on the missingness (MNAR). Under such circumstances, methods such as multiple imputation that only leverage observed information to impute missing values can, by contrast, introduce bias (Hughes *et al*, 2019). It should be noted that analysis in the thesis was conducted following the CCA which could also generate biased estimates if the missing data were MAR or MNAR (Altman & Bland, 2007). Future efforts are needed to collect more variables for each study cohort and extensively investigate possible mechanism behind the missing pattern, and address this issue before subsequent analysis. Another challenge lies in the prohibitive computational burden especially in the setting of large-scale omics analysis such as GWAS with millions of tests being conducted and predictive modelling where hundreds of equally-sized bootstrap samples are generated. There is a pressing need to develop analytical tools with high computational efficiency to combine methods such as multiple imputation where multiple imputed samples were generated with analyses like GWAS and bootstrapping.

Summary of strengths & limitations for statistical analysis

Strengths:

* Effect estimates derived from Cox regression models had clear interpretation.

* Covariates included in the multivariable analysis (age at diagnosis, sex and AJCC stage) increased statistical power.

*A Martingale residual-based approach, transformed from the Cox model, was used in the GWAS to estimate effects of genetic variants on CRC survival. This approach provides more conservative results with lower type I error rates and can improve computational efficiency.

*The method used to estimate statistical power does not rely on the contiguous alternative assumption, and it can potentially provide more accurate estimates for expected large effects.

Limitations:

*The proportional hazards assumption might not be valid for some genetic variants. Future efforts may consider estimating effects of genetic variants using other methods, such as the Weibull model, and comparing the findings with estimates from the Cox model. Potential time-varying genetic effects should be further explored if this assumption is violated.

*Potential competing risk effects for non-CRC related death were not modelled, leading to possible overestimated CRC-specific survival rates using the Kaplan-Meier approach.

*Results of power estimation should not be considered as accurate values given that the structure of included covariates was not considered.

*For predictive modelling, only the C statistics, the Hosmer-Lemeshow test and calibration plots were adopted to evaluate model performance. These metrics have their own limitations. The C statistics has no clear interpretation when assessing the incremental value after adding a new predictor. Spurious findings could be generated from the Hosmer-Lemeshow test due to arbitrary categorisation of risk groups, especially when the sample size is large. The calibration plot cannot provide quantitative assessment of model calibration. Thus, more novel metrics should be employed to extensively evaluate prediction models developed in this thesis.

*Possible mechanisms of missing data were not investigated due to limited data. Complete case analysis was used to deal with missing data, which could lead to potential bias.

6.3 Interpretation of main findings

6.3.1 Systematic literature review

Main findings

The systematic literature review included a total of 83 primary model development studies and 52 external validation studies that investigated the predictive value of candidate prognostic factors of CRC. The main predictors used in these published prediction models included age at diagnosis, sex, AJCC stage, TNM stage, tumour grade and biomarkers such as CEA. Somatic genetic alterations, for example *BRAF* mutation and microRNA markers, were only investigated in 3 published prediction models (Goossens-Beumer *et al*, 2015b; Manceau *et al*, 2014; Zhang *et al*, 2013). No prediction models used germline genetic markers in predicting CRC survival. This finding could be due to the limited evidence supporting associations between specific germline genetic variants and survival outcomes of CRC, which points to the main focus of this thesis.

As for the predictive performance of published models identified in this review, I found that most models showed low to modest discriminative performance (evaluated by C statistics). In addition, the majority of included models were subject to potential risk of bias. The main sources of risk of bias stemmed from loss to follow-up of the study cohort and methodological flaws in data analysis, for example the lack of internal validation when reporting the model performance. I identified eight models (Basingstoke preoperative score, Fong score, Iwatsuki score, Memorial Sloan Katherine Cancer Center nomogram, Nordinger score, Peritoneal Surface Disease Severity Score, Kanemistu nomogram and Valentini nomogram) that had been validated in at least two external datasets and conducted meta-analyses to evaluate the pooled external performance of these models. Meta-analyses found significant discriminative ability (the 95% CI of the C statistic excluding the null) for five out of eight models (Basingstoke score, Fong score, Nordinger score, Peritoneal Surface Disease Severity Score and Valentini nomogram) in predicting six survival outcomes of CRC patients (details in Chapter 2). Among these five models, the Fong score was externally validated four times. It used seven predictors (positive resection margin, extrahepatic lesion, metastases-free period, number of metastases, the largest size of metastasis, CEA and lesion of regional lymph nodes for primary tumour), and the meta-analysis found significant discriminative ability for the score to predict recurrence-free and overall survival of CRC patients with liver metastasis after curative resection. All the five models used clinic-pathological predictors that were not available in the SOCCS or the UK Biobank cohorts; therefore, I was unable to further validate these models.

6.3.2 Candidate association studies

Validation and predictive modelling of published genetic variants associated with CRC survival

Main findings

A total of 43 genetic variants previously reported to be associated with CRC survival were identified by searching the GWAS catalogue and their associations with overall and CRC-specific survival were validated in the SOCCS study. However, no significant associations between any individual variant or the polygenic risk score that combined all 43 variants and survival outcomes of CRC were observed after correcting for multiple testing. Although small effects cannot be confidently excluded due to limited sample size, these findings indicated no major effects of these previously identified variants on prognosis of CRC patients.

The results of this section suggested poor reproducibility of previous GWASs. This could be due to different characteristics of the study cohort. For example, the previous GWAS by Pander et al. focused on only stage IV patients who had received certain treatment strategies (Pander *et al*, 2015). Different study outcomes could also potentially explain these inconsistent findings. Although different outcomes such as DFS and RFS show generally good concordance with OS (Sargent *et al*, 2005), statistical power varies across these outcomes due to different number of events. It is worth mentioning that 41 of the 43 (95%) variants (except for rs209489 and rs885036) did not reach GWAS significance ($p<5x10^{-8}$) in the original report. Moreover, 39 out of the 43 variants (91%) were identified by GWASs with relatively small sample sizes (N<1,000), pointing to possible false positive findings or overestimation of the genetic effects in the published GWASs. With respect to associations not corrected for multiple testing, three genetic variants (rs17026425, rs6854845 and rs17057166) were identified to be associated with overall survival of CRC patients from the SOCCS study at p<0.05 and showed the same direction of effects compared to the original GWAS reports. These variants are discussed in the sections below.

Rs17026425

Xu et al. reported a significant association between the A allele of the variant rs17026425 and inferior overall survival of rectal cancer patients in a Canadian cohort (Xu *et al*, 2015). In addition to a similar effect on overall survival of all CRC patients from SOCCS, I also observed a suggestive association (uncorrected p<0.05) between this variant and overall survival of rectal cancer patients based on the results of a stratified analysis. Interestingly, neither the analysis in this thesis nor in the original GWAS by Xu et al. found a significant association of this variant among colon cancer patients, indicating that this effect may be more prominent in rectal cancer. This variant is an intron variant in the IQ motif containing M (*IQCM*) gene located in chromosome 6. According to the Human Protein Atlas (URL6-2), this gene is highly expressed only in testis tissue. Intriguingly, our results found a significant association between this variant and overall survival of male CRC patients from the SOCCS study. It is worth noting that this variant is located in the binding region of the JUN/JUND transcription factors which are highly expressed in human CRC tissue (Wang *et al*, 2000). The results of this section therefore merit further investigation in the potential biological function of this locus.

Rs6854845

This is an intergenic variant located in the super-enhancer—a cluster of *cis*-regulatory elements with high density of transcriptional factors—closest to the *BTC* gene which encodes the Betacellulin protein. This protein is a member of the Epidermal Growth Factors (EGF) and serves as a ligand for the EGF receptor. Nagaoka et al. found that the *BTC* gene is highly expressed in human CRC tissue (Nagaoka *et al*, 2016). Additionally, this gene showed significantly higher level of expression in wild-type *KRAS* CRC cases compared to *KRAS* mutated cases (Nagaoka *et al*, 2016), indicating potential different roles of the *BTC* gene in carcinogenesis and progression of the two types of CRC. However, currently there has been a paucity of evidence supporting a direct association between the variant rs6854845 and expression of the *BTC* gene. Cong et al. conducted an *in vitro* investigation by generating G>T mutation (T as mutated allele) for rs6854845 in colon cells using the Crispr/Cas9 technique (Cong *et al*, 2019). They observed that colon cells with this mutation showed significantly altered chromosomal structure of the super-enhancer compared with those without the mutation. In addition, significantly higher expression of several key genes near this variant, including the *CXCL*2, 3, 5, 6, 8, *EREG* and *EPGN* genes, was found in colon cells with the mutated rs6854845 than in cells of wild-type (Cong *et al*, 2019). However, no significant difference was found for the expression of the *BTC* gene (Cong *et al*, 2019). Among these genes with altered expression, previous evidence found potential associations between highly expressed *CXCL* 2 and 3 genes in CRC tumour tissue and improved overall survival of patients (Lv & Li, 2019). As for the *EREG* gene, Qu et al. reported that highly expressed *EREG* gene mediated through promoter demethylation can activate the Epidermal Growth Factor Receptor (EGFR) pathway in CRC carcinogenesis (Qu *et al*, 2016). Thus far, there has been limited evidence showing the roles of *CXCL* 5, 6, 8 genes and the *EPGN* gene in CRC carcinogenesis or progression. Our findings along with the original GWAS by Xu et al. justify further investigation in the biological role of rs6854845 in CRC development and progression.

Rs17057166

This variant was originally related to disease-free survival of rectal cancer patients in the GWAS by Xu et al (Xu *et al*, 2015). In this thesis, I observed suggestive association between this variant and overall survival for all CRC patients from the SOCCS study. However, no association was detected in the stratified analysis among rectal cancer patients. Although a possible recessive effect was found for this variant in SOCCS, I failed to validate it using the UK Biobank cohort. This variant is an intron variant located in the *LINC01847* gene. Thus far, there has been a dearth of evidence regarding the biological function of this variant or the gene.

Combined effect and predictive value

In addition to investigating individual effect of each variant, I also explored the combined predictive value of the 43 variants. Multivariable regression found that approximately half (42%) of the variants showed opposite direction of effects compared to the original GWASs, which is close to what is expected by chance (p=0.34 of a Chi^2 test). Moreover, none of the 43 variants remained in the model after applying feature selection using both the LASSO and backward selection methods. This underpins the absence of meaningful predictive value for any of the included variants. The predictive performance of the 43 variants was evaluated using both the UK Biobank cohort from which the prediction model was derived and the SOCCS study as an external validation dataset. In the UK Biobank cohort, although positive point estimates of C statistics (>0.5) were observed, the 95% confidence intervals included the null (0.5) after internal validation using bootstrapping. This means that a model with significantly positive discriminative ability cannot be trained and derived using the 43 genetic variants to predict survival outcomes of CRC. External validation was then conducted using the fitted 43-variant model in the SOCCS study. The model showed no predictive value in SOCCS given that the observed survival estimates remained approximately unchanged as the predicted survival estimates increased. Moreover, a negative point estimate of C statistic (0.499) was observed when using the model to predict CRC-specific survival in SOCCS. In addition, the 43 variants showed no added predictive value on the basis of other known prognostic factors including age at diagnosis, AJCC stage and tumour grade.

In summary, analysis of this section found that genetic variants identified by previous GWASs to be associated with CRC survival are unable to efficiently predict survival outcomes of CRC patients in two external patient cohorts. This agrees with findings from the previous replication analyses that none of the 43 genetic variants remained significantly associated with CRC survival in SOCCS. It also points to the findings from the systematic literature review that no genetic variants have been employed by published prediction models. Given possible false positive or over-estimated genetic effects identified from previous small-scale GWASs, future larger GWASs and meta-analyses combining previous GWASs are needed to identify genetic variants that are robustly associated with CRC survival and can potentially be applied to improve predicting survival outcomes of CRC.

CRC-risk variants

Main findings

In this section, I investigated potential subsequent effects of 128 common CRC-risk variants identified from previous meta-analyses of GWAS studies (Huyghe *et al*, 2019; Law *et al*, 2019), on survival outcomes using the SOCCS study. Under an additive genetic model, none of the CRC-risk variants showed a significant association with survival outcomes after correcting for multiple testing. The polygenic risk score representing the overall genetic susceptibility to CRC was not associated with survival outcomes either. No signals were detected in stratified analyses after correcting for multiple testing. These findings indicate that currently known CRC-risk variants as a group have limited influence on subsequent survival outcomes after diagnosis. The heritable components of the observed variation of survival outcomes may have distinct genetic architecture that warrants separate GWASs to identify survival-related genetic loci. However, individual small effects of each CRC-risk variant cannot be excluded due to the limited sample size. As mentioned in Chapter 4 (page 88), the additive genetic model has limited statistical power to detect potential recessive genetic effects (Pereira *et al*, 2009). Sensitivity analysis using a recessive genetic model was conducted for the 128 genetic variants. Two variants (rs7495132 and rs10161980) remained significantly associated with CRC survival in SOCCS after correcting for multiple testing and their recessive effects were also observed in UK

Biobank. Given that these two variants failed to reach statistical significance in the additive analysis, they are more likely to follow a recessive mode of inheritance.

Rs7495132

I found that patients from the SOCCS study with the TT genotype of rs7495132 had significantly worse CRC-specific survival. However, no significant recessive effect was found for this variant on overall survival in either SOCCS or UK Biobank. The T allele of this variant was identified as a CRC-risk increasing allele in the original GWAS meta-analysis. Located in chromosome 15, rs7495132 is an intron variant of the *CRTC3* gene which encodes the protein CREB regulated transcription co-activator 3 (CRTC3). Previous evidence shows that the CRTC3 protein can regulate energy balance and is associated with weight gain in a mouse model (Song *et al*, 2010). According to the results of Genotype-Tissue Expression (GTEx) project (URL6-3), the T allele of rs7495132 is associated with lower expression of the *CRTC3* gene in human musculoskeletal tissue and subcutaneous adipose tissue. However, no significant association is present in colonic tissue based on the GTEx results. Thus far, the impact of adiposity on prognosis of CRC patients is still controversial (Silva *et al*, 2019). Some studies reported that the adiposity is a potential protective factor for CRC survival (Asghari-Jafarabadi *et al*, 2009), whereas other studies support a detrimental prognostic effect of adiposity (Haydon *et al*, 2006). Various factors such as different treatment strategies, varied definition of adiposity and methodological biases could be behind these inconsistent findings. There is a pressing need for future research to further reveal the biological function of rs7495132 in the adiposity-mediated pathway involved in CRC progression.

Rs10161980

In addition to rs7495132, I identified another variant, rs10161980, associated with overall survival in SOCCS and CRC-specific survival in UK Biobank. This variant, located in chromosome 13, is an intron variant of the *AL139383.1* gene. There has been a paucity of evidence revealing associations between this variant and expression of any genes. The biological function of the *AL139383.1* gene also remains to be understood.

Comparison to previous studies

There have been two studies with similar study design that investigated prognostic effect of smaller numbers of CRC-risk variants (Abuli *et al*, 2013; Smith *et al*, 2015). Smith et al. included 2,083 CRC patients as a discovery dataset and 5,552 patients as a validation dataset (Smith *et al*, 2015). They identified only one variant rs9929218 in chromosome 16 that was significantly associated with CRC survival in both cohorts. This variant is an intron variant that lies in the *CDH1* gene which encodes the protein E-cadherin. E-cadherin is a widely known tumour suppressor that plays an essential role in cell-cell adhesion. Evidence has shown that loss of function of E-cadherin can cause tumour progression and metastasis (Takeichi, 1991). Rs9929218 is in strong LD with rs16260 that affects *CDH1* expression (Li *et al*, 2000). Therefore, rs9929218 may influence CRC progression by mediating the *CDH1* expression. In this thesis, however, a nominally significant (p<0.05) yet opposite effect of this variant was observed on overall survival of patients in SOCCS under the additive genetic model. Smith et al. reported that patients with AA genotype had significantly inferior overall survival, but in this thesis, I observed a potential protective effect of the A allele on overall survival of patients from SOCCS. No significant recessive genetic effect was found on either overall or CRC-specific survival in my analysis. A few possible explanations could be behind these inconsistent findings. Firstly, Smith et al. used a discovery cohort of stage IV CRC patients. Colorectal cancer with stronger metastatic potential may have different genetic determinants compared to CRCs diagnosed at early stages. Notably, in stratified analysis of 784 stage IV CRC patients from SOCCS, I failed to observe a significant association between this variant and CRC survival either. Secondly, Smith et al. found that rs9929218 was significantly correlated with treatment response of chemotherapy and therefore they adjusted for treatment strategy in their survival analysis. However, this thesis is limited by treatment data being unavailable and therefore this effect could not be further explored using our study cohorts. Investigation of this variant using well-documented large cohorts should still be considered. With respect to the other study by Abuli et al. (Abuli *et al*, 2013), a total of 16 CRC-risk variants were investigated in 1,235 CRC patients. However, no significant associations were found after correcting for multiple testing.

Variants previously linked with survival outcomes of other cancers

Main findings

The second hypothesis is that genetic variants associated with survival outcomes of other cancers may also affect prognosis of CRC. I included 82 variants reported by previous GWASs and tested their associations with overall and CRC-specific survival of patients from the SOCCS study. Overall, none of the variants remained statistically significant after correcting for multiple testing, although four variants (rs1728400, rs17693104, rs6797464 and rs823920) were found to be associated with CRC survival at nominal significance (uncorrected p<0.05). These findings indicated that major effects of these genetic variants on CRC survival are unlikely. Different cancer types may have distinct genetic determinants in terms of survival outcomes after diagnosis.

Rs17693104

This is an intron variant of the *SHD4B* gene in chromosome 10, was initially reported to be associated with overall survival of serous epithelial ovarian cancer. In this thesis, I found a concordant effect of the T allele on inferior overall and CRC-specific survival of CRC patients in SOCCS. Notably, earlier evidence has suggested that this variant was associated with sensitivity of Capecitabin--an agent that has been regularly administrated to CRC patients (O'Donnell *et al*, 2012). Therefore, our results indicate that rs17693104 could potentially influence CRC survival by modifying treatment response of patients who have been taking Capecitabin. The possible metabolic pathways and mechanisms should be further explored in the future. According to the results of GTEx, the T allele of rs17693104 is associated with higher expression of the *RP11-137H2.4* gene in multiple types of human tissue. This gene encodes the long non-coding RNA (LncRNA) RP11-137H2.4. Ouimet et al. conducted *in vitro* investigations to characterise biological function of the RP11-137H2.4 using human acute lymphoblastic leukaemia (ALL) cell lines (Ouimet *et al*, 2017). They found that silencing RP11-137H2.4 led to significantly increased cell apoptosis (Ouimet *et al*, 2017). Thus far, there has been limited evidence concerning the function of RP11-137H2.4 in CRC carcinogenesis or progression. These findings provided perspectives for future research to explore the possible role of rs17693104 in CRC progression by regulating the level of LncRNA RP11-137H2.4.

Rs6797464

Koster et al. identified an intron variant rs6797464 in the *MECOM* gene, which minor allele (A) was associated with worse overall survival of osteosarcoma (Koster *et al*, 2018). A concordant effect of the A allele of rs6797464 was found on CRC-specific survival in SOCCS, although this effect did not survive correction for multiple testing. The *MECOM* gene encodes the MDS1 (Myelodysplastic Syndrome 1) and EVI1 (Ecotropic Viral Integration Site 1) complex locus protein which is widely accepted as an oncoprotein. It was initially linked with the pathogenesis of leukaemia. Evidence has shown that this protein is involved in aberrant cell development in the bone marrow (Morishita *et al*, 1988; White *et al*, 2013). Recently, higher expression of the MDS1 and EVI1 complex locus protein has been identified in colon cancer (Shackelford *et al*, 2006). Studies also found that this oncoprotein could increase the resistance of tumour cells to treatment agents such as taxol (Liu *et al*, 2006). Makondi et al. conducted gene enrichment analysis, and identified the highly expressed *MECOM* gene as a potential biomarker of irinotecan resistance (Makondi *et al*, 2017). They also found that the expression of *MECOM* gene was associated with disease-free survival of CRC patients (Makondi *et al*, 2017). However, there has been a dearth of evidence supporting the direct link between the germline variant rs6797464 and the *MECOM* gene expression, which merits further exploration.

Rs1728400 and rs823920

With respect to the other two variants with suggestive associations with CRC survival, rs1728400 was initially identified to be linked with overall survival of breast cancer patients (Rafiq *et al*, 2014). In this thesis, I found a protective effect of the minor allele A on overall survival of CRC patients in SOCCS. However, the effect allele was not reported in the original GWAS by Rafiq et al. This variant, located in chromosome 16, is an intergenic variant near the *LINC00917* and the *AC092327.1* genes, whose biological function has not been well characterised. Rs823920 was initially reported to be associated with overall survival of pancreatic cancer patients (Tang *et al*, 2017). I observed a concordant detrimental effect of the minor allele G on CRC-specific survival. This variant is an intergenic variant close to the *ARL2BPP7* and the *MTND3P4* gene. There has been a lack of evidence concerning the biological function of this variant too. Given the fact that these suggestive associations did not survive FDR correction and could be chance findings, further replication is needed.

6.3.3 Genome-wide association study

Main findings

In this section, a total of 8,328,632 autosomal genetic variants were investigated to identify potential novel genetic determinants on survival outcomes of CRC. As shown in the QQ plots, the overall genetic effect of the whole genome on CRC survival appears to be small. This may be attributed to possible low heritability of the survival outcome as a trait that is influenced by a large number of environmental factors. Low statistical power could also be behind the observed lack of genetic signals. Given the large number of imputed variants and the relatively small sample size (N=5,675), the variant-based heritability was not estimated as recommended by previous studies (Yang *et al*, 2015). Hence, there is compelling rationale for future efforts to combining multiple GWAS datasets on CRC survival by meta-analysis to reach sufficient statistical power.

In the analysis combining CRC patients of all stages, I identified one variant rs143664541 in chromosome 6 associated with overall survival of CRC patients in SOCCS at the threshold of GWAS significance ($p<5x10^{-8}$). Although effects of concordant direction were found in external validation datasets including the UK Biobank and other three published clinical trials, no significant association was observed in a meta-analysis. As for CRC-specific survival, two variants (rs143664541 and rs75809467) were identified at GWAS significance. Rs75809467 was not associated with overall survival in external datasets. Due to limited data availability, external validation was only conducted for overall survival. I also performed stratified GWASs in locally advanced (stage II/III) and metastatic (stage IV) CRC patients separately. For stage II/III patients, I observed a variant rs323694 in chromosome 5 that was significantly associated with CRC-specific survival ($p<5x10^{-8}$). No significant signals were detected in the analysis within stage IV patients. In terms of the gene based analysis, I observed significant enrichment of signals in the *CCDC135* gene associated with CRC-specific survival after correcting for multiple testing. Genes involved in the biosynthetic process of galactolipids and up-regulating the differentiation of adipocytes were found to harbour enriched genetic signals associated with CRC survival.

Rs143664541

This is an intergenic variant located near the *FRK* gene. According to the National

Center for Biotechnology Information (NCBI) gene database (URL6-4), the *FRK* gene encodes the Fyn-related kinase which may suppress cell growth by intervening in the G1 and S phase of the cell cycle (URL6-4). Currently, there is no evidence that shows that rs143664541 can influence the *FRK* gene expression. The exact biological implication of this variant remains unclear. According to the results from the PhenoScanner (Staley *et al*, 2016) (URL6-5) where genotype-phenotype associations from the UK Biobank full cohort are indexed, the minor allele (A) of this variant is significantly correlated with higher risk of death from oesophagogastric diseases ($p=9.4x10^{-6}$). By searching the *FRK* gene using the PhenoScanncer, I identified 360 associations ($p<5x10^{-8}$) between genetic variants within the gene and different traits among which 80 (22%) traits were deaths due to different causes. However, I failed to observe significant enrichment of genetic signals in this gene on CRC survival outcomes using the gene based analysis. Future research is expected to uncover biological function of this gene. In respect to the variant rs143664541, although statistically significant associations were found between the variant and overall and CRC-specific survival in SOCCS, these associations failed to replicate in external datasets. However, a concordant direction of effects was found across multiple external datasets. These findings indicate possible over-estimation of the genetic effect of this variant in SOCCS. It is worth mentioning that this variant is imputed in SOCCS instead of directly genotyped. Moreover, this variant has a low minor allele frequency in SOCCS (1.4%) and even lower in general European population (0.3% in the 1000 Genome Project). Although an info score greater than 0.80 across all the datasets for this variant suggests acceptable imputation accuracy, imputation error could still bias the observed genetic effect. As shown in the locus zoom plot (Figure 5.25), there is poor LD structure near this variant which restrains selection of any proxies of this variant. Future efforts are needed to re-estimate the effect of this variant on CRC survival based on genotyped or sequence data.

Rs75809467

As to the other variant I identified from combined analysis of all CRC patients, rs75809467 is located in the non-coding transcript exon of the *BTF3P4* gene which is a pseudogene with unclear biological function. This variant was significantly associated with CRC-specific survival in SOCCS. Although validation analysis on the same outcome was not performed due to data availability, an opposite direction of effect was observed from the UK Biobank and the three clinical trials overall survival (see Figure 5.25). Given the general concordance of direction of effects for these two outcomes, it is likely that the genetic effect of the minor allele (T) on CRC-specific survival in SOCCS is not real. Future well-designed studies with large sample sizes are still expected to investigate this variant.

Rs323694

With respect to stratified GWAS in stage II/III CRC patients, the variant rs323694 was identified to be significantly associated with CRC-specific survival ($p<5x10^{-8}$). Given that the stage data have not been released in UK Biobank and cause of death for each patient was unavailable in the three clinical trials, I was unable to conduct validation analysis for this variant. There is a pressing need for the observed effect of this variant to be validated in independent cohorts of stage II/III CRC patients before it can be used to assist predicting survival outcomes. Rs323694 is an intergenic variant near the *IRX2* and *LOC100506858* gene. Currently, the biological implication of this variant has been poorly understood. It is worth noting that no significant association was found between this variant and CRC survival among stage IV patients. I conducted stratified GWASs in stage II/III and stage IV patients separately based on the assumption that the genetic background could be different for CRC patients with varied metastatic potential. A recent analysis based on exome-sequencing data of metastatic CRC patients showed that most CRCs metastasise before the primary tumour is clinically detectable (Hu *et al*, 2019). This finding provides further evidence that CRCs with high metastatic potential could be a distinct disease subtype compared with ones with lower metastatic potential. Therefore, GWASs stratified by stage should be considered in future investigations.

Gene based analysis

By conducting gene based analysis, I identified significant enrichment of genetic signals in the *CCDC135* gene associated with CRC-specific survival. This gene is located in chromosome 16 and encodes the Coiled-coil domain-containing protein 135 (CCDC135). According to the gene database provided by NCBI (URL6-4), the CCDC135 protein can regulate germ cells differentiation during spermatogenesis. Thus far there has been no evidence showing any roles of CCDC135 in CRC carcinogenesis and progression. Based on the data from the Human Protein Atlas (URL6-6), the *CCDC135* gene is highly expressed only in testis. The Human Protein Atlas also incorporates the data from The Cancer Genome Atlas (TCGA), and analysis using different cancer tissues found that the *CCDC135* expression is enriched in endometrial cancer. Survival analysis suggested that higher expression of *CCDC135* in endometrial cancer tissue was associated with significantly favourable survival outcome. As for CRC, the *CCDC135* gene was highly expressed in 143/454 tumour samples, but higher expression of this gene had no significant effect on overall survival of CRC patients from the TCGA project (URL6-6). Future research is needed to investigate whether the *CCDC135* gene can serve as a prognostic indicator for CRC patients.

Gene-set based analysis

With regard to the results of gene-set based analysis, significant enrichment of genetic signals was found in two sets of genes. The first set of genes are involved in the biosynthetic process of galactolipids. Galactolipids are a subtype of glycolipid with galactose as the sugar group. Hou et al. revealed that galactolipids can suppress inflammatory mediators and serve as a potential anti-cancer agent for melanoma using a mouse model (Hou *et al*, 2007). Another study conducted by Yang et al. also leveraged mice with implanted melanoma (Hou *et al*, 2007), and they found that plant galactolipid can suppress lung metastasis of melanoma by decreasing tumour necrosis factor (TNF) α mediated pulmonary vascular permeability (Hou *et al*, 2007). Currently there has been limited research evidence supporting possible anti-cancer roles of galactolipids in CRC. The second gene-set identified, is involved in up-regulating the differentiation of adipocytes. There has been an abundance of evidence supporting the role of adipocytes in carcinogenesis and metastasis of multiple human malignancies. Firstly, studies have shown that dysfunctional adipocytes can directly down-regulate the inflammatory-immune-angiogenic response system and hence promote cancer cell proliferation and metastasis (Nieman *et al*, 2013). Secondly, adipocytes can secrete factors such as TNF α and interleukin-6 that participate in key pathways in CRC carcinogenesis and progression (Hodge *et al*, 2005; Pikarsky *et al*, 2004). Thirdly, adipocytes can regulate the tumour microenvironment of CRC and subsequently contribute to CRC invasion (Tabuso *et al*, 2017). Although the molecular mechanisms of adipocytes affecting CRC progression have been underpinned by cumulated evidence, epidemiological studies have yielded inconsistent results regarding the prognostic effect of adiposity on CRC survival, as described in the previous section discussing CRC-risk variants. Future research is still needed to clearly explain the role of adiposity in CRC progression. It should be noted that findings of enriched genetic signals in genes and gene-sets in this thesis have not been verified in external datasets. Therefore, validation studies are needed to confirm these results before further characterisation of detailed molecular mechanisms.

6.4 Conclusions and recommendations

This thesis sets out to explore impact of germline genetic variations on survival outcomes of CRC patients using multiple patient cohorts including SOCCS, UK Biobank and three published clinical trials. Overall, the inheritable genetic components, represented by germline genetic variants throughout the genome, may contribute to only a relatively small part of variance in survival outcomes of CRC patients. However, several genetic variants were identified to be possibly associated with CRC survival and merit future investigation. Outlined below are main conclusions and recommendations derived from each part of the thesis.

6.4.1 Systematic literature review

In the first part of the thesis, I systematically reviewed published prediction models on survival outcomes of CRC including 83 model development studies and 52 external validation studies. None of the reviewed models included germline genetic variants as predictors. Somatic mutations such as the *KRAS* and *BRAF* mutations were used by a few models. Most published prediction models have not been validated in external datasets and are subject to potential sources of bias which mainly include cohort attrition and methodological flaws. There have been eight models (Basingstoke preoperative score, Fong score, Iwatsuki score, Memorial Sloan Katherine Cancer Center nomogram, Nordinger score, Peritoneal Surface Disease Severity Score, Kanemistu nomogram and Valentini nomogram) that had been validated in multiple external datasets. Meta-analyses of the model performance metrics (C statistics) showed low to modest discriminative performance.

In order to further improve prediction, future investigations should explore the added predictive value of germline genetic variants that are associated with CRC survival such as genetic variants identified from this thesis and other studies once the reported associations have been confirmed by independent studies. Other novel predictors such as microRNAs should also be integrated. Besides developing new models, published prediction models should be validated by more independent efforts. The real-world impact and cost-effectiveness of published models are also expected to be studied before they can be routinely applied in clinical practice. Careful consideration of factors used to predict survival outcomes of CRC is necessary and this systematic review can guide the selection of prognostic predictors in the future.

6.4.2 Candidate association studies

Validation of genetic variants previously reported to be associated with CRC survival

In this section, I investigated 43 genetic variants identified by previous GWASs that were associated with survival outcomes of CRC. Overall, no genetic variants were found to be significantly associated with either overall or CRC-specific survival after correction for multiple testing. As for the combined effect of the 43 variants, I did not observed significant association between a polygenic risk score and CRC survival of patients from the SOCCS cohort. In addition, the 43 variants combined showed no predictive value either used alone or together with other known non-genetic factors. Our results suggested poor reproducibility of previously identified variants. Previous findings of genetic effects may be false positive. Given the sample size of the SOCCS cohort, I concluded that previously identified variants associated with CRC have no major effect on survival outcomes of CRC patients, although small effects for each individual variant cannot be confidently excluded.

Although none of the 43 variants survived multiple testing, three genetic variants (rs17026425, rs6854845 and rs17057166) were identified with concordant direction of effects compared to the original GWASs at a nominal significance level ($p<0.05$). Validation studies of large sample size may be considered in the future. Large GWASs with sufficient statistical power are needed to identify genetic variants robustly associated with CRC survival.

I also tested another two groups of genetic variants in this section based on two distinct hypotheses. Firstly, I examined associations between 128 CRC-risk variants and survival outcomes in SOCCS. Using an additive genetic model, no genetic variants were significantly associated with CRC survival after correcting for multiple testing. Similarly, the CRC-risk polygenic score was not associated with survival outcomes in SOCCS either. However, potential recessive effects were observed for two CRC-risk variants (rs7495132 in the *CRTC3* gene and rs10161980 in the

AL139383.1 gene) using the SOCCS study and significant recessive effects were also detected in the UK Biobank. The second hypothesis was to test whether genetic variants associated with survival outcomes of other cancers may also affect prognosis of CRC. However, no significant signals were found among the 82 included variants and survival outcomes of CRC in SOCCS.

Future studies should focus on further validating the potential recessive effects of the two CRC-risk variants, especially in other populations. Given the fact that little is known regarding the biological function of these two variants, our results also merit characterisation of possible molecular mechanisms and biological pathways in which these two variants are involved. As for genetic variants associated survival outcomes of other cancers, our findings indicated limited effects of them on CRC survival, and therefore pointed to the presence of possible distinct genetic architectures of survival outcomes across different cancers which should be investigated by separate GWASs in the future.

6.4.3 Genome-wide association study

In this section, more than eight million autosomal genetic variants were scanned to identify novel genetic loci associated with survival outcomes of CRC. The overall distribution of statistical significance revealed possible low heritability for survival outcomes of CRC. However, lack of statistical power could also explain this finding. Using the SOCCS cohort as the discovery set, two variants (rs143664541 near the *FRK* gene and rs75809467 in the *BTF3P4* gene) were found to be significantly associated with survival outcomes of CRC ($p<5x10^{-8}$). Meta-analysis combining effect estimates from the UK Biobank cohort and three clinical trials was used to validate these two findings. However, no significant associations were detected for either of the two variants ($p<0.05$), although concordant effect estimates of rs143664541 were found across the validation datasets compared to findings in SOCCS. By conducting stratified GWAS in locally advanced CRC patients (stage II/III), another variant (rs323694 near the *IRX2* gene) was observed to be significantly associated with survival outcomes of CRC ($p<5x10^{-8}$). In terms of the results of gene based analysis, I observed significant enrichment of genetic signals in the *CCDC135* gene, and for gene-set based analysis, two sets of genes involved in biosynthetic process of galactolipids and up-regulating the differentiation of adipocytes respectively were found to harbour significant enrichment of genetic signals in relation to survival outcomes of CRC patients.

Considering that CRC-specific survival was not investigated in the validation datasets and that heterogeneous study designs and variable structures were adopted by these datasets, findings from the GWAS using the SOCCS cohort should be further validated by large well-documented cohorts in the future. These genetic variants, genes and gene sets, once confirmed by future validation, should then be explored in terms of their biological implications in CRC progression and metastasis. The results of this section also call for collaborative efforts of aggregating large study cohorts and performing meta-analysis combining multiple cohorts to obtain sufficient statistical power for more discoveries of genetic variants that can potentially affect survival outcomes of CRC patients.

6.5 Implications for clinical practice and future policy

Thus far, genetic determinants of long-term survival outcomes of CRC patients remain poorly understood. The systematic review in this thesis identified a dearth of prediction models that used germline genetic markers as predictors. Published prediction models mainly adopted well-established clinic-pathological factors associated with CRC survival, mainly including tumour stage and histological features. In the meta-analysis the eight models that had been validated in at least two external datasets were found to have significantly positive discriminative performance. These models could potentially assist in selecting CRC patients with expected worse prognosis and informing possibly more intensive treatment strategy. However, these identified models should be further investigated in terms of their real-word impact and cost-effectiveness by conducting model impact studies and health-economic modelling. Current clinical guidelines have not recommended any of these prediction models to be routinely applied, and our systematic review supports no change to current recommendations.

Currently, there have been no germline genetic markers listed as prognostic factors for CRC in clinical guidelines or recommendations of official organisations such as the Canadian Cancer Society. This is due to lack of robust evidence supporting associations between any genetic variants and CRC survival. In candidate genetic association studies of this thesis, I identified two variants (rs7495132 and rs10161980) with possible recessive genetic effects on CRC survival, but further validation on these associations and exploration in their potential predictive value are needed before any recommendations can be made. Genome-wide association analysis also revealed suggestive evidence on two variants (rs143664541 and rs323694) which merit further validation in other datasets. Findings in this thesis provide possible candidates for future investigation in terms of both biological function and clinical utility. However, it is still premature to integrate any of these germline genetic variants into management of CRC patients.

Chapter 7 References

Ab Mutalib NS, Md Yusof NF, Abdul SN, Jamal R (2017) Pharmacogenomics DNA Biomarkers in Colorectal Cancer: Current Update. *Front Pharmacol* 8**:** 736

Abuli A, Lozano JJ, Rodriguez-Soler M, Jover R, Bessa X, Munoz J, et al. (2013) Genetic susceptibility variants associated with colorectal cancer prognosis. *Carcinogenesis* 34(10)**:** 2286-91

Al-Sohaily S, Biankin A, Leong R, Kohonen-Corish M, Warusavitarne J (2012) Molecular pathways in colorectal cancer. *Journal of gastroenterology and hepatology* 27(9)**:** 1423-31

Al-Tassan NA, Whiffin N, Hosking FJ, Palles C, Farrington SM, Dobbins SE, et al. (2015) A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports* 5**:** 10442

Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Niksic M, et al. (2018) Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* 391(10125)**:** 1023-1075

Altman DG, Bland JM (2007) Missing data. *Bmj* 334(7590)**:** 424

Amstutz U, Froehlich TK, Largiader CR (2011) Dihydropyrimidine dehydrogenase gene as a major predictor of severe 5-fluorouracil toxicity. *Pharmacogenomics* 12(9)**:** 1321-36

Amundadottir LT, Thorvaldsson S, Gudbjartsson DF, Sulem P, Kristjansson K, Arnason S, et al. (2004) Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS medicine* 1(3)**:** e65

Andersen PK, Geskus RB, de Witte T, Putter H (2012) Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* 41(3)**:** 861-70

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT (2010) Data quality control in genetic case-control association studies. *Nat Protoc* 5(9)**:** 1564-73

Andreyev HJ, Norman AR, Cunningham D, Oates J, Dix BR, Iacopetta BJ, et al. (2001) Kirsten ras mutations in patients with colorectal cancer: the 'RASCAL II' study. *British journal of cancer* 85(5)**:** 692-6

Araghi M, Soerjomataram I, Bardot A, Ferlay J, Cabasag CJ, Morrison DS, et al. (2019) Changes in colorectal cancer incidence in seven high-income countries: a population-based study. *Lancet Gastroenterol Hepatol* 4(7)**:** 511-518

Ardekani GS, Jafarnejad SM, Tan L, Saeedi A, Li G (2012) The Prognostic Value of BRAF Mutation in Colorectal Cancer and Melanoma: A Systematic Review and Meta-Analysis. *PloS one* 7(10)

Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F (2017) Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66(4)**:** 683-691

Arostegui I, Gonzalez N, Fernandez-de-Larrea N, Lazaro-Aramburu S, Bare M, Redondo M, et al. (2018) 185 Combining statistical techniques to predict postsurgical risk of 1-year mortality for patients with colon cancer. *Clin Epidemiol* 10**:** 235-251

Asghari-Jafarabadi M, Hajizadeh E, Kazemnejad A, Fatemi SR (2009) Site-specific evaluation of prognostic factors on survival in Iranian colorectal cancer patients: a competing risks survival analysis. *Asian Pacific journal of cancer prevention : APJCP* 10(5)**:** 815-21

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23(10)**:** 1294-6

Aune D, Lau R, Chan DS, Vieira R, Greenwood DC, Kampman E, et al. (2012) Dairy products and colorectal cancer risk: a systematic review and meta-analysis of cohort studies. *Annals of oncology : official journal of the European Society for Medical Oncology* 23(1)**:** 37-45

Austin PC, Pencinca MJ, Steyerberg EW (2017) 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. *Statistical methods in medical research* 26(3)**:** 1053-1077

Azad AK, Bairati I, Qiu X, Girgis H, Cheng L, Waggott D, et al. (2016) A genome-wide association study of non-HPV-related head and neck squamous cell carcinoma identifies prognostic genetic sequence variants in the MAP-kinase and hormone pathways. *Cancer epidemiology* 42**:** 173-80

Baghestani AR, Gohari MR, Orooji A, Pourhoseingholi MA, Zali MR (2015) Evaluation of parametric models by the prediction error in colorectal cancer survival analysis. *Gastroenterology and hepatology from bed to bench* 8(3)**:** 183-7

Barnetson RA, Tenesa A, Farrington SM, Nicholl ID, Cetnarskyj R, Porteous ME, et al. (2006) Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *N Engl J Med* 354(26)**:** 2751-63

Barras D (2015) BRAF Mutation in Colorectal Cancer: An Update. *Biomark Cancer* 7(Suppl 1)**:** 9-12

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1)**:** 289-300

Benson AB, 3rd, Schrag D, Somerfield MR, Cohen AM, Figueredo AT, Flynn PJ, et al. (2004) American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 22(16)**:** 3408-19

Bernstein TE, Endreseth BH, Romundstad P, Wibe A, Norwegian Colorectal Cancer G (2009) Circumferential resection margin as a prognostic factor in rectal cancer. *The British journal of surgery* 96(11)**:** 1348-57

Bertario L, Russo A, Sala P, Eboli M, Radice P, Presciuttini S, et al. (1999) Survival of patients with hereditary colorectal cancer: comparison of HNPCC and colorectal cancer in FAP patients with sporadic colorectal cancer. *International journal of cancer* 80(2)**:** 183-7

Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G (2000) One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of epidemiology and biostatistics* 5(4)**:** 251-3

Blanche P, Dartigues JF, Jacqmin-Gadda H (2013) Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biom J* 55(5)**:** 687-704

Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P (2008) Smoking and colorectal cancer: a meta-analysis. *JAMA* 300(23)**:** 2765-78

Bowles TL, Hu CY, You NY, Skibber JM, Rodriguez-Bigas MA, Chang GJ (2013) 10 An individualized conditional survival calculator for patients with rectal cancer. *Dis Colon Rectum* 56(5)**:** 551-9

Boyd AP, Kittelson JM, Gillen DL (2012) Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. *Stat Med* 31(28)**:** 3504-15

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 68(6)**:** 394-424

Breiman L, Friedman J, Olshen R, Stone C Classification and Regression Trees. 1984 Monterey, CA Wadsworth & Brooks: Cole Advanced Books & Software

Brush J, Boyd K, Chappell F, Crawford F, Dozier M, Fenwick E, et al. (2011) 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)**:** 1-192, iii-iv

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726)**:** 203-209

Carroll KJ (2003) On the use and utility of the Weibull model in the analysis of survival data. *Controlled clinical trials* 24(6)**:** 682-701

Chao A, Thun MJ, Connell CJ, McCullough ML, Jacobs EJ, Flanders WD, et al. (2005) Meat consumption and risk of colorectal cancer. *JAMA* 293(2)**:** 172-82

Chen HS, Sheen-Chen SM (2000) Obstruction and perforation in colorectal adenocarcinoma: an analysis of prognosis and current trends. *Surgery* 127(4)**:** 370-6

Chou WC, Wu MH, Chang PH, Hsu HC, Chang GJ, Huang WK, et al. (2018) A Prognostic Model Based on Circulating Tumour Cells is Useful for Identifying the Poorest Survival Outcome in Patients with Metastatic Colorectal Cancer. *International journal of biological sciences* 14(2)**:** 137-146

Chung S, Low SK, Zembutsu H, Takahashi A, Kubo M, Sasa M, et al. (2013) A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients. *Breast cancer research : BCR* 15(5)**:** R81

Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. (2014) External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology* 14**:** 40

Compton C (2019) Pathology and prognostic determinants of colorectal cancer. *UpToDate* Accessed on Oct,2019

Cong Z, Li Q, Yang Y, Guo X, Cui L, You T (2019) The SNP of rs6854845 suppresses transcription via the DNA looping structure alteration of super-enhancer in colon cells. *Biochemical and biophysical research communications* 514(3)**:** 734-741

Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2)**:** 187-202

Crowson CS, Atkinson EJ, Therneau TM (2016) Assessing calibration of prognostic risk scores. *Statistical methods in medical research* 25(4)**:** 1692-706

de Leeuw CA, Mooij JM, Heskes T, Posthuma D (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 11(4)**:** e1004219

De Sousa EMF, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, et al. (2013) Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med* 19(5)**:** 614-8

Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. (2017) A guide to systematic review and meta-analysis of prediction model performance. *Bmj* 356**:** i6460

Delaneau O, Marchini J, Zagury JF (2011) A linear complexity phasing method for thousands of genomes. *Nature methods* 9(2)**:** 179-81

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. *Controlled clinical trials* 7(3)**:** 177-88

Des Guetz G, Uzzan B, Nicolas P, Cucherat M, Morere JF, Benamouzig R, et al. (2006) Microvessel density and VEGF expression are prognostic factors in colorectal cancer. Meta-analysis of the literature. *British journal of cancer* 94(12)**:** 1823-32

Dettori JR (2011) Loss to follow-up. *Evidence-based spine-care journal* 2(1)**:** 7-10

Diouf M, Chibaudel B, Filleron T, Tournigand C, Hug de Larauze M, Garcia-Larnicol ML, et al. (2014) 25 Could baseline health-related quality of life (QoL) predict overall survival in metastatic colorectal cancer? The results of the GERCOR OPTIMOX 1 study. *Health & Quality of Life Outcomes* 12**:** 69

Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, et al. (2012)

Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nature genetics* 44(7)**:** 770-6

Eberle MA, Ng PC, Kuhn K, Zhou L, Peiffer DA, Galver L, et al. (2007) Power to detect risk alleles using genome-wide tag SNP panels. *PLoS genetics* 3(10)**:** 1827-37

Edge S, Byrd D, Compton C, Fritz A, Greene F, Trotti A (2010) AJCC cancer staging manual 7th edition. *NY: Springer*

Ellis LM, Takahashi Y, Liu W, Shaheen RM (2000) Vascular endothelial growth factor in human colon cancer: biology and therapeutic implications. *The oncologist* 5 Suppl 1**:** 11-5

Feakins BG, McFadden EC, Farmer AJ, Stevens RJ (2018) Standard and competing risk analysis of the effect of albuminuria on cardiovascular and cancer mortality in patients with type 2 diabetes mellitus. *Diagnostic and prognostic research* 2**:** 13

Fearon ER (2011) Molecular genetics of colorectal cancer. *Annu Rev Pathol* 6**:** 479-507

Fedirko V, Tramacere I, Bagnardi V, Rota M, Scotti L, Islami F, et al. (2011) Alcohol drinking and colorectal cancer risk: an overall and dose-response meta-analysis of published studies. *Annals of oncology : official journal of the European Society for Medical Oncology* 22(9)**:** 1958-72

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer* 136(5)**:** E359-86

Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* 94(446)**:** 496-509

Fodde R (2002) The APC gene in colorectal cancer. *European journal of cancer* 38(7)**:** 867-71

Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. (2017) Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 186(9)**:** 1026-1034

Gao S, Zhao ZY, Wu R, Zhang Y, Zhang ZY (2018) Prognostic value of microRNAs in colorectal cancer: a meta-analysis. *Cancer management and research* 10**:** 907-929

Garlan F, Laurent-Puig P, Sefrioui D, Siauve N, Didelot A, Sarafan-Vasseur N, et al. (2017) Early Evaluation of Circulating Tumor DNA as Marker of Therapeutic Efficacy in Metastatic Colorectal Cancer Patients (PLACOL Study). *Clinical cancer research : an official journal of the American Association for Cancer Research* 23(18)**:** 5416-5425

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. (2015) A global reference for human genetic variation. *Nature* 526(7571)**:** 68-74

Ghazi S, Berg E, Lindblom A, Lindforss U, Low-Risk Colorectal Cancer Study G (2013) Clinicopathological analysis of colorectal cancer: a comparison between emergency and elective surgical cases. *World journal of surgical oncology* 11**:** 133

Ghesquieres H, Slager SL, Jardin F, Veron AS, Asmann YW, Maurer MJ, et al. (2015) Genome-Wide Association Study of Event-Free Survival in Diffuse Large B-Cell Lymphoma Treated With Immunochemotherapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 33(33)**:** 3930-7

Goossens-Beumer IJ, Derr RS, Buermans HP, Goeman JJ, Bohringer S, Morreau H, et al. (2015a) 40 MicroRNA classifier and nomogram for metastasis prediction in colon cancer. *Cancer Epidemiol Biomarkers Prev* 24(1)**:** 187-97

Goossens-Beumer IJ, Derr RS, Buermans HP, Goeman JJ, Bohringer S, Morreau H, et al. (2015b) MicroRNA classifier and nomogram for metastasis prediction in colon cancer. *Cancer Epidemiology, Biomarkers & Prevention* 24(1)**:** 187-97

Guastadisegni C, Colafranceschi M, Ottini L, Dogliotti E (2010) Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. *European journal of cancer* 46(15): 2788-98

Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. (2015) The consensus molecular subtypes of colorectal cancer. *Nat Med* 21(11): 1350-6

Guo Q, Schmidt MK, Kraft P, Canisius S, Chen C, Khan S, et al. (2015) Identification of novel genetic markers of breast cancer survival. *Journal of the National Cancer Institute* 107(5)

Guyot P, Welton NJ, Ouwens MJ, Ades AE (2011) Survival time outcomes in randomized, controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 14(5): 640-6

Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15(4): 361-87

Harrell Jr FE (2015) *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*: Springer

Haydon AM, Macinnis RJ, English DR, Giles GG (2006) Effect of physical activity and body size on survival after diagnosis with colorectal cancer. *Gut* 55(1): 62-7

He Y, Li X, Gasevic D, Brunt E, McLachlan F, Millenson M, et al. (2018a) Statins and Multiple Noncardiovascular Outcomes: Umbrella Review of Meta-analyses of Observational Studies and Randomized Controlled Trials. *Ann Intern Med* 169(8): 543-553

He Y, Ong Y, Li X, Din FV, Brown E, Timofeeva M, et al. (2019a) Performance of prediction models on survival outcomes of colorectal cancer with surgical resection: A systematic review and meta-analysis. *Surg Oncol* 29: 196-202

He Y, Theodoratou E, Li X, Din FVN, Vaughan-Shaw P, Svinti V, et al. (2019b) Effects of common genetic variants associated with colorectal cancer risk on survival outcomes after diagnosis: A large population-based cohort study. *International journal of cancer* 145(9): 2427-2432

He Y, Timofeeva M, Farrington SM, Vaughan-Shaw P, Svinti V, Walker M, et al. (2018b) Exploring causality in the association between circulating 25-hydroxyvitamin D and colorectal cancer risk: a large Mendelian randomisation study. *BMC Med* 16(1): 142

Heinze G, Wallisch C, Dunkler D (2018) Variable selection - A review and recommendations for the practicing statistician. *Biom J* 60(3): 431-449

Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL (2017) Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet* 95: 1 22 1-1 22 23

Higgins JP, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Stat Med* 21(11): 1539-58

Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. *Bmj* 327(7414): 557-60

Higgins JP, Thompson SG, Spiegelhalter DJ (2009) A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 172(1): 137-159

Hodge DR, Hurt EM, Farrar WL (2005) The role of IL-6 and STAT3 in inflammation and cancer. *European journal of cancer* 41(16): 2502-12

Hosmer DW, Lemeshow S (2000) Applied Logistic Regression (2nd Edition). New York, NY: John Wiley & Sons.

Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) *Applied logistic regression*. Vol. 398: John Wiley & Sons

Hou CC, Chen YP, Wu JH, Huang CC, Wang SY, Yang NS, et al. (2007) A galactolipid possesses novel cancer chemopreventive effects by suppressing inflammatory mediators and mouse B16 melanoma. *Cancer Res* 67(14): 6907-15

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* 5(6): e1000529

Hsieh FY, Lavori PW (2000) Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled clinical trials* 21(6): 552-60

Hu YY, Zheng MH, Zhang R, Liang YM, Han H (2012) Notch signaling pathway and cancer metastasis. *Adv Exp Med Biol* 727: 186-98

Hu Z, Ding J, Ma Z, Sun R, Seoane JA, Scott Shaffer J, et al. (2019) Quantitative evidence for early metastatic seeding in colorectal cancer. *Nature genetics* 51(7): 1113-1122

Huang L, Wang C, Rosenberg NA (2009) The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am J Hum Genet* 85(5): 692-8

Hueman M, Wang H, Henson D, Chen D (2019) Expanding the TNM for cancers of the colon and rectum using machine learning: a demonstration. *ESMO Open* 4(3): e000518

Hughes RA, Heron J, Sterne JAC, Tilling K (2019) Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International journal of epidemiology* 48(4): 1294-1304

Huncharek M, Muscat J, Kupelnick B (2009) Colorectal cancer risk and dietary intake of calcium, vitamin D, and dairy products: a meta-analysis of 26,335 cases from 60 observational studies. *Nutr Cancer* 61(1): 47-69

Hunter K, Welch DR, Liu ET (2003) Genetic background is an important determinant of metastatic potential. *Nature genetics* 34(1): 23-4; author reply 25

Hunter KW, Broman KW, Voyer TL, Lukes L, Cozma D, Debies MT, et al. (2001) Predisposition to efficient mammary tumor metastatic progression is linked to the breast cancer metastasis suppressor gene Brms1. *Cancer Res* 61(24): 8866-72

Huret JL, Ahmad M, Arsaban M, Bernheim A, Cigna J, Desangles F, et al. (2013) Atlas of genetics and cytogenetics in oncology and haematology in 2013. *Nucleic Acids Res* 41(Database issue): D920-4

Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. (2019) Discovery of common and rare genetic risk variants for colorectal cancer. *Nature genetics* 51(1): 76-87

Iacopetta B, Grieu F, Joseph D, Elsaleh H (2001) A polymorphism in the enhancer region of the thymidylate synthase promoter influences the survival of colorectal cancer patients treated with 5-fluorouracil. *British journal of cancer* 85(6): 827-30

Ingui BJ, Rogers MA (2001) Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 8(4): 391-7

Inno A, Fanetti G, Di Bartolomeo M, Gori S, Maggi C, Cirillo M, et al. (2014) Role of MGMT as biomarker in colorectal cancer. *World J Clin Cases* 2(12): 835-9

IntHout J, Ioannidis JP, Rovers MM, Goeman JJ (2016) Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* 6(7): e010247

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nature genetics* 29(3): 306-9

Ioannidis JP, Tarone R, McLaughlin JK (2011) The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* 22(4): 450-6

Iveson TJ, Kerr RS, Saunders MP, Cassidy J, Hollander NH, Tabernero J, et al. (2018) 3 versus 6 months of adjuvant oxaliplatin-fluoropyrimidine combination therapy for colorectal cancer (SCOT): an international, randomised, phase 3, non-inferiority trial. *Lancet Oncol* 19(4): 562-578

Jia WH, Zhang B, Matsuo K, Shin A, Xiang YB, Jee SH, et al. (2013) Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nature genetics* 45(2): 191-6

Jiang K, Sun Y, Wang C, Ji J, Li Y, Ye Y, et al. (2015) Genome-wide association study identifies two new susceptibility loci for colorectal cancer at 5q23.3 and 17q12 in Han Chinese. *Oncotarget* 6(37): 40327-36

Jiang W, Wang PG, Zhan Y, Zhang D (2014) Prognostic value of p16 promoter hypermethylation in colorectal cancer: a meta-analysis. *Cancer investigation* 32(2): 43-52

Jiang X, Finucane HK, Schumacher FR, Schmit SL, Tyrer JP, Han Y, et al. (2019) Shared heritability and functional enrichment across six solid cancers. *Nat Commun* 10(1): 431

Jinesh GG, Sambandam V, Vijayaraghavan S, Balaji K, Mukherjee S (2018) Molecular genetics and cellular events of K-Ras-driven tumorigenesis. *Oncogene* 37(7): 839-846

Johnson CM, Wei C, Ensor JE, Smolenski DJ, Amos CI, Levin B, et al. (2013) Meta-analyses of colorectal cancer risk factors. *Cancer Causes Control* 24(6): 1207-22

Johnson JR, Lacey JV, Jr., Lazovich D, Geller MA, Schairer C, Schatzkin A, et al. (2009) Menopausal hormone therapy and risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 18(1): 196-203

Joshi PK, Pirastu N, Kentistou KA, Fischer K, Hofer E, Schraut KE, et al. (2017) Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nat Commun* 8(1): 910

Kaler AS, Purcell LC (2019) Estimation of a significance threshold for genome-wide association studies. *BMC genomics* 20(1): 618

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 53(282): 457-481

Karrison T, Kocherginsky M (2018) Restricted mean survival time: Does covariate adjustment improve precision in randomized clinical trials? *Clinical trials* 15(2): 178-188

Kavvoura FK, McQueen MB, Khoury MJ, Tanzi RE, Bertram L, Ioannidis JP (2008) Evaluation of the potential excess of statistically significant findings in published genetic association studies: application to Alzheimer's disease. *Am J Epidemiol* 168(8): 855-65

Kawai K, Ishihara S, Yamaguchi H, Sunami E, Kitayama J, Miyata H, et al. (2015) 63 Nomograms for predicting the prognosis of stage IV colorectal cancer after curative resection: a multicenter retrospective study. *Eur J Surg Oncol* 41(4): 457-65

Kerns SL, West CM, Andreassen CN, Barnett GC, Bentzen SM, Burnet NG, et al. (2014) Radiogenomics: the search for genetic predictors of radiotherapy response. *Future Oncol* 10(15): 2391-406

Kerr RS, Love S, Segelov E, Johnstone E, Falcon B, Hewett P, et al. (2016) Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (QUASAR 2): an open-label, randomised phase 3 trial. *Lancet Oncol* 17(11): 1543-1557

Khan S, Fagerholm R, Kadalayil L, Tapper W, Aittomaki K, Liu J, et al. (2018) Meta-analysis of three genome-wide association studies identifies two loci that predict survival and treatment outcome in breast cancer. *Oncotarget* 9(3): 4249-4257

Khan S, Fagerholm R, Rafiq S, Tapper W, Aittomaki K, Liu J, et al. (2015) Polymorphism at 19q13.41 Predicts Breast Cancer Survival Specifically after Endocrine Therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* 21(18): 4086-4096

Knijn N, Mogk SC, Teerenstra S, Simmer F, Nagtegaal ID (2016) Perineural Invasion is a Strong Prognostic Factor in Colorectal Cancer: A Systematic Review. *The American journal of surgical pathology* 40(1): 103-12

Koster R, Panagiotou OA, Wheeler WA, Karlins E, Gastier-Foster JM, Caminada de Toledo SR, et al. (2018) Genome-wide association study identifies the GLDC/IL33 locus associated with survival of osteosarcoma patients. *International journal of cancer* 142(8)**:** 1594-1601

Kramer AA, Zimmerman JE (2007) Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Critical care medicine* 35(9)**:** 2052-6

Kudryavtseva AV, Lipatova AV, Zaretsky AR, Moskalev AA, Fedorova MS, Rasskazova AS, et al. (2016) Important molecular genetic markers of colorectal cancer. *Oncotarget* 7(33)**:** 53959-53983

Kyriakos M (1985) The President's cancer, the Dukes classification, and confusion. *Arch Pathol Lab Med* 109(12)**:** 1063-6

Lamain–de Ruiter M, Kwee A, Naaktgeboren CA, Franx A, Moons KG, Koster MP (2017) Prediction models for the risk of gestational diabetes: a systematic review. *Diagnostic and prognostic research* 1(1)**:** 3

Laohavinij S, Maneechavakajorn J, Techatanol P (2010) Prognostic factors for survival in colorectal cancer patients. *J Med Assoc Thai* 93(10)**:** 1156-66

Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, et al. (2019) Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun* 10(1)**:** 2154

Lenz HJ, Ou FS, Venook AP, Hochster HS, Niedzwiecki D, Goldberg RM, et al. (2019) Impact of Consensus Molecular Subtype on Survival in Patients With Metastatic Colorectal Cancer: Results From CALGB/SWOG 80405 (Alliance). *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 37(22)**:** 1876-1885

Lettre G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol* 31(4)**:** 358-62

Lewontin RC (1964) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* 49(1)**:** 49-67

Li J, Guo BC, Sun LR, Wang JW, Fu XH, Zhang SZ, et al. (2014) TNM staging of colorectal cancer should be reconsidered by T stage weighting. *World journal of gastroenterology* 20(17)**:** 5104-12

Li LC, Chui RM, Sasaki M, Nakajima K, Perinchery G, Au HC, et al. (2000) A single nucleotide polymorphism in the E-cadherin gene promoter alters transcriptional activities. *Cancer Res* 60(4)**:** 873-6

Li XL, Zhou J, Chen ZR, Chng WJ (2015) P53 mutations in colorectal cancer - molecular pathogenesis and pharmacological reactivation. *World journal of gastroenterology* 21(1)**:** 84-93

Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 1(6)**:** 417-425

Lifsted T, Le Voyer T, Williams M, Muller W, Klein-Szanto A, Buetow KH, et al. (1998) Identification of inbred mouse strains harboring genetic modifiers of mammary tumor age of onset and metastatic progression. *International journal of cancer* 77(4)**:** 640-4

Lindstrom LS, Hall P, Hartman M, Wiklund F, Gronberg H, Czene K (2007) Familial concordance in cancer survival: a Swedish population-based study. *Lancet Oncol* 8(11)**:** 1001-6

Liu Y, Chen L, Ko TC, Fields AP, Thompson EA (2006) Evi1 is a survival factor which conveys resistance to both TGFbeta- and taxol-mediated cell death via PI3K/AKT. *Oncogene* 25(25)**:** 3565-75

Lopez I, L PO, Tucci P, Alvarez-Valin F, R AC, Marin M (2012) Different mutation profiles associated to P53 accumulation in colorectal cancer. *Gene* 499(1)**:** 81-7

Low SK, Chung S, Takahashi A, Zembutsu H, Mushiroda T, Kubo M, et al. (2013) Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan. *Cancer science* 104(8)**:** 1074-82

Lunn M, McNeil D (1995) Applying Cox regression to competing risks. *Biometrics* 51(2)**:** 524-32

Lv J, Li L (2019) Hub Genes and Key Pathway Identification in Colorectal Cancer Based on Bioinformatic Analysis. *BioMed research international* 2019**:** 1545680

Maddams J, Utley M, Moller H (2012) Projections of cancer prevalence in the United Kingdom, 2010-2040. *British journal of cancer* 107(7)**:** 1195-202

Makondi PT, Chu CM, Wei PL, Chang YJ (2017) Prediction of novel target genes and pathways involved in irinotecan-resistant colorectal cancer. *PloS one* 12(7)**:** e0180616

Manceau G, Imbeaud S, Thiebaut R, Liebaert F, Fontaine K, Rousseau F, et al. (2014) Hsa-miR-31-3p expression is linked to progression-free survival in patients with KRAS wild-type metastatic colorectal cancer treated with anti-EGFR therapy. *Clinical Cancer Research* 20(12)**:** 3338-47

Mao X, Young BD, Lu YJ (2007) The application of single nucleotide polymorphism microarrays in cancer research. *Curr Genomics* 8(4)**:** 219-28

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11(7)**:** 499-511

Marigorta UM, Rodriguez JA, Gibson G, Navarro A (2018) Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet* 34(7)**:** 504-517

Marley AR, Nan H (2016) Epidemiology of colorectal cancer. *International journal of molecular epidemiology and genetics* 7(3)**:** 105-114

McGeoch L, Saunders CL, Griffin SJ, Emery JD, Walter FM, Thompson DJ, et al. (2019) Risk Prediction Models for Colorectal Cancer Incorporating Common Genetic Variants: A Systematic Review. *Cancer Epidemiol Biomarkers Prev* 28(10)**:** 1580-1593

McKeigue P (2018) Quantifying performance of a diagnostic test as the expected information for discrimination: Relation to the C-statistic. *Statistical methods in medical research***:** 962280218776989

McKeigue P (2019) Quantifying performance of a diagnostic test as the expected information for discrimination: Relation to the C-statistic. *Statistical methods in medical research* 28(6)**:** 1841-1851

Mefford J, Witte JS (2012) The Covariate's Dilemma. *PLoS genetics* 8(11)**:** e1003096

Midgley RS, McConkey CC, Johnstone EC, Dunn JA, Smith JL, Grumett SA, et al. (2010) Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 28(30)**:** 4575-80

Mlecnik B, Van den Eynde M, Bindea G, Church SE, Vasaturo A, Fredriksen T, et al. (2018) 175 Comprehensive Intrametastatic Immune Quantification and Major Impact of Immunoscore on Survival. *J Natl Cancer Inst* 110(1)**:** 01

Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 151(4)**:** 264-9, W64

Molinari F, Frattini M (2013) Functions and Regulation of the PTEN Gene in Colorectal Cancer. *Front Oncol* 3**:** 326

Montazeri Z, Li X, Nyiraneza C, Ma X, Timofeeva M, Svinti V, et al. (2019) Systematic meta-analyses, field synopsis and global assessment of the evidence of genetic association studies in colorectal cancer. *Gut*

Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al.

(2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 162(1)**:** W1-73

Moons KG, Altman DG, Vergouwe Y, Royston P (2009) Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj* 338**:** b606

Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. (2014) Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 11(10)**:** e1001744

Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. (2019) PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 170(1)**:** W1-W33

Moore KN, Tritchler D, Kaufman KM, Lankes H, Quinn MCJ, Ovarian Cancer Association C, et al. (2017) Genome-wide association study evaluating single-nucleotide polymorphisms and outcomes in patients with advanced stage serous ovarian or primary peritoneal cancer: An NRG Oncology/Gynecologic Oncology Group study. *Gynecologic oncology* 147(2)**:** 396-401

Morishita K, Parker DS, Mucenski ML, Jenkins NA, Copeland NG, Ihle JN (1988) Retroviral activation of a novel gene encoding a zinc finger protein in IL-3-dependent myeloid leukemia cell lines. *Cell* 54(6)**:** 831-40

Morson B, Sobin L (1976) Histological Typing of Intestinal Tumours (International Histological Classification of Tumours, No. 15). 1st ed Geneva, Switzerland: World Health Organization.

Munkholm P (2003) Review article: the incidence and prevalence of colorectal cancer in inflammatory bowel disease. *Aliment Pharmacol Ther* 18 Suppl 2**:** 1-5

Munro AJ, Lain S, Lane DP (2005) P53 abnormalities and outcomes in colorectal cancer: a systematic review. *British journal of cancer* 92(3)**:** 434-44

Nagaoka T, Kitaura K, Miyata Y, Kumagai K, Kaneda G, Kanazawa H, et al. (2016) Downregulation of epidermal growth factor receptor family receptors and ligands in a mutant K-ras group of patients with colorectal cancer. *Molecular medicine reports* 13(4)**:** 3514-20

NCI (2018) Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1975-2016), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2019, based on the November 2018 submission.

Negri E, Franceschi S, Parpinel M, La Vecchia C (1998) Fiber intake and risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 7(8)**:** 667-71

Nguyen HT, Duong HQ (2018) The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy. *Oncology letters* 16(1)**:** 9-18

Nieman KM, Romero IL, Van Houten B, Lengyel E (2013) Adipose tissue and adipocytes support tumorigenesis and metastasis. *Biochimica et biophysica acta* 1831(10)**:** 1533-41

Nitsche U, Zimmermann A, Spath C, Muller T, Maak M, Schuster T, et al. (2013) Mucinous and signet-ring cell colorectal cancers differ from classical adenocarcinomas in tumor biology and prognosis. *Ann Surg* 258(5)**:** 775-82; discussion 782-3

Noffsinger AE (2009) Serrated polyps and colorectal cancer: new pathway to malignancy. *Annu Rev Pathol* 4**:** 343-64

O'Donnell PH, Stark AL, Gamazon ER, Wheeler HE, McIlwee BE, Gorsic L, et al. (2012) Identification of novel germline polymorphisms governing capecitabine sensitivity. *Cancer* 118(16)**:** 4063-73

Oba K, Paoletti X, Alberts S, Bang YJ, Benedetti J, Bleiberg H, et al. (2013) Disease-free survival as a surrogate for overall survival in adjuvant trials of gastric cancer: a meta-analysis. *Journal of the National Cancer Institute* 105(21): 1600-7

Ouimet M, Drouin S, Lajoie M, Caron M, St-Onge P, Gioia R, et al. (2017) A childhood acute lymphoblastic leukemia-specific lncRNA implicated in prednisolone resistance, cell proliferation, and migration. *Oncotarget* 8(5): 7477-7488

Owzar K, Li Z, Cox N, Jung SH (2012) Power and sample size calculations for SNP association studies with censored time-to-event outcomes. *Genet Epidemiol* 36(6): 538-48

Ozaki T, Nakagawara A (2011) Role of p53 in Cell Death and Human Cancers. *Cancers (Basel)* 3(1): 994-1013

Pages F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R, et al. (2005) Effector memory T cells, early metastasis, and survival in colorectal cancer. *N Engl J Med* 353(25): 2654-66

Pander J, van Huis-Tanja L, Bohringer S, van der Straaten T, Gelderblom H, Punt C, et al. (2015) Genome Wide Association Study for Predictors of Progression Free Survival in Patients on Capecitabine, Oxaliplatin, Bevacizumab and Cetuximab in First-Line Therapy of Metastatic Colorectal Cancer. *PloS one* 10(7): e0131091

Park S, Hendry DJ (2015) Reassessing Schoenfeld residual tests of proportional hazards in political science event history analyses. *American Journal of Political Science* 59(4): 1072-1087

Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 32(4): 381-5

Pei YF, Li J, Zhang L, Papasian CJ, Deng HW (2008) Analyses and comparison of accuracy of different genotype imputation methods. *PloS one* 3(10): e3551

Peltomaki P (2001) Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human molecular genetics* 10(7): 735-40

Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P (2012) Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol* 176(6): 473-81

Peng F, Hu D, Lin X, Chen G, Liang B, Chen Y, et al. (2018) 174 An in-depth prognostic analysis of baseline blood lipids in predicting postoperative colorectal cancer mortality: The FIESTA study. *Cancer Epidemiol* 52: 148-157

Penney ME, Parfrey PS, Savas S, Yilmaz YE (2019) A genome-wide association study identifies single nucleotide polymorphisms associated with time-to-metastasis in colorectal cancer. *BMC cancer* 19(1): 133

Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP (2009) Discovery properties of genome-wide association signals from cumulatively combined data sets. *Am J Epidemiol* 170(10): 1197-206

Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, et al. (2013) Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* 144(4): 799-807 e24

Peto R, Peto J (1972) Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)* 135(2): 185-198

Petrelli F, Coinu A, Cabiddu M, Borgonovo K, Lonati V, Ghilardi M, et al. (2015) Prognostic factors for survival with bevacizumab-based therapy in colorectal cancer patients: a systematic review and pooled analysis of 11,585 patients. *Medical oncology (Northwood, London, England)* 32(2): 456

Petrelli F, Tomasello G, Borgonovo K, Ghidini M, Turati L, Dallera P, et al. (2017) Prognostic Survival Associated With Left-Sided vs Right-Sided Colon Cancer A Systematic Review and Meta-analysis. *JAMA oncology* 3(2): 211-219

Phipps AI, Passarelli MN, Chan AT, Harrison TA, Jeon J, Hutter CM, et al. (2016) Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis. *Carcinogenesis* 37(1): 87-95

Pikarsky E, Porat RM, Stein I, Abramovitch R, Amit S, Kasem S, et al. (2004) NF-kappaB functions as a tumour promoter in inflammation-associated cancer. *Nature* 431(7007): 461-6

Pino MS, Chung DC (2010) The chromosomal instability pathway in colon cancer. *Gastroenterology* 138(6): 2059-72

Qu X, Sandmann T, Frierson H, Jr., Fu L, Fuentes E, Walter K, et al. (2016) Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter. *Oncogene* 35(50): 6403-6415

Rafiq S, Khan S, Tapper W, Collins A, Upstill-Goddard R, Gerty S, et al. (2014) A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis. *PloS one* 9(12): e101488

Rahbari NN, Aigner M, Thorlund K, Mollberg N, Motschall E, Jensen K, et al. (2010) Meta-analysis shows that detection of circulating tumor cells indicates poor prognosis in patients with colorectal cancer. *Gastroenterology* 138(5): 1714-26

Ramaswamy S, Ross KN, Lander ES, Golub TR (2003) A molecular signature of metastasis in primary solid tumors. *Nature genetics* 33(1): 49-54

Ranganathan S, Nakai K, Schonbach C (2018) *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*: Elsevier

Rausei S, Iovino D, Tenconi S, Mangano A, Inversini D, Boni L, et al. (2013) Impact of lymph node ratio on survival of colorectal cancer patients. *Int J Surg* 11 Suppl 1: S95-9

Real LM, Ruiz A, Gayan J, Gonzalez-Perez A, Saez ME, Ramirez-Lorca R, et al. (2014) A colorectal cancer susceptibility new variant at 4q26 in the Spanish population identified by genome-wide association analysis. *PloS one* 9(6): e101178

Rees M, Tekkis PP, Welsh FK, O'Rourke T, John TG (2008a) 105 Evaluation of long-term survival after hepatic resection for metastatic colorectal cancer: a multifactorial model of 929 patients. *Ann Surg* 247(1): 125-35

Rees M, Tekkis PP, Welsh FK, O'Rourke T, John TG (2008b) Evaluation of long-term survival after hepatic resection for metastatic colorectal cancer: a multifactorial model of 929 patients. *Annals of Surgery* 247(1): 125-35

Reynisson OH. Comparing different methods to perform GWAS on censored data. Master, University of Iceland 2018

Robins JM, Wang N (2000) Inference for imputation estimators. *Biometrika* 87(1): 113-124

Roder D, Karapetis CS, Olver I, Keefe D, Padbury R, Moore J, et al. (2019) Time from diagnosis to treatment of colorectal cancer in a South Australian clinical registry cohort: how it varies and relates to survival. *BMJ open* 9(9): e031421

Rogers AC, Winter DC, Heeney A, Gibbons D, Lugli A, Puppa G, et al. (2016) Systematic review and meta-analysis of the impact of tumour budding in colorectal cancer. *British journal of cancer* 115(7): 831-40

Roshyara NR, Horn K, Kirsten H, Ahnert P, Scholz M (2016) Comparing performance of modern genotype imputation methods in different ethnicities. *Scientific reports* 6: 34386

Rothwell PM, Wilson M, Elwin CE, Norrving B, Algra A, Warlow CP, et al. (2010) Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *Lancet* 376(9754): 1741-50

Sampson JR, Jones S, Dolwani S, Cheadle JP (2005) MutYH (MYH) and colorectal cancer. *Biochem Soc Trans* 33(Pt 4): 679-83

Sankila R, Aaltonen LA, Jarvinen HJ, Mecklin JP (1996) Better survival rates in patients with MLH1-associated hereditary colorectal cancer. *Gastroenterology* 110(3): 682-7

Sargent DJ, Wieand HS, Haller DG, Gray R, Benedetti JK, Buyse M, et al. (2005) Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 23(34): 8664-70

Sato Y, Yamamoto N, Kunitoh H, Ohe Y, Minami H, Laird NM, et al. (2011) Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel. *J Thorac Oncol* 6(1): 132-8

Schisterman EF, Cole SR, Platt RW (2009) Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 20(4): 488-95

Schmid D, Leitzmann MF (2014) Association between physical activity and mortality among breast cancer and colorectal cancer survivors: a systematic review and meta-analysis. *Annals of oncology : official journal of the European Society for Medical Oncology* 25(7): 1293-311

Schmit SL, Edlund CK, Schumacher FR, Gong J, Harrison TA, Huyghe JR, et al. (2018) Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *Journal of the National Cancer Institute*

Schmit SL, Schumacher FR, Edlund CK, Conti DV, Raskin L, Lejbkowicz F, et al. (2014) A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. *Carcinogenesis* 35(11): 2512-9

Schmoll HJ, Van Cutsem E, Stein A, Valentini V, Glimelius B, Haustermans K, et al. (2012) ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. *Annals of oncology : official journal of the European Society for Medical Oncology* 23(10): 2479-516

Schoenfeld D (1980) Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* 67(1): 145-153

Shackelford D, Kenific C, Blusztajn A, Waxman S, Ren R (2006) Targeted degradation of the AML1/MDS1/EVI1 oncoprotein by arsenic trioxide. *Cancer Res* 66(23): 11360-9

Shu XO, Long J, Lu W, Li C, Chen WY, Delahanty R, et al. (2012) Novel genetic markers of breast cancer survival identified by a genome-wide association study. *Cancer Res* 72(5): 1182-9

Silva A, Faria G, Araujo A, Monteiro MP (2019) Impact of adiposity on staging and prognosis of colorectal cancer. *Critical reviews in oncology/hematology* 145: 102857

Smit HA, Pinart M, Anto JM, Keil T, Bousquet J, Carlsen KH, et al. (2015) Childhood asthma prediction models: a systematic review. *Lancet Respir Med* 3(12): 973-84

Smith CG, Fisher D, Harris R, Maughan TS, Phipps AI, Richman S, et al. (2015) Analyses of 7,635 Patients with Colorectal Cancer Using Independent Training and Validation Cohorts Show That rs9929218 in CDH1 Is a Prognostic Marker of Survival. *Clinical cancer research : an official journal of the American Association for Cancer Research* 21(15): 3453-61

Song N, Choi JY, Sung H, Jeon S, Chung S, Park SK, et al. (2015) Prediction of breast cancer survival using clinical and genetic markers by tumor subtypes. *PloS one* 10(4): e0122413

Song Y, Altarejos J, Goodarzi MO, Inoue H, Guo X, Berdeaux R, et al. (2010) CRTC3 links catecholamine signalling to energy balance. *Nature* 468(7326): 933-9

Spindler BA, Bergquist JR, Thiels CA, Habermann EB, Kelley SR, Larson DW, et al. (2017) Incorporation of CEA Improves Risk Stratification in Stage II Colon Cancer. *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract* 21(5): 770-777

Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. (2016) PhenoScanner: a database of human genotype-phenotype associations.

Bioinformatics 32(20)**:** 3207-3209

Steyerberg EW (2019) Overfitting and optimism in prediction models. In *Clinical prediction models*, pp 95-112. Springer

Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 54(8)**:** 774-81

Study C, Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, et al. (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics* 40(12)**:** 1426-35

Syed H, Jorgensen AL, Morris AP (2016) Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. *Pharmacogenomics* 17(8)**:** 907-15

Szulkin R, Karlsson R, Whitington T, Aly M, Gronberg H, Eeles RA, et al. (2015) Genome-wide association study of prostate cancer-specific survival. *Cancer Epidemiol Biomarkers Prev* 24(11)**:** 1796-800

Tabuso M, Homer-Vanniasinkam S, Adya R, Arasaradnam RP (2017) Role of tissue microenvironment resident adipocytes in colon cancer. *World journal of gastroenterology* 23(32)**:** 5829-5835

Taieb J, Le Malicot K, Shi Q, Penault-Llorca F, Bouche O, Tabernero J, et al. (2017) Prognostic Value of BRAF and KRAS Mutations in MSI and MSS Stage III Colon Cancer. *Journal of the National Cancer Institute* 109(5)

Takakura Y, Okajima M, Kanemitsu Y, Kuroda S, Egi H, Hinoi T, et al. (2011) 186 External validation of two nomograms for predicting patient survival after hepatic resection for metastatic colorectal cancer. *World J Surg* 35(10)**:** 2275-82

Takeichi M (1991) Cadherin cell adhesion receptors as a morphogenetic regulator. *Science* 251(5000)**:** 1451-5

Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D (2019) Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 20(8)**:** 467-484

Tang H, Wei P, Chang P, Li Y, Yan D, Liu C, et al. (2017) Genetic polymorphisms associated with pancreatic cancer survival: a genome-wide association study. *International journal of cancer* 141(4)**:** 678-686

Tang S, Pan Y, Wang Y, Hu L, Cao S, Chu M, et al. (2015) Genome-wide association study of survival in early-stage non-small cell lung cancer. *Annals of surgical oncology* 22(2)**:** 630-5

Tanikawa C, Kamatani Y, Takahashi A, Momozawa Y, Leveque K, Nagayama S, et al. (2018) GWAS identifies two novel colorectal cancer loci at 16q24.1 and 20q13.12. *Carcinogenesis* 39(5)**:** 652-660

Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, et al. (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature genetics* 40(5)**:** 631-7

Theodoratou E, Farrington SM, Tenesa A, McNeill G, Cetnarskyj R, Barnetson RA, et al. (2008) Dietary vitamin B6 intake and the risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 17(1)**:** 171-82

Therneau TM, Grambsch PM, Fleming TR (1990) Martingale-based residuals for survival models. *Biometrika* 77(1)**:** 147-160

Thirunavukarasu P, Sukumar S, Sathaiah M, Mahan M, Pragatheeshwar KD, Pingpank JF, et al. (2011) C-stage in colon cancer: implications of carcinoembryonic antigen biomarker in staging, prognosis, and management. *Journal of the National Cancer Institute* 103(8)**:** 689-97

Timmers PR, Mounier N, Lall K, Fischer K, Ning Z, Feng X, et al. (2019) Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife* 8

Ting JC, Ye Y, Thomas GH, Ruczinski I, Pevsner J (2006) Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC Bioinformatics* 7**:** 25

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* 39(8)**:** 984-8

Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, et al. (2011) Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS genetics* 7(6)**:** e1002105

Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature genetics* 40(5)**:** 623-30

Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global cancer statistics, 2012. *CA: a cancer journal for clinicians* 65(2)**:** 87-108

Tuohy TM, Rowe KG, Mineau GP, Pimentel R, Burt RW, Samadder NJ (2014) Risk of colorectal cancer and adenomas in the families of patients with adenomas: a population-based study in Utah. *Cancer* 120(1)**:** 35-42

Tyczynski J, Demaret E, Parkin D (2003) Standards and guidelines for cancer registration in Europe. *IARC Technical Publication* 40**:** 69-73

Van Cutsem E, Cervantes A, Adam R, Sobrero A, Van Krieken JH, Aderka D, et al. (2016) ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Annals of oncology : official journal of the European Society for Medical Oncology* 27(8)**:** 1386-422

van Giessen A, Peters J, Wilcher B, Hyde C, Moons C, de Wit A, et al. (2017) Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 20(4)**:** 718-726

van Meer S, Leufkens AM, Bueno-de-Mesquita HB, van Duijnhoven FJB, van Oijen MGH, Siersema PD (2013) Role of dietary factors in survival and mortality in colorectal cancer: a systematic review. *Nutrition Reviews* 71(9)**:** 631-641

van Walraven C, Davis D, Forster AJ, Wells GA (2004) Time-dependent bias was common in survival analyses published in leading clinical journals. *J Clin Epidemiol* 57(7)**:** 672-82

Venderbosch S, Nagtegaal ID, Maughan TS, Smith CG, Cheadle JP, Fisher D, et al. (2014) Mismatch repair status and BRAF mutation status in metastatic colorectal cancer patients: a pooled analysis of the CAIRO, CAIRO2, COIN, and FOCUS studies. *Clinical cancer research : an official journal of the American Association for Cancer Research* 20(20)**:** 5322-30

Vickers AJ (2011) Prediction models in cancer care. *CA Cancer J Clin* 61(5)**:** 315-26

Virani S, Bilheem S, Chansaard W, Chitapanarux I, Daoprasert K, Khuanchana S, et al. (2017) National and Subnational Population-Based Incidence of Cancer in Thailand: Assessing Cancers with the Highest Burdens. *Cancers (Basel)* 9(8)

Wang H, Birkenbach M, Hart J (2000) Expression of Jun family members in human colorectal adenocarcinoma. *Carcinogenesis* 21(7)**:** 1313-7

Wang H, Burnett T, Kono S, Haiman CA, Iwasaki M, Wilkens LR, et al. (2014) Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat Commun* 5**:** 4613

Wang M, Gu D, Du M, Xu Z, Zhang S, Zhu L, et al. (2016) Common genetic variation in ETV6 is associated with colorectal cancer susceptibility. *Nat Commun* 7**:** 11478

Watanabe K, Taskesen E, van Bochoven A, Posthuma D (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 8(1)**:** 1826

Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, et al.

(2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature genetics* 38(7): 787-93

Whiffin N, Hosking FJ, Farrington SM, Palles C, Dobbins SE, Zgaga L, et al. (2014) Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Human molecular genetics* 23(17): 4729-37

White DJ, Unwin RD, Bindels E, Pierce A, Teng HY, Muter J, et al. (2013) Phosphorylation of the leukemic oncoprotein EVI1 on serine 196 modulates DNA binding, transcriptional repression and transforming ability. *PloS one* 8(6): e66510

Wolin KY, Yan Y, Colditz GA, Lee IM (2009) Physical activity and colon cancer prevention: a meta-analysis. *British journal of cancer* 100(4): 611-6

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318(5853): 1108-13

Wu C, Li D, Jia W, Hu Z, Zhou Y, Yu D, et al. (2013) Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. *Nature genetics* 45(6): 632-8

Wu C, Xu B, Yuan P, Miao X, Liu Y, Guan Y, et al. (2010) Genome-wide interrogation identifies YAP1 variants associated with survival of small-cell lung cancer patients. *Cancer Res* 70(23): 9721-9

Xu W, Xu J, Shestopaloff K, Dicks E, Green J, Parfrey P, et al. (2015) A genome wide association study on Newfoundland colorectal cancer patients' survival outcomes. *Biomark Res* 3: 6

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics* 47(10): 1114-20

Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. (2011) Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* 19(7): 807-12

Yeh CH, Bellon M, Nicot C (2018) FBXW7: a critical tumor suppressor of human cancers. *Molecular cancer* 17(1): 115

Yesupriya A, Evangelou E, Kavvoura FK, Patsopoulos NA, Clyne M, Walsh MC, et al. (2008) Reporting of human genome epidemiology (HuGE) association studies: an empirical assessment. *BMC medical research methodology* 8: 31

Yoon KA, Jung MK, Lee D, Bae K, Joo JN, Lee GK, et al. (2014) Genetic variations associated with postoperative recurrence in stage I non-small cell lung cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* 20(12): 3272-9

Yuan H, Dong Q, Zheng Ba, Hu X, Xu J-B, Tu S (2017) Lymphovascular invasion is a high risk factor for stage I/II colorectal cancer : a systematic review and meta-analysis. *Oncotarget* 8(28): 46565-46579

Zeng C, Matsuda K, Jia WH, Chang J, Kweon SS, Xiang YB, et al. (2016) Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology* 150(7): 1633-1645

Zhang B, Jia WH, Matsuda K, Kweon SS, Matsuo K, Xiang YB, et al. (2014) Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nature genetics* 46(6): 533-42

Zhang JX, Song W, Chen ZH, Wei JH, Liao YJ, Lei J, et al. (2013) Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis.[Erratum appears in Lancet Oncol. 2014 Jan;15(1):e4]. *Lancet Oncology* 14(13): 1295-306

Zhang L, Cao F, Zhang G, Shi L, Chen S, Zhang Z, et al. (2019) Trends in and Predictions of Colorectal Cancer Incidence and Mortality in China From 1990 to 2025.

Front Oncol 9**:** 98

Ziv E, Dean E, Hu D, Martino A, Serie D, Curtin K, et al. (2015) Genome-wide association study identifies variants at 16p13 associated with survival in multiple myeloma patients. *Nat Commun* 6**:** 7539

List of URLs

1-1 https://gco.iarc.fr/today/fact-sheets-cancers.

1-2 https://www.isdscotland.org/Health-Topics/Cancer/Cancer-Statistics/Colorectal/#summary

1-3 http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-One

1-4 https://www.wcrf-uk.org/uk/preventing-cancer/cancer-types/bowel-cancer

1-5 https://www.fascrs.org/patients/disease-condition/hereditary-colorectal-cancer-0

1-6 https://www.cancerresearchuk.org/about-cancer/bowel-cancer/stages-types-and-grades

1-7 https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html

1-8 http://atlasgeneticsoncology.org/Tumors/colonID5006.html

1-9 https://www.cancer.ca/en/cancer-information/cancer-type/colorectal/prognosis-and-survival/?region=on

1-10 https://www.ebi.ac.uk/gwas/

1-11 https://www.uptodate.com/contents/pathology-and-prognostic-determinants-of-colorectal-cancer

1-12 https://www.nice.org.uk/guidance/cg131

4-1 https://emea.support.illumina.com/

4-2 https://www.internationalgenome.org/

4-3 https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

4-4 https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/

4-5 https://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf

4-6 http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/DNA-Extraction-at-UK-Biobank-October-2014-1.pdf

4-7 http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/Affymetrix-UKB_WCSGAX-Genotype-Data-Generation-1.pdf

4-8 https://jmarchini.org/software/

4-9 http://www.ensembl.org/Homo_sapiens/Tools/

4-10 https://cran.r-project.org/web/packages/survival/index.html

4-11 https://cran.r-project.org/web/packages/survminer/index.html

4-12 https://www.r-project.org/

4-13 https://cran.r-project.org/web/packages/survSNP/index.html

4-14 https://cran.r-project.org/web/packages/glmnet/index.html

4-15 https://cran.r-project.org/web/packages/rms/index.html

4-16 https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf

4-17 https://cran.r-project.org/web/packages/ResourceSelection/ResourceSelection.pdf

4-18 https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#frequentist_tests

4-19 https://cran.r-project.org/web/packages/qqman/qqman.pdf

4-20 https://www.bioconductor.org/packages/devel/bioc/manuals/GWASTools/man/GWASTools.pdf

4-21 https://rdrr.io/cran/GenABEL/src/R/estlambda.R

4-22 https://cran.r-project.org/web/packages/meta/meta.pdf

4-23 https://fuma.ctglab.nl/

6-1 https://documentation.sas.com/

6-2 https://www.proteinatlas.org/ENSG00000234828-IQCM/tissue

6-3 https://gtexportal.org/home/

6-4 https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=2444

6-5 http://www.phenoscanner.medschl.cam.ac.uk/

6-6 https://www.proteinatlas.org/ENSG00000159625-DRC7/tissue