Development and Assessment of a Genetic and Environmental Risk Model for Colorectal Cancer

Sarah Briggs

Wadham College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Abstract

Colorectal cancer (CRC) is the third most common cancer globally, and is readily prevented by screening. Colonoscopy is the gold standard screening tool but is invasive, expensive, and time consuming. 'Risk-stratified' approaches to screening, using genetic or non-genetic risk predictors to guide screening decisions, may improve screening outcomes and utilise limited resources more effectively.

The aim of this thesis was to develop and evaluate risk prediction models for CRC, to improve our ability to predict CRC risk. An initial search for novel genetic risk loci began with a new genome-wide association study (GWAS), following which meta-analysis of 5 new GWAS with 10 existing GWAS identified 31 new risk loci for CRC. Genome-wide linkage analysis combined with whole genome sequencing in two early-onset CRC families, and three families with multiple adenomas, identified several new possible risk loci for CRC including a potentially pathogenic variant in *GFI-1*.

Six polygenic risk scores (PRS) for CRC were then developed and tested in UK Biobank, with SNPs and weights derived from the GWAS meta-analysis. Evaluation of PRS performance showed that a genome-wide approach using LDpred2 software performed best, with a top age- and sex-adjusted C-statistic of 0.733 (95% confidence interval 0.710-0.753) in logistic regression. The PRS performed less well in women compared to men, and in minority ethnic UK Biobank participants.

Integrated genetic and non-genetic models were then developed, combining the top-performing PRS with the non-genetic QCancer-10 (Colorectal Cancer) model, which is based on primary care data. QCancer-10+PRS models modestly improved performance compared to QCancer-10 alone, with C-statistics of 0.730 (0.720-0.741) compared to 0.693 (0.682-0.704) respectively in men, and similar improvements in women. Decision curve analysis indicated small incremental improvements in net benefit, and the absolute difference in predicted 5-year risk was small (0.15-0.3%).

This thesis improves our understanding of the genetic basis of CRC risk, and risk prediction for CRC. Given the modest improvements in risk prediction with the addition of PRS to the QCancer-10 model, and the current logistical and ethical implications of PRS testing, there is no clear justification for PRS implementation in bowel cancer screening in their current form.

Acknowledgements

I feel very lucky to have had the pleasure of working alongside enthusiastic, friendly and supportive colleagues, both at the Wellcome Centre and at HERC, through the course of my DPhil. In particular, thank you to Claire, David, Luke, Hayley, Annie, Chiara, Lennard, Emma, Emily, Enric, Laura (C), Laura (G), Ingrid, Paul, James, Paolo, Francesco, Liz (M), Murong, Patrick, Ana, Melvin, Mi Jun, Koen, John, Filipa, David, Liz (S), Rositsa, Apo, Jose, Larry. I have also had the good fortune to work with excellent colleagues outside of Oxford - in particular Philip has been generous in his assistance, and Richard initiated many thought provoking conversations.

A huge thanks to Ian many years of encouragement, debate, and some healthy skepticism, and to Sarah for welcoming me to HERC and for the pragmatic advice, kindness, and encouragement, particularly through COVID. Thanks to Julia and to James for their all their help and support. Working with them all has been a pleasure, and they have helped me to grow in confidence as an academic.

I am unendingly grateful to my parents. Thank you for the lifetime of love and encouragement, and more lately the childcare, flowers, and gardening.

Thank you, always, to Adam for his unending support, and to my girls, Iris and Agnes, for their patience, hugs, and welcome distractions.

Sarah Briggs
Wadham College, Oxford
September 2022

Preface

This thesis is a collection of research carried out by the author at the Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, between January 2017 and September 2022, under the supervision of Professor Sarah Wordsworth, Professor Ian Tomlinson, Professor Julia Hippisley-Cox, and Dr James East.

The research described is original, and no part has previously been submitted for any degree at this University or any other. The work presented was conducted by the author unless clearly stated. Noted exceptions include the genome-wide association study meta-analysis in Section 3.3 which was conducted by Dr Philip Law at the Institue for Cancer Research, and the inspection of data from the 100,000 Genomes Project in Section 3.4.2 which was undertaken by Professor Ian Tomlinson.

Parts of the work presented in Chapter 3 were published in Law et al. (*Nature Communications*, 2019). Much of the work presented in Chapters 4, 5, and 6 has been published in "Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study", Briggs et al., *British Medical Journal*, 2022;379:e071707, DOI: 10.1136/bmj-2022-071707. Figures and tables in these chapters are reproduced under Creative Commons licence BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`).

My DPhil studies were disrupted by COVID, with a return to clinical work during the first wave from April-July 2020, home-schooling my daughters during the second wave in January-March 2021, and additional childcare during required periods of self-isolation. As a result of these disruptions I amended the scope of my thesis. I had originally planned to undertake more extensive linkage analysis, but this was not feasible and so I present here a more limited project. In addition, I had intended to present a final chapter modelling the clinical and cost-effectiveness of approaches to risk stratified screening, which will now be evaluated outside of the scope of my DPhil.

Contents

List of Figures

List of Tables

List of Abbreviations

1000G . . . . . 1000 Genomes Project

AIC . . . . . Akaike's Information Criterion

ASR . . . . . Age-specific incidence rate

AUROC . . . . Area under receiver operating characteristic curve

BCSP . . . . . Bowel Cancer Screening Programme

BFDP . . . . . Bayesian False Discovery Probability

chr . . . . . Chromosome

cM . . . . . Centimorgan

CI . . . . . Confidence interval

CITL . . . . . Calibration in the large

CORGI . . . . Colorectal Tumour Gene Identification

CRC . . . . . Colorectal cancer

CSG . . . . . Cancer susceptibility gene

DALY . . . . . Disability-adjusted life-year

DCA . . . . . Decision Curve Analysis

DNA . . . . . Deoxyribonucleic acid

DSR . . . . . Directly standardised incidence rate

E/O . . . . . Expected to observed ratio

EPIC . . . . . The European Prospective Investigation into Cancer and Nutrition Study

EPP . . . . . Events per predictor

ExAC . . . . . Exome Aggregation Consortium

FAP . . . . . Famial adenomatous polyposis

FCCTX . . . . Familial colorectal cancer type X

FIT . . . . . Faecal immunochemical test

gFOBT . . . . Guaiac faecal occult blood test

GWAS . . . . . Genome-wide association study

IARC . . . . . International Agency for Cancer Research

IBD . . . . . Identity by descent

ICD-9 . . . . . International Classification of Diseases, Ninth Revision

ICD-10 . . . . International Classification of Diseases, Tenth Revision

IQR . . . . . Interquartile range

LD . . . . . linkage disequilibrium

LOH . . . . . loss of heterozygosity

LDpred2-inf . LDpred2 infinitesimal model

LDpred2-grid LDpred2 grid non-sparse model

LDpred2-grid-sp LDpred2 grid sparse model

LP . . . . . Linear predictor

MAF . . . . . Minor allele frequency

MAR . . . . . Missing at random

MCAR . . . . Missing completely at random

MET . . . . . Metabolic equivalent of task

MNAR . . . . Missing not at random

MMR . . . . . mismatch repair

MR . . . . . Mendelian randomisation

MSI . . . . . microsatellite unstable

MSS . . . . . microsatellite stable

NHS . . . . . National Health Service

NSAID . . . . Non-steroidal anti-inflammatory drug

NANSAID . . Non-aspirin non-steroidal anti-inflammatory drugs

NICE . . . . . National Institute for Health and Care Excellence

NGS . . . . . Next generation sequencing

NSC . . . . . National Screening Committee

ONS . . . . . Office for National Statistics

OR . . . . . Odds radio

PCA . . . . . Principal components analysis

PCR . . . . . Polymerase chain reaction

PCs . . . . . Principle components

PoBI . . . . . People of the British Isles

PPY . . . . . Per person year

PRS . . . . . Polygenic risk score

QC . . . . . Quality control

RAF . . . . . Risk allele frequency

RAF . . . . . Randomised controlled trial

SES . . . . . Socioeconomic status

SNP . . . . . Single nucleotide polymorphism

TCGA . . . . . The Cancer Genome Atlas

TRIPOD . . . Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

UC . . . . . ulcerative colitis

UK . . . . . United Kingdom

UKRI . . . . . UK Research Innovation

UKB . . . . . UK Biobank

VUS . . . . . Variant of uncertain significance

WCRF . . . . World Cancer Research Fund

WHO . . . . . World Health Organisation

Introduction

Colorectal cancer (CRC) is common and to a large degree, preventable. Many high-income countries undertake screening for CRC, including a well-established bowel cancer screening programme in the UK, with the aim of both preventing early non-cancerous lesions from progressing, or identifying an established cancer early when there are likely to be more treatment options. However, current screening approaches are fairly crude, and outside of surveillance of high risk families, applied to the population in a blanket fashion generally limited by simple parameters such as age, failing to address differences in cancer risk across the population. The idea of a more nuanced approach, variously called 'personalised' or 'risk-stratified' screening, is highly appealing. Under this scenario, assessment of an individual's risk (which might incorporate both their genetic predisposition and environmental exposures) is used to inform screening decisions.

In this thesis, I seek to further our understanding of risk-stratified prevention as applied to colorectal cancer and bowel cancer screening. There has been an explosion in interest in risk-stratified approaches across the field of medicine in the last ten years. Therefore, a key question for my thesis is whether polygenic risk scores (PRS) can predict colorectal cancer risk to a level which would be clinically useful, and whether adding PRS to a non-genetic risk model would improve risk prediction above non-genetic risk prediction alone in a clinically meaningful way.

Since I began my DPhil studies five years ago there has been considerable progress in this field, especially with large genome-wide association consortia efforts creating a 5-fold increase in the number of risk polymorphisms identified. A number of new CRC risk scores have been published, as well as several studies of clinical and cost-effectiveness of risk-stratified screening. In addition, bowel cancer

screening in England has changed, with the introduction of faecal immunochemical stool testing and an end to one-off flexible sigmoidoscopy. Beyond this, the healthcare system faces disruption caused by the COVID pandemic, with a need to optimise resource use.

In this thesis I aim to improve prediction of cancer risk using genetic and non-genetic risk, to facilitate longitudinal approaches to risk-stratified cancer screening. Chapter 1 reviews the existing literature and sets my research in context. Chapter 2 outlines the main methodologies used.

In Chapter 3 I search for new risk loci for CRC by genotyping a new colorectal cancer cohort, and evaluating this in a new genome-wide association study (GWAS). This newly genotyped dataset forms part of a new GWAS-meta-analysis, which is used ultimately used to derive polygenic risk scores. In this chapter I also use linkage analysis to try to identify novel colorectal cancer risk loci in a several early onset CRC and polyposis pedigrees.

My risk model development and validation work is conducted in UK Biobank, and I explore this dataset in Chapter 4. In Chapter 5 I develop and evaluate several new polygenic risk scores, comparing the performance of GWAS-significant scores with genome-wide approaches. Next, I evaluate the improvement in performance obtained by adding polygenic risk scores to the QCancer-10 non-genetic risk model (Chapter 6).

My discussion section reflects on my research findings, specifically highlighting the current limitations and future potential of risk-stratified screening. I conclude with final thoughts from the thesis and suggests areas in which future research efforts should be focused.

1 Background

In this chapter I summarise the existing literature in relation to my thesis. I begin by giving some background to colorectal cancer (CRC) and prevention through screening. After this I explore our understanding of risk for CRC, including genetic risk, hereditability of CRC, and epidemiological risk for CRC. I then discuss the background to risk-stratified screening, and existing risk-prediction models for CRC risk.

1.1 Colorectal cancer

Colorectal cancer is the fourth most common cancer in the UK, with approximately 43,000 cases annually [1]. Globally there were an estimated 2.17 million cases in 2019, and 1.09 million deaths [2].

The global incidence of CRC has increased over the last 30 years. The latest Global Burden of Disease study [2] reported an age-standardised incidence rates (ASIRs) of 26.7 per 100,000 in 2019, compared to 22.2 per 100,000 person years in 1990. Meanwhile, mortality and disability-adjusted life-years (DALYs) have reduced (Figure 1.1). These changing patterns inevitably vary within different demographic groups and the rising incidence is largely attributable to males, in whom age-standardised mortality rates and DALYs have actually worsened (Figure

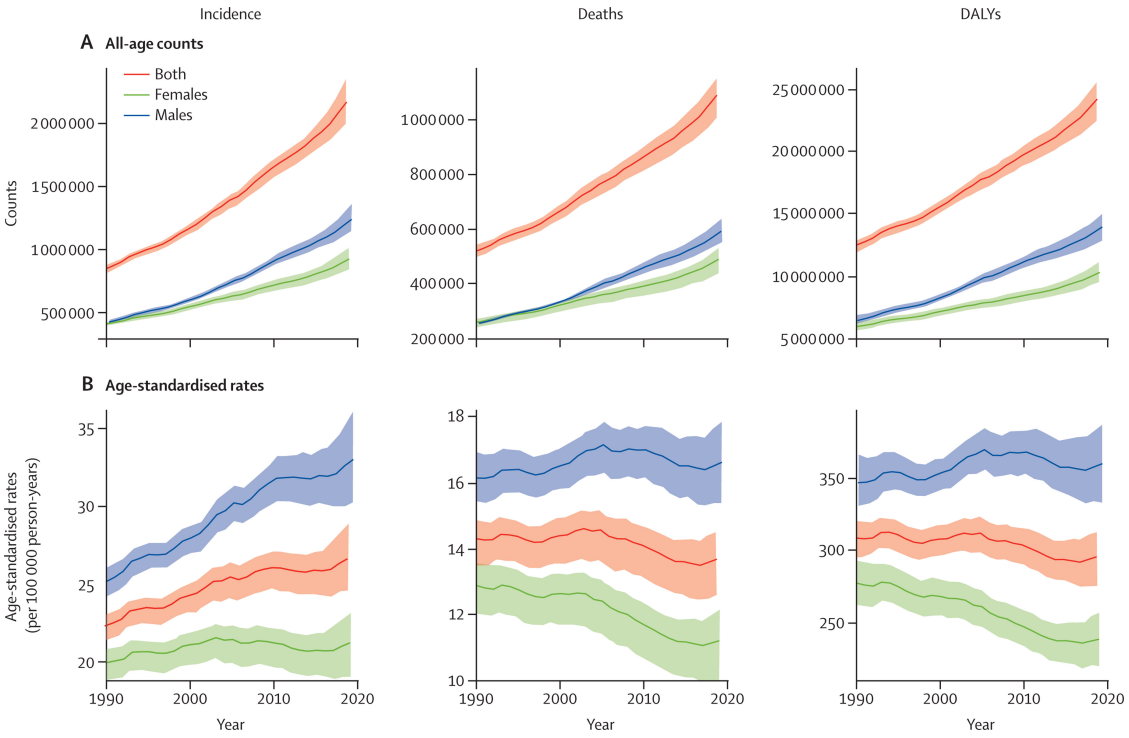

Figure 1.1: Global trends in colorectal cancer from 1990-2019. Figure taken from "The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017", GBD 2017 Colorectal Cancer Collaborators, DOI: 10.1016/S2468-1253(19)30345-0, CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`)

1.1B). Improvements in incidence and mortality are restricted to the highest quintile of the socio-demographic index internationally [2].

The burden of CRC is shifting internationally, with incidence increasing in low and middle-income countries since 1990 as lifestyles become more westernised, the largest increases being in regions of Asia and Latin America. Meanwhile in high-income countries, ASIRs have held steady or decreased, and mortality has fallen [2, 3]. The 2019 distribution of age-standardised incidence, mortality and DALYs globally is shown in Figure 1.2, highlighting higher incidence in high-income countries.

Colorectal cancer incidence has been rising in the under 50's in many high-income countries, including the UK, often despite stable or decreasing incidence in older age groups [4]. However, absolute numbers of cancers in this age group remain low [5]. Globally this pattern is again restricted to those in the highest socio-demographic index quintile, with ASIRs increasing across all age groups in other quintiles [2].

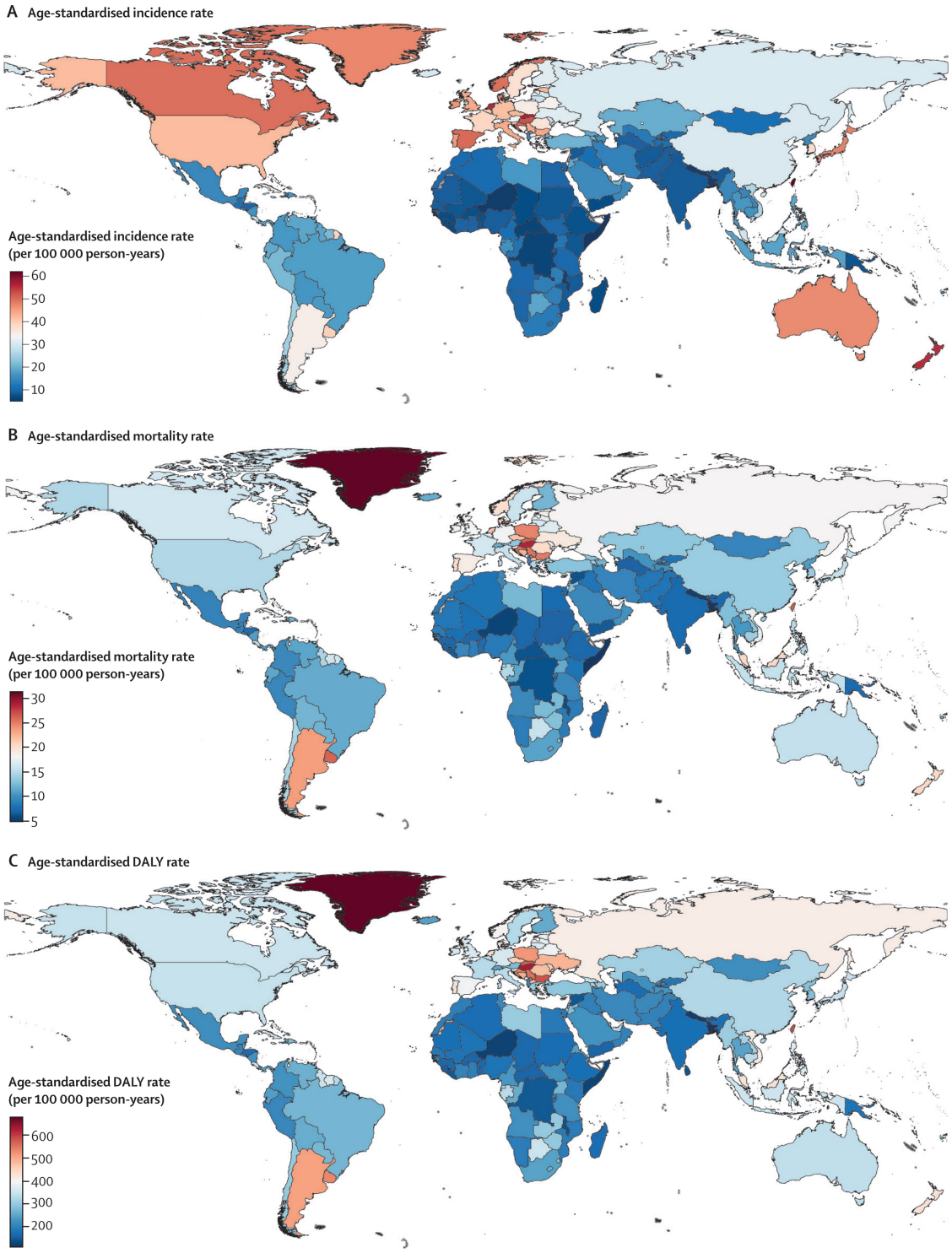

Figure 1.2: Global geographical distribution of colorectal cancer incidence, mortality and disability-adjusted life-years in 2019. Figure taken from "The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017", GBD 2017 Colorectal Cancer Collaborators, DOI: 10.1016/S2468-1253(19)30345-0, CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`)

1. Background

Most sporadic colorectal cancers are adenocarcinomas, which form from glandular tissue. These cancers nearly always arise from benign polyps (tubular adenomas or serrated polyps). Lockhart-Mummery and Dukes first suggested that CRC arose from adenomatous tissue, rather than normal colonic mucosa, in 1927 [6], and subsequently Jackman and Mayo [7] proposed the 'adenoma-carcinoma sequence'. Over 30 years later, Stryker et al. [8] reported a retrospective analysis following the natural history of colonic polyps $\geq$ 1cm in 226 patients followed up radiographically. They observed tumour growth in 37% of polyps, and 47% were ultimately excised, with invasive adenocarcinoma observed in 21 patients, many of whom already had advanced disease.

We now know that most CRCs arise due to genetic changes in stem-like cells in the colorectal epithelium [9]. The progressive genetic changes accompanying polyp progression were first identified by Vogelstein et al. [10], who observed increasing frequency of *ras* mutations, FAP-linked sequences on chromosome 5, and allelic deletions of chromosome 17 in small adenomas, large adenomas, and colorectal cancers respectively. Fearon and Vogelstein [11] proposed a genetic model for colorectal tumorigenesis, and since then several key pathogenic pathways have been identified, with characteristic genomic and mutational patterns (Figure 1.3, [12]).

Colorectal cancer is heterogeneous, reflecting both varied molecular pathways and the influence of the tumour microenvironment. It can be classified into subtypes based on molecular and morphological characteristics. Most cancers (~70-80%) follow the chromosomal instability (CIN) pathway, in which errors in chromosomal arrangement and separation during cell division (mitosis) result in abnormal structure or number of chromosomes. These cancers frequently show chromosomal abnormalities including aneuploidy and chromosomal rearrangement, somatic copy number alterations (SCNA), and loss of heterozygosity at tumour suppressor genes. This pathways is initiated by *APC* mutation, with subsequent mutations in key tumour suppressors and oncogenes including *KRAS*, *TP53*, *SMAD4* and *PIK3CA* [12, 13].

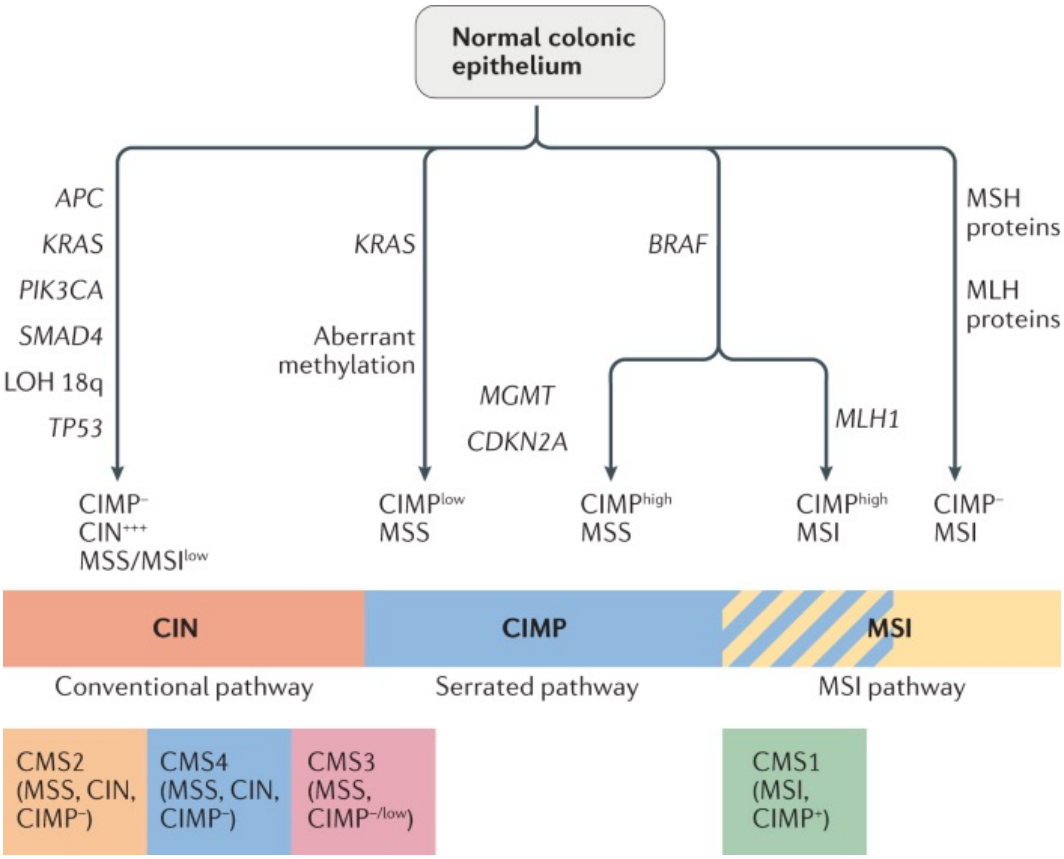

Figure 1.3: Molecular pathways of colorectal cancer. CIMP CpG island methylation pathway, CIN chromosomal instability, CMS consensus molecular subtype, MSI microsatellite instability, MSS microsatellite stable. Figure reprinted by permission from Springer Nature: Nature Reviews Immunology. The inflammatory pathogenesis of colorectal cancer. Schmitt et al. Copyright (2021)

The microsatellite instability (MSI) pathway also leads to genomic instability, and is defined by defects in the MMR genes. In somatic cancers this is secondary to silencing of MMR genes (typically *MLH1*) by hypermethylation of gene promotors [14, 15].

Twenty to thirty percent of CRCs follow the serrated adenoma pathway, with several characteristic subdivisions (see Figure 1.3). This pathway is often initiated by activation of the MAPK pathway (through either *KRAS* or *BRAF* mutation), and is characterised by the presence of CpG island methylation pathway (CIMP) mutations. *MGMT*, *CDKN2A* and *MLH1* are frequently silenced in these tumours.

Several projects have offered additional CRC classifications. For example, The Cancer Genome Atlas Study (TCGA) derived three groups of CRC, based on their genomic characteristics: CIN (~84% of tumours), hypermutated tumours (MMR-

deficient, MSI-high, *MLH1*-silenced, CIMP-high and *BRAF* mutant, ~13%), and ultramutated tumours (high number of C>A transversions with *POLE* or more rarely *POLD1* mutations; ~3%) [16]. The Consensus Molecular Subtype (CMS) Consortium identified four consensus molecular groups based on gene expression data (Figure 1.3) [17]. CMS1 represent an MSI-Immune subtype (~14%); CMS2, the canonical SCNA-high tumour subtype (~37%); CMS3, a metabolically dysregulated subtype (~13%); and CMS4, a mesenchymal subtype with TGF-β activation, angiogenesis, and stromal invasion. In addition, 13% of tumours in the study had mixed features, representing a transition phenotype and intratumoural heterogeneity. The CMS subtypes correspond to clinical characteristics: CMS1 tumours are largely right sided and more common in females, whilst CMS4 tumours tend to be more advanced, with poorer outcomes.

Colorectal cancer progression is influenced by the tumour microenvironment, with inflammation playing a key role. Many cell types within the tumour stroma, including fibroblasts, immune cells, vascular cells, and nerves, can positively or negatively influence tumour progression through their interaction with tumour cells [12]. In some cases this is related to the tumour's genetic make-up. For example *POLE* ultramutated tumours are highly immunogenic due to the high numbers of epitopes produced, with high numbers of CD8+ lymphocyte infiltration, and tend to have a relatively good prognosis [16, 18].

1.2 Colorectal cancer prevention

Colorectal cancer risk, at both a population and an individual level, can be reduced in a number of ways. Colorectal cancer is one of relatively few diseases with good evidence for the effectiveness of screening programmes. These have played a key role in reducing CRC incidence internationally over recent decades.

1.2.1 Bowel cancer screening

The principle of screening for chronic disease were first laid out over 50 years ago in a seminal paper by Wilson and Junger [19]. Ten criteria for screening were specified

Table 1.1: Wilson and Junger's principles of screening

Criteria	
1.	The condition is an important health problem;
2.	There should be an accepted treatment for those with the disease;
3.	Facilities for diagnosis and treatment should be available;
4.	There should be a recognised latent or early symptomatic stage;
5.	A suitable test or examination should exist;
6.	The test must be acceptable to the screening population;
7.	The natural history of the condition including the transition from latent to overt disease should be understood;
8.	There should be an agreed policy on who to treat;
9.	The cost of case finding, diagnosis and treatment should be economically balanced in relation to medical care as a whole;
10.	Case-finding should be continuous, rather than 'once and for all'.

to facilitate the identification of appropriate conditions for screening (Table 1.1), and methodologies for doing so in a public health context. The slow progression of CRC from benign adenoma to malignancy, occurring over many years, affords opportunities for prevention through screening [20]. Over the course of the late 20th century, the ten criteria have been fulfilled for CRC.

The primary aim of bowel cancer screening is to prevent cancer by detection and removal of precursor lesions, but screening also allows the down-staging of disease by identifying cancers earlier, leading to better cancer outcomes. Early demonstrations of the utility of bowel cancer screening included 26,000 individuals screened by rigid sigmoidoscopy (the visual inspection of the sigmoid colon through a rigid tube) by Hertz, Deddish, and Day [21] between 1946 and 1954, and subsequently a non-randomised 20,000+ patient study of rigid sigmoidoscopy and polypectomy (removal of polyps), conducted over 25 years from 1948 by Gilbertsen and Nelms [22]. These demonstrated large reductions in cancer incidence, and improvements in survival [23]. The development of colonoscopes, and of polypectomy through these, paved the way for the first randomised controlled trials (RCTs) [23].

Guaiac faecal occult blood (gFOBT) stool testing (where guaiac-coated cards are used to detect the presence of a small amount of blood on application of a stool sample), was found to identify individuals at risk of CRC. With subsequent advances in colonoscopy techniques, this led to the first RCTs of two-stage colorectal cancer

screening (that is, stool test followed by colonoscopy) in the UK, United States (US) and Denmark [24–26]. Using annual or biennial gFOBT testing, these trials demonstrated significant reductions in CRC-associated mortality, which was durable in long term follow-up, though there was no effect on all-cause mortality [27].

However, gFOBT has relatively limited sensitivity for advanced neoplasia at 11-37% (in comparison to 89-95% for the gold standard, colonoscopy) [28–30]. Meta-analyses suggested that whilst cancer-related mortality was reduced, there was no effect on cancer incidence [31]. Guaiac FOBT has sub-optimal uptake, requiring dietary modification (as it is not specific for human globin), and sampling over sequential days to improve sensitivity. The development of faecal immunochemical testing (FIT), an alternative human-specific quantitative test for occult faecal blood, which is easier to administer, has increased uptake by up to 10%, and is more sensitive than gFOBT for advanced neoplasia [28, 32–34]. Sensitivity of FIT for advanced adenoma is 0.43 (95% CI, 0.40-0.46). Sensitivity for cancer increases as might be expected with tumour stage (50% for Dukes A compared to 78% for Dukes C or D), and is greater for distal relative to proximal advanced neoplasia (30% and 16% respectively) [30, 32]. A further method of stool testing, incorporating faecal DNA with FIT, is more sensitive than FIT alone, but is less specific, with significantly higher numbers proceeding to colonoscopy, and it is more expensive than other stool tests [29, 35].

Colonoscopic screening is the gold standard, yet evidence from RCTs to support its effects on CRC mortality have appeared only recently. The non-randomised National Polyp Study in the 1980's demonstrated that removing adenomas by colonoscopy significantly reduced the incidence of colorectal cancer in colonoscopic screening, and longer term follow-up suggested cancer related mortality was reduced by about 50% [36, 37]. Colonoscopic screening was endorsed in several countries (notably in the US and Canada) on the basis of this and several other non-randomised studies. European guidelines did not recommend colonoscopic screening in part due to the lack of RCTs. Several RCTs of colonoscopy are now completing follow-up. In a multinational European study, screening of 55-64 year olds in Sweden,

Norway, Poland and the Netherlands resulted in high detection rates for both proximal and distal colon cancers, but participation rates were variable - from 60.7% in Norway to just 22.9% in the Netherlands [38]. Endoscopist performance (measured by caecal intubation and adenoma detection rates) was also variable, and this may have an impact upon outcomes in longer term follow-up. In a trial of colonoscopy compared to biannual FIT, participation was lower with colonoscopic screening, and whilst cancer detection rates were comparable, adenoma detection was better with colonoscopy [39].

Several additional screening modalities are also under evaluation, including colon capsule endoscopy (CCE) and magnetic resonance colonography [40]. A randomised trial of CCE to triage patients after positive stool testing is currently underway [41], and CCE may prove useful for further investigation where colonoscopy is incomplete [42].

Besides considering the clinical effectiveness of alternative screening methods, evaluation of cost-effectiveness allows health care decision makers to assess value for money for their populations and health systems. Health economic analyses of CRC screening have consistently demonstrated that it is cost effective when compared against no screening [43–45]. This may apply even in some resource limited settings [46], though there must be sufficient resources to treat detected cancers in this context.

Notably, beyond clinical- and cost-effectiveness, it is likely that in the future the environmental impact of different approaches will also be considered when prioritising screening strategies (which will require an entirely new evidence base). The NHS and World Health Organisation have committed to rapidly decarbonising healthcare in the UK and across the world [47, 48]. Endoscopy is a heavily resource intensive process [49], and institutions such as the National Institute for Health and Care Excellence (NICE) have begun to evaluate incorporating environmental sustainability into clinical guidelines [50].

Given the range of methods available and differing healthcare contexts, different approaches to screening have unsurprisingly been adopted internationally. Colonoscopic screening is invasive, time-consuming, and expensive; many countries have therefore adopted the two-stage approach of stool testing followed by colonoscopy. Screening for asymptomatic disease may be 'organised' or 'opportunistic' [51]. Organised screening requires the issuing of screening invitations from population-based registries, and centralised oversight of follow-up and quality assurance. This needs considerable infrastructural support and investment, rendering it out of reach in many countries. Opportunistic screening is offered on a more ad-hoc basis, with less formal oversight and more variability. Broadly, uptake of the two approaches internationally reflects differing healthcare philosophies, with 'socialised' planned approaches contrasting with free market approaches. Organised programmes aim to improve the health of the population as a whole, whilst opportunistic screening focuses on the health of the individual [51].

In the UK, a recognition that the screening outcomes demonstrated in RCTs might not be achievable in a national programme led to a pilot study of gFOBT-based bowel cancer screening to assess feasibility and effectiveness in a 'real' population prior to implementation [52]. The pilot study began in 2000, recruiting almost 500,000 individuals, with an uptake of 56%, and positive predictive values of 10.9% for cancer and 35% for adenoma [53]. This ultimately led to roll-out of the national bowel cancer screening programme for 60-74 year olds in 2006. Evaluation after the first million tests confirmed uptake in line with the pilot [54].

Bowel cancer screening in the UK is now devolved, and FIT was introduced in Scotland in 2017. There have been several modifications to bowel screening in England since its inception. Following a successful randomised trial, one-off flexible sigmoidoscopy ('Bowelscope') at 55 was offered, albeit with regional variation, in 2013 [55]. Subsequently, a pilot study of FIT-based testing demonstrated improved detection rates and uptake [56]. Health economic analysis of FIT-based screening using this pilot data demonstrated that FIT was more cost-effective than gFOBT at all FIT thresholds evaluated [57]. Cost savings and QALYs gained increased

as the FIT threshold was lowered, but with a considerable increase in the number of additional colonoscopies performed. FIT replaced gFOBT in 2019, resulting in an increase in uptake from around 60% to 67%.

In 2019, The NHS Long Term Plan committed to lowering the screening age to 50 [58], and this roll-out is currently in progress. Bowelscope was withdrawn in 2021, in part due to capacity constraints as a result of the commitment to lower the screening age [59]. In Europe, the EU council recommended population screening of 50-74 year olds in 2003 [60]. However, implementation has been highly variable, and by 2019 only 3 countries (Slovenia, Ireland, and France) had population screening programmes covering this age range in place [61].

Financial resources and colonoscopy capacity are variable across the EU, as are CRC incidence and mortality. Most countries have instigated organised screening programmes using faecal testing, whilst others including Germany and Austria use opportunistic colonoscopy [40]. Bowel cancer screening in the US is generally opportunistic, though some healthcare providers do have semi-organised systems [51]. There have, however, been clear guidelines for screening since the late 1990's, initially with gFOBT or sigmoidoscopy, and subsequently with colonoscopy [23, 62]. Currently the US Preventive Services Task Force guidelines recommend choosing screening modality (stool testing, colonoscopy or CT colonography) based on patient and organisational factors and preferences, whilst the US Multi-Society Task Force recommend colonoscopy every 10 years or annual FIT as the modalities of choice. In 2021 both recommended extending screening age to begin at 45 [30, 63].

Epidemiological evidence suggests that screening has reduced CRC incidence in relevant age groups in recent years. In a recent comparison of CRC incidence, stage, and mortality across Europe since 2000, substantial reductions in CRC ASIRs were evident in countries with established screening programmes. Meanwhile, ASIRs increased in most countries without screening, whilst successful implementation of screening in this period was reflected in a peak and then fall in ASIRs [64].

In addition to population screening, enhanced screening pathways exist for those with higher risk based on family history, or identification of a germline mutation in a cancer risk gene, which I will discuss further below [65].

1.2.2 Genetic risk and heritability

The extent to which genetic variation contributes to risk of complex traits is variable, but often considerable. Colorectal cancer estimates from twin studies (evaluating the cancer risk in sibling-pairs of monozygotic and dizygotic twins), put heritability at between 35-40%, with a familial relative risk of 2.2 [66, 67]. Around 5% (and perhaps as much as 10%) of CRCs overall are thought to be secondary to inherited cancer syndromes, most commonly Lynch Syndrome, with the remaining heritable risk accounted for by lower penetrance cancer susceptibility genes (CSGs), and polygenic risk [68–70]. Prevalence of CSG mutation increases in early onset CRC, at around 13% (with some estimates of up to 26% prevalence), while around 28% have a family history of CRC (though estimates vary) [71–73].

The most common highly penetrant CSGs, *APC*, *MLH1*, and *MSH2*, were identified through familial genetic linkage studies in the late 20th century, around the same time as the discovery of *BRCA1* and *BRCA2* in breast cancer in the 'first-wave' of cancer gene discovery [74–79]. Further linkage efforts have identified additional causative CRC and polyposis genes, including *SMAD4*, *BMPR1A*, *STK11*, *PTEN*, *GREM1*, *POLE* and *POLD1* [80–86]. In the most recent of these discoveries, combining linkage with next-generation sequencing (NGS) has permitted more rapid identification of the causal variants. Candidate gene studies focusing on DNA repair pathways have also uncovered rare causative variants such as *MUTYH*, *MSH6* [87, 88]. In the age of NGS, novel candidate risk genes are regularly reported, but their pathogenicity is often far more difficult to demonstrate [79]. It is essential that pathogenicity is rigorously proven, and cases replicated, before they are incorporated into clinical testing [89].

Hereditary CRC syndromes are broadly defined by the presence or absence of polyposis (the presence of tens or hundreds of polyps within the colon). Most of the

Mendelian genes identified cause polyposis, and specific genes tend to predispose to particular polyp types, as shown in Figure 1.4 [9, 90]. The most common mutations in non-polyposis syndrome are in mismatch repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*, or epigenetic silencing of *MSH2* secondary to a 3' deletion of *EPCAM*) which causes Lynch syndrome, responsible for around 3% of CRC cases [69]. These are dominantly inherited, though notably individuals with mutations in both copies of a Lynch gene have congenital mismatch repair deficiency (CMMRD) and are predisposed to colorectal polyps in addition to CRC [9].

In polyposis syndromes the most commonly affected genes are *APC* and *MUTYH* (the latter showing recessive inheritance). Many additional rarer or less penetrant genes have now been identified (Figure 1.4). Most recently, bi-allelic loss of *MDB4*, a base-excision repair gene (not shown in the figure), has been identified as a cause of syndromic adenomatous polyposis with uveal melanoma and acute myeloid leukaemia [91]. Age of onset, risk of additional cancers, and non-neoplastic characteristics vary by gene, and CRC penetrance varies, up to a 100% lifetime risk in familial adenomatous polyposis (FAP) [68].

As shown in Figure 1.4, Mendelian CRC genes affect a number of key functional pathways implicated in sporadic CRC pathogenesis. This includes DNA repair (for example mismatch repair in Lynch syndrome, or proofreading errors with *POLE* or *POLD1* mutations in Polymerase Proofreading Associated Polyposis); the bone morphogenetic protein (BMP) signalling pathway which controls stem-cell differentiation in colorectal crypts, Wnt signalling, and mTOR. Several of these genes are commonly inactivated in somatic tumours including *APC* (mutated in almost all CRCs as discussed above), *MLH1* (~15% of CRCs), *SMAD4* (~10%), and *POLE* (~2-3%) [9].

In this thesis I use genetic linkage studies to search for new risk loci for early onset CRC and polyposis. The mechanisms of inheritance underpinning linkage studies were first elucidated in the late 19th and early 20th century. Gregor Mendel's seminal paper, 'Experiments in Plant Hybridization', set out his principles of inheritance, which stated that units of inheritance assorted independently. The ratios of two

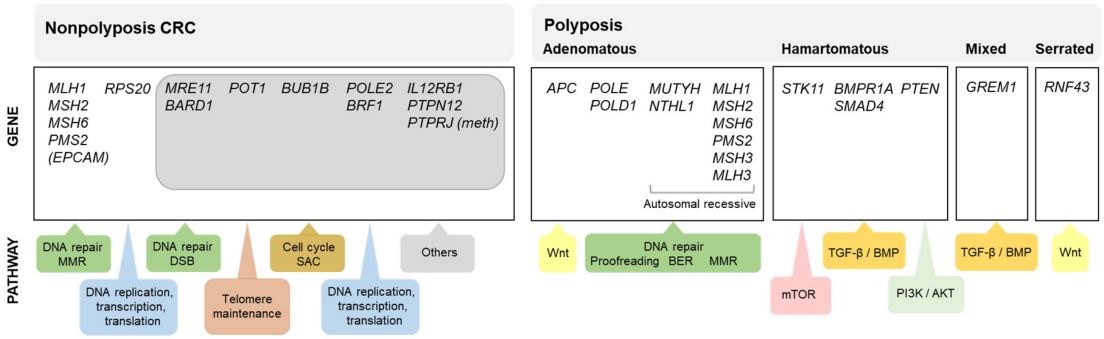

Figure 1.4: Familial colorectal cancer phenotypes, genes and affected pathways. BER, base excision repair; DSB, double strand breaks; meth, promoter hypermethylation; MMR, DNA mismatch repair; SAC, spindle assembly checkpoint. Figure taken from "Dominantly Inherited Hereditary Nonpolyposis Colorectal Cancer Not Caused by MMR Genes", Terradas et al., DOI: 10.3390/jcm9061954, CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`)

or more phenotypes in future generations would therefore be predictable [92]. Subsequently, scientists began to note situations in which Mendelian assumptions were broken. Bateson, Saunders, and Punnett [93] observed, in experiments crossing sweet peas with differing petal colour and pollen grain shape, that rather than seeing independent assortment of these characteristics, there was a marked excess of parental phenotypes in the offspring. They hypothesised that these alleles must be linked, but were unclear how this occurred.

Shortly after this, Thomas Hunt Morgan was experimenting with fruit flies, and identified a highly unusual fly within his colonies with white eyes instead of red. In crossbreeding this fly with normal flies, Morgan observed that there were no white-eyed females among the progeny, concluding that eye colour inheritance must be linked to sex inheritance [94]. Morgan hypothesised (building on work by Jannsens describing interweaving and merging of chromatids during meiosis [95]) that,

> *"Instead of random segregation in Mendel's sense we find 'association of factors' that are located near together in the chromosomes."*

He observed that when chromosomes split, genetic material contained within short distances would be more likely to locate on the same side, whilst for distant regions segregation would be more random [96], and thus identified genetic linkage [97].

The statistical approach of linkage analysis has been a powerful tool in genetic mapping, determining the order of loci on each chromosome. The recombination fraction derived from linkage was long used to describe genetic distance between loci, a 1% recombination fraction equating to one centimorgan (cM). Following shared genetic regions through family pedigrees (genetic linkage), identifying those common to affected individuals, allows mapping of disease traits, and this was the main approach to disease mapping in the latter twentieth century.

Beyond the CSGs identified through linkage studies, significant or suggestive peaks have also been found for CRC and polyposis syndromes at 1q22–q24.2, 3q21-24, 4q13.1, 4q21.1, 6p21.1-p12.1, 7q31, 9q22.32-31.1, 9q33.3–q34.3, 10p15.3–p15.1, 10q21.2–q23.1, 11q23, 12q24.32, 14q24.3–q31.1, 15q22.31, 18p11.2 and recessive loci at 8q13.2 and 9q31.1 [98–111].

Over the first decade of the 21st century, there was an increasing focus on the common disease-common variant hypothesis (this is, common phenotypes are the result of the additive - or multiplicative - effects of many common variants [112]), which built on classic population genetics theory [113]. Genome-wide association studies (GWAS), which use single nucleotide polymorphisms (SNPs) to map associations with disease phenotype, may be considered a complementary approach to linkage in the discovery of risk loci, this time focusing on common, low penetrance variation. In GWAS, polymorphism frequencies are compared between cases and controls to assess association with disease (described in more detail in Chapters 2 and 3). A locus is considered to be associated with disease at a significance threshold of $p<5x10^{-8}$. Small initial studies identified loci with the largest effect size (with ORs typically >1.2, the "low-hanging fruit"), and as GWAS sample size has increased, with meta-analysis of multiple studies, less common variants and those with smaller effects sizes (OR < 1.1) have been identified. Over the last 20 years, GWAS have identified many common loci for common diseases. For CRC, the number currently stands at 205 [114].

Notably, most GWAS variants have no clear functional role themselves, mapping to non-coding regions of the genome. The SNPs may exert their effects in a number

of different indirect ways, with much work conducted to functionally annotate GWAS 'hits' over the last decade (discussed in more detail in Section 3.5). Candidate GWAS loci tend to cluster within particular cellular pathways, highlighting mechanisms of disease. Law et al. [115] identified BMP/TGF-β signalling pathways, *MYC*-related mechanisms, and chromosomal integrity and DNA repair genes as key pathways in CRC pathogenesis.

Despite these efforts, a large proportion of heritability remains unaccounted for, termed 'missing heritability'. Dudbridge [116] argued that most of this will be common variation with very small effect, as yet undetected by GWAS. The most recent estimates for CRC based on GWAS data, suggest that 19% of CRC heritability can be explained by all common variation, accounting for 73% of familial relative risk (FRR), whilst the 205 identified CRC risk SNPs explain just 19.7% of FRR [114]. Alternatively unidentified rare mutations with strong effects may have a significant role [117], however recent endeavours have notably failed to identify additional high penetrance CRC CSGs, with the proposed genes contributing a small amount (if anything) to CRC heritability [90, 118].

Common diseases most likely display allelic and locus heterogeneity, that is, the causative genes vary between individuals, as does the allele within a given gene [117]. Early onset CRC may also be oligogenic, as a result of the interaction of multiple variants with moderate penetrance, which is supported by increased numbers of rare, functionally damaging variants in genes involved in known pathways for colorectal carcinogenesis in early onset cases [118].

1.2.3 Non-genetic risk

Recent estimates suggest that over half of CRC cases and deaths are caused by the effects of potentially modifiable risk factors [119]. In the Global Burden of Disease Colorectal Cancer study, the main contributors to CRC DALYs were smoking and alcohol intake in high-income countries, and low dietary intake of milk and calcium in Asia and sub-Saharan Africa (Figure 1.5) [2]. Relative contributions varied by sex, with smoking, alcohol and raised BMI contributing to a higher proportion of

global DALYs in males. The high prevalence of these life-style related risk factors in men probably contributes to the divergence in cancer incidence between sexes noted in Section 1.1 (indeed age-standardised CRC incidence was higher for women than men when first reported in the 1960's) [2, 120].

In a US study, the population-attributable fraction (PAF) for CRC was highest for alcohol intake (17.1% in men and 8.1% in women), low physical activity (15.7% in men and 16.8% in women) and cigarette smoking (13.5% in men and 9.7% in women). Low fibre intake had a PAF of 9.3% and 11.3%, processed meat consumption for 10.3% and 5.8% respectively, red meat consumption for 6.6% and 3.9%, and excess body weight accounted 5.1% and 5.4% in men and women respectively [119].

A recent umbrella review of meta-analyses of observational studies found evidence of association for only a limited range of dietary components across eleven cancer types [121]. Colorectal cancer was the most extensively studied, with 221 meta-analyses, and along with breast cancer, was the sole cancer for which strong evidence of dietary associations was found (that is, having statistically strong association and no evidence of bias). Strong evidence was found for a positive association of CRC risk with alcohol (summary relative risk (RR) 1.07 per 10g/day, [approximately 1 drink], 95% CI 0.83–0.90), and protective effects of dairy products (RR 0.87 per 400g/day [approximately 2 servings], 95% CI 0.83–0.90), seen also for milk and calcium intake, and whole grains (RR 0.84 per 90g/day [approximately 3 servings], 95% CI 0.78–0.90). These results are supported by the latest World Cancer Research Fund/American Institute for Cancer Research report, 'Diet, nutrition physical activity and cancer' (hereafter identified as the 'WCRF report' for brevity) [122].

The association of red meat (beef, lamb, pork and goat) and processed meat (preserved through curing, smoking, salting, or the addition of chemicals) with CRC has been highly publicised and controversial [123]. However, the International Agency for Research on Cancer (IARC) concluded that processed meat is 'carcinogenic to humans', and red meat 'probably carcinogenic to humans' [124]. These associations were judged to have suggestive evidence only in the review by Papadimitriou et al. [121], as significance was $\sim 10^{-4}$, failing to meet their stringent threshold of 10^{-6}.

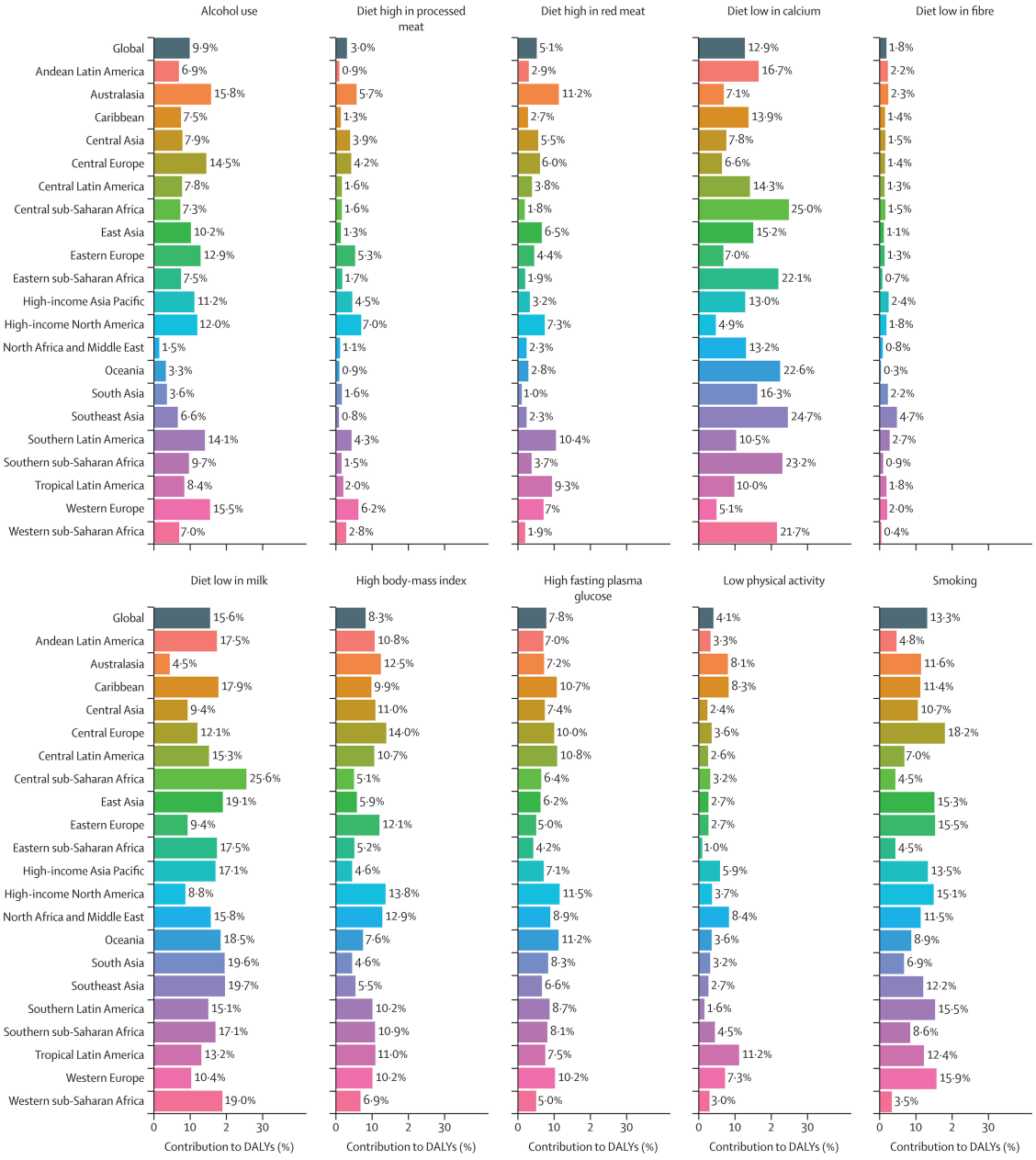

Figure 1.5: Global and regional percentage contributions of risk factors to colorectal cancer DALYs in 2019. Figure taken from "The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017", GBD 2017 Colorectal Cancer Collaborators, DOI: 10.1016/S2468-1253(19)30345-0, CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`)

However in the WCRF report, processed meat had 'convincing strong evidence' for an association with CRC risk, and red meat 'probable strong evidence' [122]. Dose-response meta-analysis of 15 studies (31,551 cases) demonstrated a 12% risk increase per 100g/day (RR 1.12, 95% CI 1.04-1.21, I^2 70.5%, $p_{het} < 0.0001$) [122, 125].

Foods high in dietary fibre also have strong suggestive evidence of a protective effect in the WCRF report, and protective effects of calcium and dairy intake are supported [122]. Other putative dietary risk factors with more limited suggestive evidence include vegetable or fruit intake, saturated fat intake, haem-iron containing foods, foods containing vitamin C, fish consumption, multivitamin use, and vitamin D [121, 122].

Aside from dietary intake, other nutritional factors - attained adult height and body fatness - have strong convincing evidence of a positive association with CRC risk [122]. Adult height is thought to be a composite measure of nutritional, hormonal, environmental and genetic exposures in early life. Dose-response meta-analysis of 13 studies (65,880 cases) showed a 5% increase in CRC risk for every 5cm height increase (RR 1.05, 95% CI 1.02-1.07, $I^2 = 90\%$, $p_{het} < 0.001$), with the association maintained in analysis by sex and by site (colon and rectum) [122, 126]. A causal effect for height and body composition is supported by Mendelian randomisation studies [127–129].

Increased levels of physical activity were also noted to be protective for colon cancer risk (with no clear conclusion for rectal cancer) by the WCRF. Physical activity is defined as 'any bodily movement produced by skeletal muscles that results in energy expenditure' [130]. It covers all aspects of daily movement, and can be divided into different domains: household, occupational, active travel, and leisure. In a meta-analysis of total physical activity in 12 studies (8,396 cases), a 20% decreased risk of colon cancer was seen in the highest compared to lowest activity groups (RR 0.80, 95% CI 0.72-0.88, I^2 39%, $p_{het} = 0.06$) [122].

Various health exposures are also associated with increased cancer risk. Most notably, inflammatory bowel disease (ulcerative colitis (UC), and Crohn's Disease) increases risk for CRC secondary to persistent inflammation [131, 132]. For UC, this

risk appears to be decreasing over time [133, 134], potentially due to improvements in disease control and surveillance, and widening use of immunomodulators [133–137]. Whilst the UC-associated CRC risk was previously thought to be much higher than with Crohn's, estimates now suggest risk is broadly similar [138]. Colorectal cancer risk is significantly higher risk if diagnosed in childhood [138–141].

Drug exposures, particularly non-steroidal anti-inflammatory drugs (NSAIDs), may also modulate CRC risk. There is strong trial-based evidence that the use of low dose (75mg) aspirin for five years reduces CRC incidence with a latency of 5-10 years [142, 143], and some evidence for non-aspirin NSAIDs in CRC prevention [144]. Metformin, statins, and bisphosphonates, have also been suggested in meta-analyses to have protective effects for CRC and adenoma risk [145–150].

Other environmental exposures may also have a role in CRC risk, including fine particular matter from air pollution (PM2.5), which has been suggested to increase both CRC incidence and mortality [151, 152].

1.3 Risk-stratified screening

Since the pilot study of organised population screening in England over twenty years ago, concerns have been raised about the ability of clinical services, particularly endoscopy capacity, to cope with this workload [52, 153]. Bowel screening services in England (and the UK more broadly) are under significant pressure. Colonoscopy capacity has not sufficiently increased to meet increasing demand [154]. This has been exacerbated by constraints in healthcare spending, workforce deficits, and the COVID-19 pandemic. Screening services were temporarily suspended during 2020, and whilst bowel cancer screening uptake has been relatively resilient in terms of recorded FIT participation (presumably because this can be done at home), colonoscopy capacity was significantly reduced [155]. Consequently, whilst the National Screening Committee recommended that FIT testing be introduced at 20μg/g (in line with a number of other countries), it is being introduced at 120μg/g to avoid overwhelming screening capacity [154]. This experience is common to

other programmes - the Netherlands introduced FIT at a threshold of 15 μg/g and increased this to 47μg/g due to service pressures [156].

Over recent years the interest in 'personalising' approaches to cancer prevention has grown. Stratified screening, in which the type of screening test and/or frequency of screening is varied depending on individual risk, offers a more nuanced approach. This focuses resources on higher risk individuals, and can potentially improve detection rates, prevent over-investigation of those at lower risk, and improve cost-effectiveness [157, 158]. Risk assessment in screening could also help informed consent and shared decision making by improving participants understanding of risk.

The English National Screening Committee has recently expanded its remit to be able to consider targeted and stratified screening programmes. Targeted screening, that is, screening for those at higher risk for a condition (for example as a result of a CSG mutation, or ulcerative colitis) were previously overseen by NICE, and stratified screening was not explicitly considered. Stratified screening has become a priority [159], with the four Chief Medical Officers of the UK now recommending

> *"A nationally delivered, proactive approach to screening, offering testing which varies in frequency and modality (type of test offered), according to the level of individual risk. This is designed to achieve a more favourable balance of benefits and harms at individual as well as population level. Stratified screening can be used to complement both targeted and population screening programmes."*
>
> UK National Screening Committee, 2022

Targeted, or 'enhanced', screening programmes for CRC have been in place for some time for those at increased familial risk, and in those with CSG mutations screening guidance is gene specific. For example, the recommended starting age for biannual colonoscopic screening in Lynch Syndrome is 10 years earlier for *MLH1* and *MSH2* carriers, compared to *MSH6* or *PMS2* [65]. Risk-stratified screening represents an extension to both population and targeted screening programmes, as it is likely that some individuals at the highest level of risk would be eligible for targeted screening based on current risk eligibility thresholds.

Risk-based screening can be considered in two broad forms. Cross-sectional risk-based screening predicts the risk of advanced neoplasia at the present moment, and could be used to select individuals for colonoscopy (with or without faecal testing). This might be especially helpful in detecting advanced adenomas, for which faecal testing is less predictive than for cancers. Longitudinal screening on the other hand considers future risk of advanced neoplasia, using this to guide screening pathways, and inform participants of their risk [160].

In addition to risk stratification based on common genetic variation, there is also an argument for incorporating screening for hereditary CSGs (in particular for Lynch genes *MLH1* and *MSH2*) into population genetic testing [79]. Identification of specific mutations facilitates cascade-testing of other family members, and risk-reduction guidance, generally carried out through specialised familial cancer clinics. In hereditary CRC syndromes, risk reduction may include colonoscopic surveillance, often from a young age (for example 12-14 years in FAP), pharmacoprevention with aspirin in Lynch syndrome, or prophylactic surgery [65].

Lynch syndrome is relatively common, identified in 2-3% of unselected CRC cases and 16% of MMR-deficient cases, many of whom do not meet traditional testing criteria such as the Bethesda guidelines [161–163]. Screening for Lynch syndrome is now recommended in all sporadic CRCs [164]. The population prevalence of Lynch syndrome is estimated to be around 1:280 [165]. Just 10% of prevalent carriers are thought to have been identified, and even with testing of cancers, it could take decades to identify all of these families [79]. Clinical and financial arguments for population screening for CSGs are currently being evaluated, with studies in breast cancer leading the way [166, 167].

1.4 Predicting colorectal cancer risk

Risk prediction models in healthcare aim to predict the likelihood or absolute risk of an individual developing a given condition. Uses of these models include guiding prevention measures, such as screening or chemoprevention, directing therapeutic decision making, or guiding monitoring and follow-up strategies [168]. Thousands

of risk prediction models have been developed across many phenotypes, yet few are used in practice, in part because of a lack of external validation studies [169]. Royston et al. [170] note that to be useful clinically, a risk model should be reliable, accurate (that is, they should have good discriminative ability and be well calibrated), be generalisable, and be demonstrated to be clinically effective. Clear guidelines now exist for the development and reporting of clinical risk prediction models, and of PRS specifically [168, 171].

1.4.1 Non-genetic risk models

The earliest models for CRC incorporated non-genetic risk factors. Usher-Smith et al. [172] and Smith et al. [173] undertook systematic reviews of non-genetic risk prediction models, and externally validated these in UK Biobank (UKB) and The European Prospective Investigation into Cancer and Nutrition (EPIC) cohorts respectively. Thirteen studies were included, the majority developed in Europe or the US, summarised in Table 1.2 [174–185]. Many of these contained well-recognised epidemiological risk factors discussed above.

Some studies performed very poorly in external validation (this is, no better than chance), while the top-performing models were the QCancer-10 (Colorectal Cancer) models for males and females, with a C-statistic of 0.70 in men and 0.66 in women [172, 182]. The lower performance in women is echoed in several of the other studies.

Table 1.2: Non-genetic risk models validated in UK Biobank (based on Tables in Usher-Smith et al. [172] and Smith et al. [173]). Smith et al. also included models which are not included here as they were for right sided cancer in men and rectal cancer in women only. In addition the Steffen et al. [184] paper also included models for colon and rectal cancer separately; only the CRC model is included here.

Study	Population	Predictors	UKB UsherSmith	UKB Smith	EPIC Smith	Reference
Colditz (M)	USA	Education, BMI, FHx, screening, IBD	0.68 (0.66-0.70)	0.68 (0.66-0.70)	0.67 (0.64-0.70)	Colditz, Cancer Causes Control, 2000
Colditz (F)	USA	Education, BMI, FHx, screening, IBD, oral contraceptives, HRT	0.5 (0.48-0.53)	0.63 (0.60-0.65)	0.65 (0.62-0.69)	Colditz, Cancer Causes Control, 2000
Driver	USA	Age, BMI, smoking, alcohol	0.67 (0.66-0.69)	0.68 (0.67-0.69)	0.67 (0.64-0.70)	Driver, Am J Med, 2007
Freedman (M)	USA	Ethnicity, BMI, FHx, colonoscopy	0.64 (0.61-0.66)	0.6 (0.58-0.62)	0.61 (0.59-0.63)	Freedman, J Clin Oncol, 2009
Freedman (F)	USA	Ethnicity, BMI, FHx, colonoscopy	0.59 (0.56-0.61)	0.58 (0.56-0.61)	0.58 (0.56-0.60)	Freedman, J Clin Oncol, 2009
Guesmi	Tunisia	Age, meat, milk	0.65 (0.63-0.66)	NA	NA	Guesmi, Tunisie Medicale, 2010
Hippisley-Cox & Coupland (M)	UK	Age, Townsend, ethnicity, BMI, smoking, alcohol, FHx, UC, polyps,diabetes, lung cancer, blood cancer, oral cancer	0.7 (0.69-0.72)	NA	NA	Hippisley-Cox, BMJ Open, 2015
Hippisley-Cox & Coupland (F)	UK	Age, ethnicity, smoking, alcohol, FHx, UC, polyps, diabetes, breast cancer, ovarian cancer, uterine cancer, cervical cancer	0.66 (0.64-0.68)	NA	NA	Hippisley-Cox, BMJ Open, 2015
Johnson (M)	Worldwide	BMI, physical activity, smoking, alcohol, HRT, aspirin, processed meat, red meat, fruit, FHx, IBD	0.49 (0.47-0.51)	NA	NA	Johnson, Cancer Causes Control, 2013
Johnson (F)	Worldwide	BMI, physical activity, smoking, alcohol, HRT, aspirin, processed meat, red meat, fruit, FHx, IBD	0.5 (0.48-0.52)	NA	NA	Johnson, Cancer Causes Control, 2013
Ma (M, Cox Regression)	Japan	Age, BMI, smoking, alcohol, physical activity	0.69 (0.68-0.71)	0.69 (0.68-0.71)	0.68 (0.65-0.70)	Ma, Cancer Epidemiology, 2010
Ma (Simple score)	Japan	Age, BMI, smoking, alcohol, physical activity	0.68 (0.67-0.70)	NA	NA	Ma, Cancer Epidemiology, 2010
Steffen	European	Age, sex, BMI, smoking, alcohol, diabetes, screening	NA	0.68 (0.67-0.69)	0.68 (0.65-0.71)	Steffen, PLOS ONE, 2013
Taylor	USA	Age, FHx	NA	0.67 (0.66-0.68)	0.67 (0.65-0.69)	Taylor, Genet Med, 2011
Tao (M)	Germany	Age, sex, FHx, smoking, alcohol, NSAIDs, colonoscopy, polyps, red meat	0.69 (0.67-0.70)	NA	NA	Tao, Clin Gastroenterol Hepatol, 2014
Tao (F)	Germany	Age, sex, FHx, smoking, alcohol, NSAIDs, colonoscopy, polyps, red meat	0.63 (0.61-0.65)	NA	NA	Tao, Clin Gastroenterol Hepatol, 2014
Wei (M)	China	BMI, smoking, alcohol, FHx	0.51 (0.49-0.53)	NA	NA	Wei, World J Gastroenterol, 2009
Wei (F)	China	BMI, smoking, alcohol, FHx	0.49 (0.47-0.51)	NA	NA	Wei, World J Gastroenterol, 2009
Wells (M)	USA	Age, ethnicity, smoking, alcohol, education, BMI, FHx, diabetes, aspirin, multivitamin, red meat, physical activity	0.61 (0.59-0.64)	0.69 (0.67-0.71)	0.7 (0.67-0.73)	Wells, J Am Board Fam Med, 2014
Wells (F)	USA	Age, ethnicity, smoking, alcohol, education, BMI, FHx, diabetes, NSAID, multivitamin, oestrogen	0.64 (0.62-0.66)	0.62 (0.60-0.64)	0.67 (0.65-0.70)	Wells, J Am Board Fam Med, 2014

UKB - UK Biobank; M - male; F - female BMI - body mass index; FHx - family history; IBD - inflammatory bowel disease;
HRT - hormone replacement threapy; UC - ulcerative colitis; NSAIDs - non-steroidal anti-inflammatory drugs; NA - Not applicable

These external validation studies were the most up to date in the literature at the time I began model development. Since then, several additional prediction models have been published. Guo et al. [186] developed a simple scoring system to predict 10-year CRC risk in a large Chinese cohort, including age, alcohol intake, diabetes, occupational sitting time, and waist circumference, with an internally validated an area under receiver operating characteristic curve (AUROC) of 0.66. A model for women developed in the US Women's Health Initiative study including age, ethnicity, family history, smoking history, weight, waist circumference, aspirin use, and a measure of general health, (with variables chosen a priori) achieved an AUROC of 0.67 for 5 year risk in split sample geographic validation [187].

Some studies focused on modifiable lifestyle factors with the aim of facilitating behaviour change [188]. Usher-Smith et al. [189] developed a lifestyle model including physical activity, BMI, alcohol, smoking, red and processed meat, vegetable and fruit intake, with effect sizes derived from published meta-analyses. Validation of their model in the EPIC-Norfolk cohort demonstrated an AUROC of 0.66 and 0.68 in men and women in predicting 10-year risk. Aleksandrova et al. [188] developed a lifestyle-based time-to-event model, the LiFeCRC score, utilising the EPIC cohort of 19-70 year olds, again predicting 10-year CRC risk, and with split-sample validation in a subset of the dataset. The study has the advantage of a large and highly phenotyped dataset, and is one of few models to include height, which is a strong predictor (and is easily quantified). This model achieved a Harrell's C-index (a measure of the ability to correctly discriminated cases from controls, see Section 2.8.8) of 0.71 overall. They also examined performance across age groups and found their model to have strongest performance in <45 year olds, concluding that this may be an optimal time for intervention.

Several studies have looked specifically at risk prediction during bowel screening, largely evaluating a cross-sectional approach. A 2018 systematic review and meta-analysis with 17 risk scores for individuals undergoing screening colonoscopy, including at least age, sex, and other risk factors or laboratory tests, reinforced that most had weak discrimination, with a maximum C-statistic of 0.70 in the

meta-analysis [190], and maximum AUROCs of 0.61 and 0.65 on external validation [191]. Kaminski et al. [192] developed a logistic regression model to predict advanced neoplasia in a large screening cohort of 40-66 year old patients, including age, sex, family history, cigarette smoking and BMI, but this again discriminated cases from controls relatively poorly, with a C-statistic of 0.62.

Imperiale et al. [193] evaluated a large number of sociodemographic, lifestyle and medical predictors in a model for advanced neoplasia in a cohort of 50-80 year olds undergoing first round of screening colonoscopy at various centres in Indiana. In random split sample validation, they obtained a relatively high C-statistic of 0.78. However several aspects of the study are methodologically flawed. Categorisation of most predictors leads to loss of information, and evaluation of a number of potential cut-points is likely to lead to over-fitting and over-estimation of performance. In addition, the high number of parameters considered relative to the number of cases (with an event-per-parameter <10), and use of random split-sample validation, mean the study is likely underpowered to reach firm conclusions about performance [194]. This level of performance may well therefore not be achieved in other cohorts.

In the context of a two-step screening process, the addition of demographic, medical or lifestyle data to the faecal test alone can improve test performance, particularly for the detection of advanced adenomas. Simply adding age and gender results in small improvements in advanced neoplasia detection [195], whilst the addition of a family history questionnaire to FIT testing failed to improve detection rates [196]. A risk prediction model including age, family history, and total calcium intake, alongside FIT, improved discrimination for advanced neoplasia compared to FIT alone (AUROC 0.76 compared to 0.69), and increased sensitivity [197]. Adding standard screening data (age, sex, screening history, and index of multiple deprivation score) to FIT (at a threshold of 160µg/g) increased the AUROC from 0.63 to 0.66, and increased sensitivity for advanced neoplasia, driven entirely by higher detection rates of advanced adenomas [198]. In an extension of this study, Cooper et al. [199] used GP data to develop a risk score for CRC diagnosis within a 2 year screening round following gFOBT in a UK-based screening cohort. The

model including gFOBT result with of age, sex, family history of gastrointestinal cancer, alcohol consumption, smoking status, IBS diagnosis, previous negative FOBT tests, and whether they had had a blood test via the GP, resulted in a C-statistic of 0.86 on internal validation. Compared to gFOBT only, the model increased sensitivity from 53.9% to 58.8%.

Many models are methodologically weak based on TRIPOD guidelines, with small sample sizes [200, 201], and extensive categorisation of data [187, 192, 202], and with few exceptions they have generally failed to improve on the performance of earlier models. I have discussed discrimination here, but notably other measures such as explained variation and calibration are important, and often poorly reported [168].

At least one clinical trial has assessed a risk-based approach to screening. The Asia-Pacific Colorectal Screening (APCS) Score [203], which incorporates age, sex, first-degree family history, BMI, and smoking, has been evaluated in the TARGET-C trial. This compared one-time colonoscopy, annual FIT (with threshold of 4µg/g), and annual risk-adapted screening in almost 20,000 Chinese screening participants. The risk-adapted approach divided participants into low or high risk groups, who subsequently underwent FIT or colonoscopy respectively. Whilst there was no significant difference in CRC detection rates among the three groups in intention-to-screen analysis, risk-adapted screening overall significantly increased detection rates for advanced neoplasia compared to FIT (OR 1.49, 95% CI 1.13–1.97), though DR was highest in the colonoscopy group. However, in the high risk group the detection rate for advanced neoplasia was greater than for the colonoscopy group (OR: 1.72, 95% CI 1.20–2.48) [204].

1.4.2 Polygenic risk scores

With the discovery of an increasing number of risk SNPs for common diseases through GWAS came the recognition that these might be used to assess genetic predisposition. Over the last 20 years, a huge amount of work has been devoted to the development of polygenic risk scores (PRS, also called genomic risk scores, GRS). Typically, PRS summarise individual risk based on the summation of risk allele

counts, weighted by their effect sizes. Whilst the effect of each SNP can be very small, the summary score can be strongly associated with the phenotype in question.

Initially, PRS included only GWAS-significant SNPs, but a growing body of evidence suggests that better predictive performance can be achieved beyond those SNPs reaching these thresholds, including the many common, small effect variants not yet detected by GWAS [171, 205, 206]. The development of methodologies evaluating which SNPs ought to be included, and how to weigh them to obtain optimal performance, is a highly active field of research [207–212]. Recently there has been a push towards clinical implementation, and the emergence of direct-to-consumer PRS testing, with which individuals can obtain genetic risk estimates outside of healthcare settings. In light of this, efforts are now being made to standardise PRS conduct and reporting, to improve a currently heterogeneous evidence base [171, 213].

The majority of CRC PRS studies to date have included genome-wide significant SNPs only, increasing in number as GWAS sample size increased, and are listed in Table 1.3 [157, 214–231]. These have largely been conducted in European populations, with several in East Asian populations. Many early CRC PRS were externally validated in UK Biobank by Saunders et al. [215], demonstrating improved performance as more SNPs were included. Notably, the predictive performance of PRS remains poor, with a maximum AUROC of 0.65 in the genome-wide LDpred model [228].

Table 1.3: Published polygenic risk scores for colorectal cancer. AUROCs are largely taken from the external validation study in UK Biobank by Saunders et al. [215], which were not adjusted for co-variates, and are presented for males and females. The Fritsche et al. [219] PRS has not been externally validated and shows derivation performance in UK Biobank (adjusted for age, sex, principal components, and cohort). The Thomas et al. [228] PRS were validated in the GERA cohort (US) and adjusted for age and sex.

Study	Derivation Population	SNPs	Base data	PRS method	AUROC (Males; Females)
Abe 2017	Japan	11	Published European and Asian GWAS	GWAS-sig, unweighted allele count	0.55 (0.54-0.56); 0.55 (0.53-0.56)
Dunlop 2013	North American, European, Australia	10	Published European GWAS	GWAS-sig, unweighted allele count	0.56 (0.55-0.57); 0.57 (0.55-0.59)
Frampton 2016	UK	37	Published European GWAS	GWAS-sig, weighted allele count	0.55 (0.53-0.56); 0.55 (0.54-0.57)
Fritsche 2020	UK (Biobank)	87	Huyghe 2019 summary statistics	P+T, weighted allele dosage	0.62 (0.60-0.63) (d)
Fritsche 2020	UK (Biobank)	27	Published European GWAS	P+T, weighted allele dosage	0.57 (0.55,0.58) (d)
Kachuri 2020		103	Huyghe 2019 summary statistics	GWAS-sig, weighted allele count	0.716
Hosono 2016	Japan	6	Published European GWAS	GWAS-sig, unweighted allele count	0.54 (0.53-0.55); 0.53 (0.52-0.55)
Hsu 2015	USA and Germany	31	Published European GWAS	GWAS-sig, weighted allele count	0.57 (0.55-0.58); 0.58 (0.57-0.6)
Huyghe 2019	European (91.7%) and East Asian (8.3%)	120	European and Asian GWAS meta-analysis	GWAS-sig, weighted allele count	0.64 (0.61-0.66); 0.62 (0.59-0.64)
Ibanez-Sanz 2017	Spain	21	Published European GWAS	GWAS-sig, unweighted allele count	0.55 (0.54-0.57); 0.56 (0.54-0.58)
Iwasaki 2017	Japan	6	Published European and Asian GWAS	GWAS-sig, weighted allele count	0.54 (0.53-0.55); 0.53 (0.52-0.55)
Jenkins 2016	Australia, Canada, USA	49	Published European GWAS	GWAS-sig, weighted allele count	0.57 (0.55-0.58); 0.57 (0.55-0.58)
Jeon 2018	Australia, Canada, Germany, Israel, and USA	63	Published European and Asian GWAS	GWAS-sig, weighted allele count	0.58 (0.57-0.6); 0.58 (0.57-0.6)
Li 2020		116	European GWAS meta-analysis	GWAS-sig, weighted allele count	0.60 (0.59-0.61)
Smith 2018	UK	41	Published mainly European GWAS	GWAS-sig, weighted allele count	0.56 (0.55-0.58); 0.57 (0.56-0.59)
Thomas 2020	USA	140	Huyghe 2019 European data	GWAS-sig, weighted allele dosage	0.629 (0.613–0.645)
Thomas 2020	USA	10,000	Huyghe 2019 European data	Penalised ridge regression	0.633 (0.617–0.648)
Thomas 2020	USA	1,180,765	Huyghe 2019 European data	LDpred (Bayesian)	0.654 (0.639–0.669)
Wang 2013	Taiwan	16	Published Asian GWAS	GWAS-sig, logistic regression	0.51 (0.49-0.52); 0.50 (0.48-0.52)
Xin 2018	China	14	Published European or Asian GWAS	GWAS-sig, unweighted allele count	0.54 (0.53-0.56); 0.53 (0.52-0.55)
Yarnall 2013	UK	14	Published European GWAS	GWAS-sig, weighted allele count	0.56 (0.54-0.57); 0.55 (0.54-0.57)

GWAS-sig - genome-wide association study significant SNPs; P+T - Pruning and thresholding; d - derivation study

Increasing sample sizes for GWAS will improve the accuracy of PRS [116, 232]. The small effect sizes of known GWAS variants, contrasting with larger estimates of heritability for many complex diseases, indicate that there are likely thousands of variants of small effect sizes which are as yet unidentified, and which may require very large GWAS sample sizes to detect [233].

However, there are limitations to the extent to which genotyped SNPs are able to account for genetic risk. As discussed above, the SNPs in a genome-wide tag-SNP chips are not usually the causative variants themselves, but are in LD with these variants (hence, tag-SNPs). These typically represent SNPs with relatively common alleles, and therefore will not tag SNPs with very rare allele frequencies. Variants with large effects on fitness will be driven to have a low minor allele frequency through natural selection [232]. It is estimated that with imputation, around 90% of low-frequency variants are covered [206], however inevitably some rare variants with a significant effect on risk will not be captured in a GWAS discovery population, and PRS will be unable to capture all genetic risk.

Despite assumptions that adding rarer variants with potentially large effects to prediction models would improve performance, this is not necessarily the case. Although rare variants are now more readily identifiable with large sequencing projects [234], and could improve variation accounted for closer towards total heritability [232], modelling suggests the addition of such variants to a polygenic model is likely to have minimal impact on PRS performance at a population level [235].

Despite the profusion of PRS studies, few are approaching clinical implementation [171]. Recent PRS have been shown in a number of phenotypes to perform sufficiently to identify individuals with risk equivalent to those with monogenic conditions, or at a level equivalent to an individual with a family history which would warrant enhanced screening [228, 236].

1.4.3 Integrated risk models

As discussed above, given the a priori limitations of the ability of PRS to predict genetic risk, and the significant non-heritable portion of variation in overall risk,

it would seem logical that utilising both genetic and non-genetic factors in risk prediction would give superior predictive performance. In particular, while genetic risk is unchanged throughout an individual's lifetime, for age-associated conditions such as CRC, age-specific predictions are important.

Integrated models combine both genetic and environmental risk predictors [171]. A number of studies have evaluated integrated models for a range of phenotypes, and generally found combined models to be superior to PRS alone (Table 1.4) [215, 217, 218, 220, 223–227, 231, 237].

Notably, age and sex show by far the greatest influence - many PRS studies adjust performance for age and sex, and realistically it seems unlikely that PRS would be implemented without integrating these simple predictors. In a further study by Kachuri et al. [238], the C-statistic for their 116-SNP age- and sex-adjusted PRS was identical to that of the model including all risk factors (family history, smoking, alcohol, diet, physical activity, body fatness, colonoscopy history). Adding the PRS to the non-genetic risk model (including age and sex) increased the C-statistic by 0.03, suggesting that in this study the risk factors beyond age and sex had little influence. The incremental benefit of adding PRS in this multi-cancer study was greater for cancers with fewer environmental risk predictors, such as testicular and thyroid cancers [238].

Table 1.4: Integrated risk models for colorectal cancer from the external validation study in UK Biobank by Saunders et al (2020), showing AUROC for PRS alone (not adjusted for co-variates) compared to models including PRS and non-genetic risk factors (NGRF). Reprinted from Cancer Prevention Research, 2020, 13(6), 509-520, Saunders et al., External Validation of Risk Prediction Models Incorporating Common Genetic Variants for Incident Colorectal Cancer Using UK Biobank, with permission from AACR.

			Males		Females	
Study	PRS SNPs	Non-genetic predictors	PRS	PRS+NGRF	PRS	PRS+NGRF
Abe 2017	11	Age, sex, FHx, BMI, smoking, referral pattern, alcohol, exercise, dietary folate	0.55 (0.54-0.56)	0.71 (0.69-0.74)	0.55 (0.53-0.56)	0.67 (0.64-0.7)
Dunlop 2013	10	Age, sex, family history	0.56 (0.55-0.57)	0.67 (0.66-0.69)	0.57 (0.55-0.59)	0.64 (0.62-0.66)
Hosono 2016	6	Age, FHx, BMI, smoking, referral pattern, alcohol, exercise, dietary folate	0.54 (0.53-0.55)	0.7 (0.69-0.72)	0.53 (0.52-0.55)	0.66 (0.64-0.67)
Ibanez-Sanz 2017	21	FHx, BMI, alcohol, exercise, red meat, vegetable intake, NSAIDs and aspirin	0.55 (0.54-0.57)	0.58 (0.56-0.59)	0.56 (0.54-0.58)	0.53 (0.52-0.55)
Iwasaki 2017	6	Age,FHx, BMI, alcohol	0.54 (0.53-0.55)	0.62 (0.6-0.63)	0.53 (0.52-0.55)	0.56 (0.54-0.58)
Jenkins 2016	49	FHx	0.57 (0.55-0.58)	0.58 (0.56-0.59)	0.57 (0.55-0.58)	0.56 (0.54-0.58)
Jeon 2018	63	Men and women separately. BMI, smoking, height, education, diabetes, alcohol, aspirin, NSAIDs, smoking, fibre, calcium, processed meat, red meat, fruit, vegetables, total energy, physical activity, HRT	0.58 (0.57-0.6)	0.6 (0.57-0.62)	0.58 (0.57-0.6)	0.59 (0.56-0.63)
Smith 2018	42	Age, FHx, BMI, smoking, diabetes, multivitamins, education, alcohol, physical activity, NSAIDs, red meat, smoking, oestrogen use	0.56 (0.55-0.58)	0.7 (0.68-0.71)	0.57 (0.56-0.59)	0.65 (0.63-0.66)
Yarnall 2013	14	FHx, BMI, alcohol, fibre, red meat, physcial activity	0.56 (0.54-0.57)	0.59 (0.57-0.6)	0.55 (0.54-0.57)	0.54 (0.53-0.56)

1.5 Conclusion

In this chapter I have summarised the background to CRC and its prevention through screening. I have also discussed our current understanding of the genetic and non-genetic aetiologies of CRC, and our ability to predict CRC risk using risk prediction models. In the next chapter I will summarise the main methodologies used in this thesis, prior to presenting the results of this research in subsequent chapters.

2
Methods

This chapter presents an overview of the datasets and methodologies used in this thesis, and the rationale for choice of these.

2.1 Ethics

Collection of samples and clinical and pathological data from patients was completed with written informed consent. All studies were approved by relevant ethical review boards at respective study centres. UK National Cancer Research Network Multi-Research Ethics Committee approvals by study are as follows: CoRGI - 17/SC/0079, VICTOR/QUASAR2 - ORECB/05/Q1605/66, SCOT - 07/S0703/136, NSCCG - 02/0/097, SOCCS - 01/0/50, Generation Scotland-Scottish Family Health Study - 05/S1401/89, Lothian Birth Cohorts 1921 - LREC/1998/4/183, Lothian Birth Cohorts 1936 - 2003/2/29. The CORSA study was approved by the ethical review committee of Medical University of Vienna (MUW, EK Nr. 703/2010) and the Ethikkommission Burgenland (KRAGES, 33/2010). Finnish studies were approved by the Finnish National Supervisory Authority for Welfare and Health, National Institute for Health and Welfare (THL/151/5.05.00/2017), the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS/408/13/03/03/09). The

PoBI study was approved by Leeds (West) Ethics Research Ethics Committee (05/Q1205/35).

The UK Biobank study has ethical approval from the North West Multi-centre Research Ethics Committee (06/MRE09/65). The work presented here was carried out under Biobank application number 8508.

2.2 GWAS datasets

For the genome-wide association study and subsequent meta-analysis undertaken in Chapter 3, 15 case-control cohorts were used. Five primary GWAS were analysed (SCOT, NSCCG-OncoArray, SOCCS/GS, SOCCS/LBC, UK Biobank), and meta-analysed with 10 previously published GWAS. I performed sample processing (see Section 2.6), genotyping and QC of the SCOT study (described below and in Chapter 3), and evaluated the use of the People of the British Isles (PoBI) study as a control dataset; ultimately The Heinz Nixdorf Recall Study provided controls for the SCOT cohort. Each of these studies is described below.

2.2.1 SCOT

The Short Course Oncology Treatment (SCOT) trial [239] was an international clinical trial of 3 versus 6 months of adjuvant oxaliplatin chemotherapy treatment for individuals with high risk stage II or stage III CRC, from which 3076 cases were included in GWAS. In the final GWAS, 4349 cancer-free control individuals were identified from The Heinz Nixdorf Recall study [240], a population-based cohort study of risk stratification for cardiac events. Both studies were genotyped on Illumina's Global Screening Array.

2.2.2 PoBI

The People of the British Isles (PoBI) study [241, 242] was designed to provide a UK control population, and further our understanding of fine-scale genetic population structure by collecting information for a cohort of individuals living in rural locations in the UK, with all four grandparents living in approximately the same location.

Samples were collected from 4371 individuals, with 2886 genotyped as part of the WTCCC2 study on the Illumina Human 1.2M-Duo chip.

2.2.3 NSCCG-OncoArray

Cases for this GWAS were obtained from two studies: the National Study of Colorectal Cancer Genetics (NSCCG, n = 6240) [243], and the Colorectal Tumour Gene Identification (CoRGI, n = 1041) consortium [244]. This dataset is enriched for genetic predisposition to CRC: cases had CRC under 58 years of age, or had family history of CRC in at least one first-degree relative. Controls for this study were from two studies: 3031 men from the PRACTICAL Consortium recruited to the UK Genetic Prostate Cancer Study (UKGPCS), and the 4488 women recruited to studies from the Breast Cancer Association Consortium (BCAC), all of whom were cancer free [245, 246]. All samples were genotyped on Illumina's OncoArray chip.

2.2.4 SOCCS/GS

Cases for this GWAS were obtained from the Study of Colorectal Cancer in Scotland (SOCCS, n = 4772), a population-based series of incident CRC cases diagnosed under 80 years [247, 248]. Cancer-free population-based controls were taken from two studies: 2221 from SOCCS, and 9937 from Generation Scotland-Scottish Family Health Study [249]. Genotyping was performed on custom Illumina Infinium arrays.

2.2.5 SOCCS/LBC

An additional 1037 cases from the SOCCS study (n = 1037) were analysed in the SOCCS/LBC GWAS, with cancer-free population based controls from the Lothian Birth Cohorts from 1921 and 1936 [250]. Genotyping was performed on custom Illumina Infinium arrays.

2.2.6 CCFR1 and CCFR2

The CCFR1 GWAS includes 1290 familial CRC cases and 1055 controls taken from the Colon Cancer Family Registry (CCFR) [251]. This is a US consortia

recruiting from 6 different sites. Most cases are diagnosed at a young age or have a positive family history, with unrelated controls matched on age and sex. The CCFR2 GWAS includes an additional 796 CCFR cases. Controls were from the Cancer Genetic Markers of Susceptibility (CGEMS) studies investigating risk of prostate and breast cancer (n = 2236) [252]. Both datasets were genotyped on Illumina Omni-express, Hap1M, or Hap1M-Duo arrays.

2.2.7 COIN

COIN cases (n = 2244) were taken from two clinical trials, COIN and COIN-B. The COIN phase 3 trial compared continuous versus intermittent chemotherapy in patients with advanced colorectal cancer, while COIN-B was a phase 2 trial evaluating intermittent chemotherapy with cetuximab [253, 254]. Cases were genotyped on Affymetrix Axiom arrays. Controls for this GWAS (n = 2162) were from UK Blood Service Control Group, which formed part of the WTCCC2 study and were genotyped on the Illumina Human 1.2M-Duo chip [255].

2.2.8 CORSA

The CORSA (COloRectal cancer Study of Austria) study includes 978 sporadic CRC cases and 855 controls with no lesions on colonoscopy recruited from four different Austrian hospitals [256]. This study was genotyped on the Affymetrix Axiom Genome-Wide CEU 1 array.

2.2.9 Croatia

The Croatia GWAS dataset includes 764 CRC cases and 460 population-based controls [257]. Samples were genotyped on the Illumina OmniExpressExome BeadChip 8v1.1 or 8v1.3 arrays.

2.2.10 DACHS

DACHS (Darmkrebs: Chancen der Verhütung durch Screening) is a population-based case-control study based in South-West Germany, including individuals over

30 years of age with first diagnoses of CRC, with controls randomly selected from population registries [258]. From this, 1195 cases and 700 controls were included in GWAS. Genotyping was performed on the Illumina OncoArray.

2.2.11 FIN

The FIN GWAS included 1172 Finnish cases and 8266 controls gathered through the Helsinki Birth Cohort, Finnish Twin Cohort, FINRISK, and Health 2000 studies [259]. Cases were genotyped on Illumina HumanOmni 2.5M8v1 arrays; controls on Illumina HumanHap 610k or 670k arrays.

2.2.12 UK1

This GWAS includes 940 CRC cases and 965 controls from the CoRGI consortium [244, 248]. All cases had a first degree relative with CRC and either: CRC diagnosed under age 75; a large or aggressive colorectal adenoma, or three or more adenomas, diagnosed under age 75; or any colorectal adenomas under age 45. Controls were partners or spouses without personal or first- or second-degree family history of CRC. Samples were genotyped on Illumina Hap240S, Hap300, Hap370, Hap550 or Omni2.5M arrays.

2.2.13 Scotland1

The Scotland1 GWAS includes 1012 early-onset CRC cases (diagnosed at 55 years of under) and 1012 age, sex and geographically matched population controls [248]. Hereditary cancer syndromes were excluded (polyposis syndrome, Lynch syndrome, or bi-allelic *MUTYH* mutation carriers). Genotyping was performed on Illumina Hap240S, Hap300, Hap370, Hap550 or Omni2.5M arrays.

2.2.14 VQ58

Cases from the VQ58 study were from two UK trials of adjuvant chemotherapy-VICTOR [260] and QUASAR2 [261] (n = 1800). Controls were taken from the Wellcome Trust Case Control Consortium 2 (WTCCC2) 1958 birth cohort (n =

2690) [262]. VICTOR and QUASAR 2 were genotyped on Illumina Hap240S, Hap300, Hap370, Hap550 or Omni2.5M arrays; the 1958 birth cohort was genotyped on the Illumina Human 1.2M-Duo Custom array.

2.3 Whole genome sequencing datasets

2.3.1 WGS500

Sixty-five genomes from CRC patients were sequenced as part of the WGS500 programme, an early large-scale sequencing project which aimed to sequence 500 individuals with diverse genetic disorders [263]. Examination of 15 of these genomes led to the discovery of *POLE* and *POLD1* as cancer susceptibility genes [86].

2.3.2 CGI-196

The 196 individuals sequenced in the Complete Genomics project were all CRC cases identified within CoRGI, who had been diagnosed under the age of 55, and had at least one first-degree relative with CRC [264]. Complete Genomics software was used to sequence and call these genomes [265].

2.3.3 Illumina-215

The individuals sequenced as part of the Illumina-215 project were ascertained through CoRGI. All had at least 5 adenomas, or CRC at <65 years of age, or were first-degree relatives of individuals with these phenotypes.

2.4 The UK Biobank dataset

UK Biobank is a well validated data source, described in detail in Fry et al. [266] and Bycroft et al. [267]. Just over 500,000 participants were recruited through 20 centres across the UK (2006 and 2010), representing 5.5% of invitees. Baseline demographics, medical, lifestyle and physical data, and blood samples, were collected at recruitment. Follow-up through linked hospital and registry data is ongoing. A detailed description of the dataset is presented in Chapter 4.

2.5 QCancer-10 (Colorectal Cancer) risk model

Given the existence of many models for CRC, there is a strong argument for updating an existing model rather than generating a new combined model. As discussed in Section 1.4.1, at the time of beginning my modelling work QCancer-10 (Colorectal Cancer) was the top-performing model in external validation in UKB, and I therefore used this as the basis for my combined model.

The QCancer-10 models [182] are a set of risk prediction algorithms for 11 common cancers, including colorectal cancer, which use Cox regression models to generate 10-year absolute risks from routinely available primary care data. They were developed with the aim of identifying high risk individuals in the population to whom prevention could be targeted. The dataset used in development was the QResearch database, which contains pseudonymised data on patients from 750 primary care practices using the Egton Medical Information Systems (EMIS) computer system across the UK, linked to cancer and mortality registries and hospitals admissions data [268]. The model for CRC has now been extended to 15 years of follow-up.

Patients aged 25-84, registered between January 1st 1998 and September 30th 2013 with English practices which had been using EMIS for at least a year were included. Individuals with no postcode-based Townsend score (a measure of population deprivation which takes into account unemployment, non-car ownership, non-home ownership, and household overcrowding, recorded for each individuals postcode [269]), and those with a prior history of the cancer in question were excluded. 565 practices (4.96 million patients) were included in the model derivation cohort, with 188 practices (1.64 million patients) in the randomly allocated validation cohort. Cases were identified through coding in any of general practice (GP) data, cancer registry, mortality registry or hospital data records.

Separate models were developed for men and women for each cancer. The potential predictors included in the model (listed in Table 2.1) were identified through existing literature or listed on the Cancer Research UK website, and were those readily found in GP records or known by patients. Many of these

Table 2.1: QCancer-10 (Colorectal Cancer) variables and additional candidate predictors considered in model development. F: included in women's model only; M: included in men's model only; BMI: body mass index; HRT: hormone replacement therapy

QCancer-10 variables	Additional variables considered
Age	Type 1 diabetes
Townsend deprivation score (M)	Manic depression or Schizophrenia
BMI (M)	Antipsychotics at baseline
Ethnic Group	HRT at baseline (F)
Smoking Status	Oral contraceptives at baseline (F)
Alcohol intake	Prior gastro-oesophageal cancer
Family history of CRC	Prior pancreatic cancer
Ulcerative colitis	Prior prostate cancer (M)
Colonic polyps	Prior renal tract cancer
Type 2 diabetes	Prior brain cancer
Prior breast cancer (F)	-
Prior uterine cancer (F)	-
Prior ovarian cancer (F)	-
Prior cervical cancer (F)	-
Prior lung cancer (M)	-
Prior blood cancer (M)	-
Prior oral cancer (M)	-

are well established as likely causes of CRC in the epidemiological literature, as discussed in Section 1.4.1. Full model specification can be found at `https://qcancer.org/15yr/colorectal/` [270]. Coding of these variables in this thesis is described in more detail in Section 4.2.3.

2.6 Laboratory methods

2.6.1 DNA extraction and quantification

For the SCOT study, patient blood was collected and stored at -20°C in EDTA tubes. I used Maxwell®16 Blood Purification Kits (Promega UK Ltd) to extract DNA from samples thawed at room temperature. I extracted DNA in 16-sample batches using the Maxwell®16 AS2000 machine. This process followed the kit instructions: I added 230μL of elution buffer and 420μL whole blood to each of the cartridges, and DNA was extracted and quantified, and stored at 4°C. I quantified DNA concentrations using picogreen quantification.

2.7 Bioinformatics

2.7.1 GWAS quality control

Inaccurate calling in GWAS studies can result in spurious associations, and given the many thousands (and following imputation, many millions) of markers tested, even low error rates can result in significant numbers of false associations. Removal of individuals and markers with higher error rates is therefore required. My QC procedure followed the standard principles set out in Anderson et al. [271], using PLINK v1.9 [272], performing per-person followed by per-SNP QC as recommended to retain as many markers as possible. A more detailed description is given in Section 3.2.1. Full details of the quality control and exclusions for the final GWAS meta-analysis can be found in Law et al. [115] (Supplementary Table 3).

2.7.2 Phasing and Imputation

Genotyping arrays provide data on hundreds of thousands of SNPs across the genome quickly and cheaply. These tag-SNPs act as proxies for nearby variants correlated through linkage dysequilibrium (see Section 2.8.1 below), which allows imputation of untyped SNPs based on known patterns of LD. Imputation allows interrogation of a far greater proportion of the genome, and allows data typed on different chips to be combined for analysis, which was required in both my GWAS and linkage studies.

Imputation uses comparison of haplotypes in genotyped and reference data (of similar ancestral background) to derive the probabilities of a given genotype at imputed loci. The genotyped data is first phased with a haplotype reference panel (using a Markov chain Monte Carlo algorithm), and then the probability of a genotype at a given non-genotyped locus is calculated, given the data at observed loci in LD [273, 274]. Imputed genotypes are recorded as probabilities, and the quality of imputation at a given locus is calculated as an INFO score. Typically, only loci reaching a given quality threshold are included in subsequent analysis.

Prior to imputation I used a pre-imputation checking script provided by Dr Will Rayner which confirms strand alignment and checks for errors [275]. Phasing

was performed using SHAPEITv2.790 [276] against a combined reference panel of UK10K and 1000 Genomes Project Phase 1 reference panel [277, 278]. For GWAS studies, imputation of missing genotypes was performed against the combined reference panel using IMPUTE2 [274]. In linkage analysis datasets, phasing and imputation was to the 1000 Genomes Project Phase 3 reference panel.

2.7.3 Linkage analysis

Genetic linkage is the increased likelihood of two or more loci to be transmitted together in meiosis due to physical proximity. Recombination during meiosis results in new combinations, or phases, of loci in the offspring. The proportion of recombinations seen between two loci in offspring is known as the recombination fraction. With two genetically distant loci, the recombination fraction is about 0.5, i.e. the loci assort independently. On the other hand, very proximal loci have a recombination fraction of near zero, as there is little chance of recombination between them.

Linkage analysis uses statistical methods to test for fewer than expected recombinations between genetic markers, and thus map the loci to regions of the genome, and test for segregation of such loci with traits of interest. These methods are broadly parametric (i.e. model-based) or non-parametric (model-free).

Parametric linkage analysis assumes that the genetic models are known, and requires specification of allele frequency, mode of inheritance and penetrance. Non-parametric linkage analysis makes fewer assumptions.

In this thesis I used Merlin [279] which uses sparse binary trees to represent gene-flow patterns throughout a pedigree, to perform non-parametric linkage analysis. Detailed linkage methods are described in more detail in Section 3.4.1.

2.8 Statistical Methods

2.8.1 Linkage Dysequilibrium

Distant loci in the genome are inherited independently, as described by Mendel (discussed in Section 1.2.2), and so are said to be in linkage equilibrium. Loci which are on the same chromosome (i.e. syntenic) are potentially separated by recombination, and thus are said to be in linkage dysequilibrium. This can be measured using the coefficient of linkage dysequilibrium, D, or using Pearson's correlation co-efficient between two loci, r^2.

2.8.2 Identity by descent

Identity by descent (IBD) is the degree of shared recent ancestry for two individuals. To calculate IBD, identity by state (IBS), a metric of relatedness based on the proportion of shared alleles at genotyped SNPs, is calculated for each pair of individuals in the sample. The mean IBS of a population is dependent on the allele frequencies of the included SNPs, while the degree of allele sharing IBS is proportional to the level of relatedness. For duplicates and mono-zygotic twins, IBD = 1; for first-degree relatives IBD = 0.5; for second-degree relatives IBD = 0.25, and for third-degree relatives IBD = 0.125. In GWAS, a cut-off of >0.1875 is recommended to exclude second-degree and higher relatives, as genotyping error and LD will result in variation around these values [271].

I removed regions of high LD [280], and pruned based on LD, to remove any correlated SNPs from the analysis prior to calculating IBD using an in-house perl script.

2.8.3 PCA

Principal components analysis (PCA) can be used to define ancestral differences within a dataset, which permits outlying individuals to be removed from analysis. In addition, ancestry can be accounted for by including the principal components in the association analysis, which adjusts for population-stratification of allele frequencies

[281]. Individuals within a population do not mate randomly (tending to mate with a nearby partner), which creates genetic structure. Environmental risk is also structured within a population, and this can confound risk estimates associated with genetic variants based on geography, leading to false-positives and inflated associations within GWAS, and optimistic performance estimates in PRS [282].

The first principal component (PC) accounts for the largest amount of variation possible in a single measure, with subsequent PCs accounting for sequentially less variation. The PC model is constructed from known populations of diverse ancestry. The PCA model is then applied to the study datasets to calculate their PCs, and cluster them with known ancestral populations. I used EIGENSOFT v5.0.2 to calculate PCs for my GWAS datasets (Chapter 3). I then used the first 4 PCs to account for population structure in my PRS modelling (Chapter 5).

2.8.4 Pearson's χ^2 test

In data described using contingency tables, Pearson's χ^2 test described the strength of association between categories, comparing differences between observed frequencies and those expected by the χ^2 distribution. Pearson's χ^2 test assumes random sampling from a distribution, a sufficiently large sample size, and independence of observations.

I used Pearson's χ^2 test to compare allele frequencies in GWAS using PLINK [272], and to describe differences in distributions of characteristics of the UK Biobank cohort.

2.8.5 Logistic regression

In logistic regression, predictor variables are regressed on the logit-transformed outcome variable (in this work, the binary outcome is CRC status), giving the log-odds of CRC status.

In this work, I used logistic regression per-SNP in GWAS association testing using both SNPTEST v2.5.6 [273] and PLINK v1.9 [272] to run association analyses corrected for co-variables (Chapter 3). In Chapter 5, I used logistic regression

to evaluate PRS using LDpred2 and bigsnpr packages, and to assess predictive performance of the PRS in test and validation cohorts.

2.8.6 Cox proportional hazards regression

Cox proportional hazard models [283] are the most commonly used models for time-to-event data. Follow-up over time is possible for all individuals up to their last follow-up (at which point they are censored), assuming that the censoring is uninformative, i.e. the probability of censoring and of experiencing the outcome are unrelated.

The hazard rate, $h(t)$, gives the likelihood of the outcome in the next instant, given it has not already occurred. The baseline hazard function, $h_0(t)$ is the risk of the outcome over time for an individual with all model variables set to 0. For individual i,

$$h(t)_i = h_0(t)exp^{(\hat{\beta}_1\chi_{1i}+\hat{\beta}_2\chi_{2i}...)}$$

$(\hat{\beta}_1\chi_1 + \hat{\beta}_2\chi_2...)$ is the linear predictor (LP), which is usually (and in this thesis) centred to the mean. The exponent of the LP therefore gives the hazard ratio compared to an average risk individual. Cox models do not estimate the shape of the baseline hazard, and are therefore considered semi-parametric models ($h_0(t)$ is non-parametric, whilst the LP is parametric).

In order to predict absolute risks, a survival function $S(t)$, describing the likelihood of surviving to time t without experiencing the event, is estimated from the cumulative hazard function, $H(t)$, the summed hazard up to time point t.

$$S(t) = e^{H(t)}$$

Survival at time t for individual i is then

$$S(t)_i = S(t)^{exp(\hat{\beta}_1\chi_{1i}+\hat{\beta}_2\chi_{2i}...)}$$

Non-parametric estimates of $S(t)$ at set time-points can be used to calculate absolute risks. I estimated $S(t)$ from fitted cox models using R function survival::survfit, which calculates the survival for an individual with covariates set at the mean values of the model dataset.

Cox models assume that the hazards associated with predictors remain constant over time. I used two methods to test this assumption. Firstly, plots of $log(-log)$survival against log(survival) [284, (pp. 78), 285], in which the data split into two groups form parallel lines if the models has a constant hazard ratio, and statistical testing of proportional hazards, in per variable and overall analysis, using survival::cox.zph with inspection and plotting of Schoenfeld residuals.

2.8.7 Meta-analysis

Summary statistics from individual genome-wide association studies can be combined to give a weighted average for effect size, with increased power leveraged by the increased sample size facilitating the identification of rarer or smaller-effect variants.

Meta-analysis can used fixed-effects or random-effects approaches. The former assumes that the included studies examine the same population, with the same outcome definition and predictor variables, whilst the random effects model is generally used to combine heterogeneous studies. Fixed effects is commonly used in GWAS meta-analyses, where individual studies tend to be quite homogenous in their populations (corrected for ancestry), and the predictors (i.e. genotypes) and outcomes (diagnosis of CRC) tend to be quite standardised. In fixed effects meta-analysis, the studies are typically weighted by the inverse of the variance of the estimates, resulting in lesser contribution from smaller studies. In this thesis, I use the META package vsn 1.7 to conduct GWAS meta-analysis [286], described in more detail in Section 3.3.1.

2.8.8 Measures of discrimination

Discrimination describes how well a model differentiates between those with and without the outcome.

The C-statistic is a measure of accuracy of prediction using a classifier [287]. It describes the probability that a randomly selected pair of affected and unaffected individuals are correctly classified. This is directly related to Somers' D_{xy} statistic [288].

Harrell's c-index measures the fraction of pairs in which survival predictions and outcomes are concordant [289]. Pairs of individuals are drawn and their observed and predicted survival times are compared. If the patient with the higher prognostic score has survived the longest this pair are concordant. All possible pairs within the data are assessed, however censoring means that pairs in which both individuals survive to the end of follow-up cannot be assessed. Values near 0.5 are equivalent to chance, whilst those closer to 1 demonstrate that the model nearly always predicts correctly.

Royston and Sauerbrei's D statistic measures the prognostic separation of survival curves [290]. The prognostic indices across participants are ordered, and the corresponding rankits (the expected values of the order statistics of the normal distribution of the same size as the dataset) are calculated. The rankits are then scaled on a factor of $\kappa = \sqrt{8/\pi} \simeq 1.596$, and Cox regression performed on the scaled rankits.

Kaplan-Meier survival curves allow visual assessment of discrimination, and are used in this thesis to assess discrimination between risk groups. The further apart the curves for different risk groups, the better the discrimination. Comparison can be made between different datasets, though differences in the underlying populations must be considered, as residual confounding (i.e. the categorisation used does not fully account for the relationship between the outcome and the prognostic index) may result in performance appearing to be worse despite a well fitted model [291]. A moderate number of risk groups is optimal for this evaluation - too few provides insufficient information, whilst many groups can be unstable with little discrimination between adjacent categories. Unequal groups allow examination of individuals with more extreme risk together, retaining those around the middle of the risk distribution with largely similar prognosis in the same groups [291]. In this thesis I used the 16th, 50th and 84th centiles as cut-points to divide the dataset into 4 risk groups, which equates to approximately the mean ± 1 standard deviation [291, 292].

I also calculated scaled Brier scores, which assess overall model fit, reflecting both discrimination and calibration [194]. The Brier score is the squared difference between the observed outcomes and predicted probabilities; it has a range of 0-0.25, with lower scores indicating better performance. However, the maximum possible score is related to outcome prevalence, $mean(prevalence) * (1 - mean(prevalence))$, and so the score can be scaled to the maximum possible score as $1 - Brier/Brier_{max}$, permitting comparison between different populations.

2.8.9 Measures of variance explained

Variance explained describes the extent to which the model accounts for the variation in outcomes. R^2 is the most commonly used measure, of which there are several different statistics available for both logistic and Cox models.

In this thesis I used Nagelkerke's R^2 for logistic regression models, which is a version of the Cox-Snell R^2, scaled to the maximum, resulting in values from 0-1 (or 0-100%). Cox-Snell R^2 is calculated as

$$1 - \left(\frac{L_{null}}{L_{model}}\right)^{\frac{2}{N}}$$

where L is likelihood and N is the number of observations.

Royston and Sauerbrei's R^2_D is a measure on the log relative hazard scale derived from their D statistic [290]. It is converted from D by the formula

$$R^2_D = \frac{D^2/\kappa^2}{\sigma^2 + D^2/\kappa^2}$$

where $\sigma^2 = \pi^2/6 \simeq 1.596$ and $\kappa = \sqrt{8/\pi} \simeq 1.596$ as above.

2.8.10 Measures of model calibration

Calibration represents how closely a model's predictions correspond to the observed data. Calibration is tested in a validation dataset (in model training datasets calibration ought to be close to perfect, as the model is fitted to the data). In this thesis I use a number of methodologies to evaluate and present calibration.

Calibration plots visually represent calibration by plotting expected vs. observed risk or survival across risk groups [169]. In plots of calibration for PRS models in Chapter 5 I plotted risk across 10 groups by PRS. In calibration plots presented in Chapter 6 I plotted risk across 10 groups by LP. For all time-to-event models these were plotted at 5, 6, 7 and 8 years of followup. I overlayed loess smoothers to more clearly visualise the relationships [293].

Calibration of Cox models can be roughly assessed by comparing Kaplan-Meier plots in validation and derivation data: in a well calibrated model these should be similar [291]. I used this approach to assess calibration of the PRS Cox Models in Chapter 5.

The calibration slope reflects the average strength of predictor effects, which ideally is equal to 1. In internal validation, the calibration slope can be used to indicate the level of shrinkage required to fit the model to new individuals from the population [284, p272]. I used this approach to shrink my prediction models prior to external validation (described in more detail in Section 5.2.6).

Calibration-in-the-large (CITL) represents the difference between the average outcomes in an external validation dataset and the average predictions, and thus for a perfectly calibrated model this will equal 0. I present CITL for PRS tested in logistic regression models in Chapter 5. CITL cannot be tested in Cox models [284, p273].

A further simple representation of model calibration is the ratio of expected to observed outcomes (E/O). I used this measure in this thesis to assess calibration in subgroup analysis in Chapter 6, where case numbers were low.

2.8.11 Recalibration of prediction models

On application of a model to a new dataset, poor calibration may be improved by recalibration or updating of the model. In this thesis I used the most simple method of recalibration which involves re-estimation of the intercept for logistic regression models or the baseline hazard for Cox models (known as recalibration-in-the-large) [284, p.364]. This is achieved by fitting the model in the validation sample with the linear predictor as an offset term.

Several additional methods are available for model updating, with increasing complexity (and which increasingly distance the new model from the original model). Extensions to this include refitting of both the intercept and slope ('logistic calibration'). Methods of updating a model for a new population beyond this include re-estimation of parameters in the model in addition to the intercept, and 'model extension' which extends the model to include additional parameters in addition to recalibration or re-estimation [284, pp. 364-366]. These were not evaluated in this thesis, in part due to time constraints, and also because with more complex model revision, further external validation of the model would be required prior to any attempt at implementation.

2.9 Software

Table 2.2: Software and programmes used in the work presented in this thesis

	Software/Package	Reference
Genetics Software	PLINK v1.90b3	Purcell and Chang [272]
	SHAPEIT v2.790	Delaneau, Marchini, and Consortium [276]
	IMPUTE2 v2.3.0	Howie, Donnelly, and Marchini [274]
	GTOOL v.0.7.5	Freeman and Marchini [294]
	EIGENSOFT	Price et al. [281]
	QCTOOLS v2	Band and Marchini [295]
	SNPTEST v2.5.6	Marchini et al. [273]
	META v1.7	Liu et al. [286]
	PEDSTATS v0.6.10	Wigginton and Abecasis [296]
	Merlin v1.1.2	Abecasis et al. [279]
	BCFtools	Li et al. [297]
R packages	R v3.6.2	R Core Team [298]
	Tidyverse v1.3.0	Wickham et al. [299]
	Bigsnpr v1.5.2	Privé et al. [300]
	Survival v3.1-8	Therneau [301]
	RMS v5.1-4	Harrell Jr [302]
	MFP	Gareth Ambler and Axel Benner [303]
	Hmisc	Harrell Jr, Charles Dupont, and others. [304]
	Epitools v0.5-10.1	Aragon [305]
	Skimr v2.1.2	Waring et al. [306]
	Lubridate v1.7.4	Grolemund and Wickham [307]
	interplot	Solt and Hu [308]
	GridExtra v2.3	Auguie [309]
	final-fit	Harrison, Drake, and Ots [310]
	simPH	Gandrud [311]

3

Genetic susceptibility to colorectal cancer

This chapter focuses on the discovery of new colorectal cancer risk loci. As discussed in Chapter 1, risk for colorectal cancer (CRC) incorporates common variation (i.e. present in >1% of the population) with small effect sizes, rarer variants with larger effect sizes, and very rare highly penetrant cancer susceptibility genes. The aim of this work was to identify new CRC risk loci across this spectrum.

3.1 Background

Genome-wide association studies (GWAS) and linkage studies have been powerful tools in the identification of risk loci for human disease, and it is these approaches that I utilise in this chapter. At the time of beginning the GWAS work presented here, around 40 CRC single nucleotide polymorphisms (SNPs) had been identified, largely through the efforts of two consortia which have brought together increasingly large sample sizes, increasing the statistical power to detect rarer and smaller effect variants [259, 312].

The identification of appropriate control datasets is a key problem as GWAS grows, with many publicly available datasets already used for existing GWAS. As a result, controls are often genotyped on different platforms, or may be from different geographical areas, which can introduce bias, an issue highlighted by the work

presented here. The first part of this chapter focuses on evaluating a potential control dataset from the People of the British Isles (PoBI) study, alongside cases taken from the Short Course Oncology Treatment (SCOT) trial. The PoBI trial recruited individuals from families who has been resident in the same geographical area for three generations [242], with genotyping on the Illumina Human 1.2M-Duo chip. The SCOT trial samples were taken from participants in a CRC adjuvant chemotherapy trial [239], and genotyped on Illumina's Global Screening Array. The GWAS work in this chapter formed part of the work of a large consortia to scale GWAS sample size further, through 5 novel GWAS of previously unstudied case cohorts, and meta-analysis of these alongside 10 previously published studies.

I then looked for rarer risk variants through linkage studies in families with CRC or multiple adenoma phenotypes. As noted in Section 1.2.2, genetic linkage studies have been used for over 30 years to identify loci resulting in hereditary cancer syndromes. Whilst early studies could hone this to genomic regions, the more recent availability of whole exome and whole genome sequencing has augmented linkage methodologies, permitting the identification of causative variants at the same time. This was successfully demonstrated in the identification of specific high penetrance variants in the exonuclease domains of *POLE* and *POLD1* in Proofreading Polymerase-Associated Polyposis [86, 313].

The linkage studies presented here are part of an ongoing linkage project of several hundred families recruited to the Colorectal Tumour Gene Identification (CORGI) study. Each family includes at least one individual with whole genome sequencing data, alongside additional family members with genome-wide array data. The CORGI study recruited individuals with a personal history of bowel cancer identified through cancer genetics clinics, alongside their families, and has provided data for several prior GWAS, linkage, and sequencing studies.

3.1.1 Chapter outline

In the first part of this chapter, I undertake a GWAS study evaluating the use of the PoBI dataset as a control for SCOT trial cases (described in Section 2.2).

Genotyping of PoBI and quality control had already been performed by the PoBI study team. I prepared the SCOT samples for genotyping and then perform QC on the data (alongside my own QC of the PoBI data), and complete association testing. I then present the results of the GWAS meta-analysis to which SCOT contributed, which has now been published in Law et al. [115]. The meta-analysis was conducted by colleagues from within the GWAS consortium led by Dr Philip Law; I reviewed the research output and contributed to interpretation of the findings. Finally, I use linkage analysis to search for lower-frequency, higher penetrance CRC risk loci in families with early onset CRC and multiple adenoma phenotypes. I present the methods and results together for each of GWAS, meta-analysis, and linkage studies, for clarity.

In my GWAS of SCOT-PoBI I hypothesised that the studies would be well matched, with both based largely in Britain [239, 242]. However, I observe a very high number of false-positive associations, the cause of which is unclear, and I detail my investigation of this problem. The subsequent GWAS meta-analysis identifies 31 new CRC risk loci, and data from this analysis (excluding the UK Biobank dataset) subsequently forms the 'Base' dataset for variant selection and effect size estimation in Chapter 5. In my linkage analysis I identify several suggestive linkage peaks and candidate genes, with a G374S substitution in *GFI1* of particular interest for further evaluation.

3.2 SCOT-PoBI genome-wide association study

3.2.1 Genome-wide association study methods

Overall quality control exclusions for the SCOT and PoBI datasets are presented in Figure 3.1 and described below. I broadly followed the standardised per-person, and per-SNP QC processes set out by Anderson et al. [271].

3.2.1.1 GWAS datasets

I extracted DNA from the SCOT samples as described in Section 2.6. Genotyping was performed at the Wellcome Centre for Human Genetics, University of Oxford. I

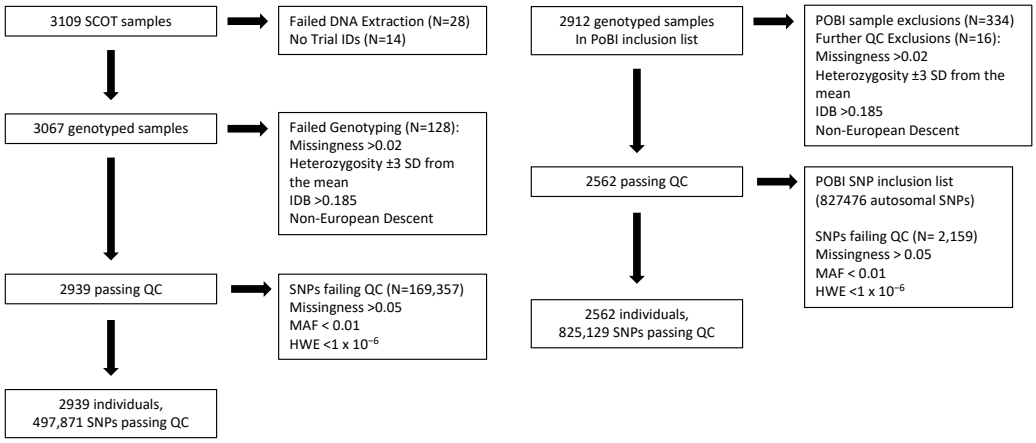

Figure 3.1: Per-person and per-SNP quality control for the SCOT-PoBI GWAS

manually curated genotyping data from the SCOT trial using Genome Studio, following Illumina's Infinium Genotyping Data Analysis guide. Examination of per-person call rates (Figure 3.2) showed six clear sample outliers, with a further two excluded when filtering for samples with call rates <98%. I then examined SNP cluster plots for SNPs with low Cluster Sep (a measure of separation between the three genotype clusters), working from the poorest upwards, identifying and manually editing SNPs which could be re-called. An example of this is given in Figure 3.3.

I aligned the genotyped SCOT samples to the positive strand using strand files for the Global Screening Array on Build 37 from Dr Will Rayner [275], flipping any SNPs on the negative strand and removing duplicates.

The PoBI dataset, comprising VCFs, and SNP- and sample-inclusion lists, was downloaded from the European Genome-Phenome Archive at the Sanger Institute [314]. I converted per chromosome VCFs to PLINK format files using PLINK v1.9 [272]. Dr Will Rayner confirmed that these aligned to Illumina's TOP strand, and I subsequently used Dr Rayner's strand file and script from

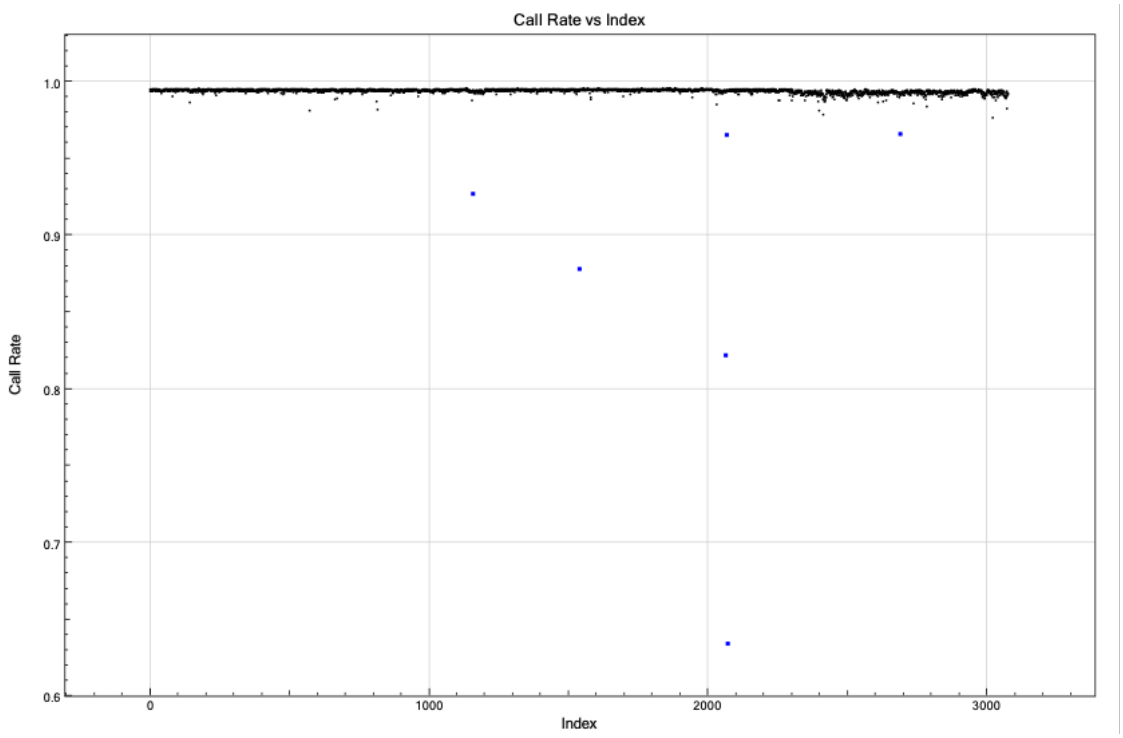

Figure 3.2: Per-person call rate across the SCOT dataset, highlighting six individuals (in blue) with low call rates, who were excluded from further analysis

Neil Robertson [275] to update strand and position to Build 37 of the Genome Reference Consortium reference genome.

3.2.1.2 Per-person quality control

Quality control inclusion lists were provided alongside the PoBI dataset. I therefore excluded 334 samples per original QC prior to analysis.

Sex discordance can potentially indicate sample handling errors, though the actual sex of participants is not significant unless it is relevant to the analyses (for example stratifying by sex). I identified individuals with sex-discordance in SCOT using PLINK v1.9 [272] but did not remove these from the analysis; sex information was not available for PoBI.

Genotype heterozygosity rate and failure rates (missingness) both reflect DNA sample quality; these samples tend to have reduced genotyping accuracy and are therefore excluded. Mean heterozygosity is calculated as (N-O)/N, where N represented number of called genotypes, and O is the number of homozygous

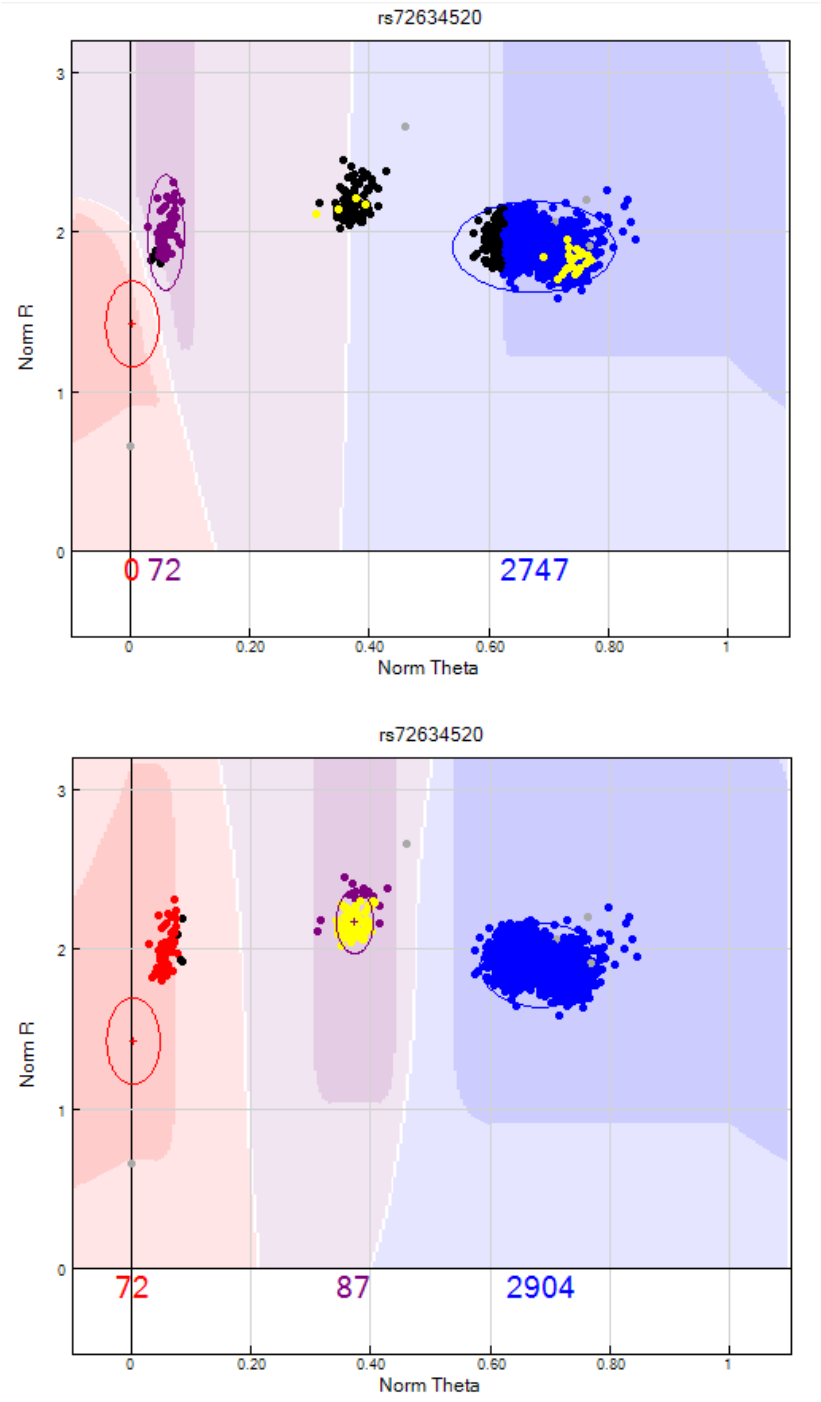

Figure 3.3: Example SNP cluster plots, showing manual re-calling of clusters. Homozygotes are coloured in red and blue, whilst heterozygotes are coloured purple. Automated calling (above) omits the heterozygous cluster, with manual recalling (below) correctly identifying the three clusters.

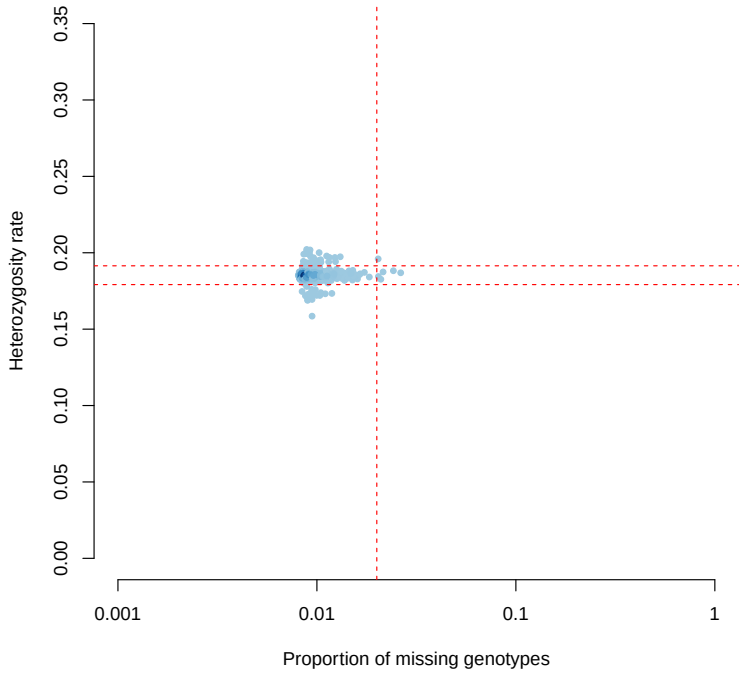

Figure 3.4: Missingness and heterozygosity for SCOT dataset. Vertical red line delineates missingness >2%; horizontal red lines delineate heterozygosity $\pm 3SD$ from the mean

genotypes for an individual. High levels of heterozygosity (i.e. more heterozygous genotypes than the average population) can indicate sample contamination, whilst low levels suggest inbreeding. I excluded individuals with missingness >2% (n = 6 for SCOT, n = 0 for PoBI) and heterozygosity ($\pm 3SD$ from the mean, n = 55 in SCOT, n = 16 in PoBI), shown in figures 3.4 and 3.5.

It is standard practice in GWAS to remove first or second-degree relatives from analyses, and any duplicate samples. Retaining these falsely inflates the proportion of these families genotypes seen in the sample, such that the GWAS may no longer represent the allele frequencies in the source population. I used identity by descent (IBD, described in Section 2.8.2) to identify these individuals, removing 10 individuals with IBD >0.185 from SCOT (none from PoBI).

Population stratification is the main source of confounding in genetic studies, caused by differences in ancestry between case and control groups [271, 315]. This is addressed firstly by selecting individuals from similar populations, followed by

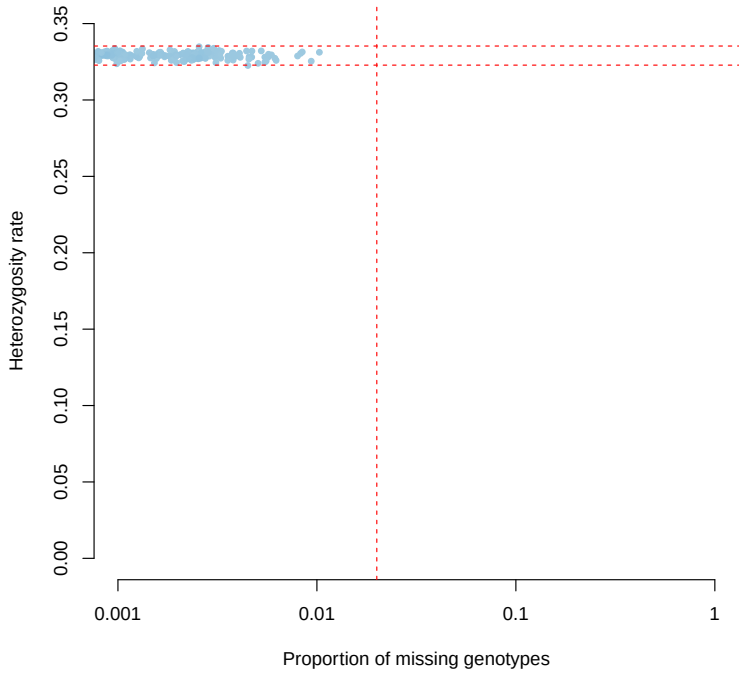

Figure 3.5: Missingness and heterozygosity for PoBI dataset. Vertical red line delineates missingness >2%; horizontal red lines delineate heterozygosity $\pm 3SD$ from the mean

evaluation of population stratification within cases and controls, and removal of individuals from ancestries not under study. Subsequently correction for within-population substructure can be made during analysis.

I assessed population structure using principal components analysis (PCA, Section 2.8.3), using SNPs included in HapMap populations, with AT/CG SNPs (which might be on the incorrect strand) removed. Prior to PCA, regions of long-range linkage dysequilibrium (LD) were excluded, and SNPs were pruned on LD (with SNPs with r^2 >0.2 removed), as correlation between SNPs can influence PCA results. I ran PCA to calculate 10 principal components using EIGENSTRAT [281].

I conducted PCA for PoBI and SCOT datasets initially with HapMap phase 3 haplotype data, which demonstrate continental-based ancestry: 30 trios of Utah residents with northern and western European ancestry (CEU); 30 Yoruba trios from Ibadan, Nigeria (YRI); 44 unrelated Tokyo-based Japanese individuals (JPT); and 45 unrelated Han Chinese individuals resident in Beijing (CHB). PoBI samples clustered

strongly with the CEU samples, whilst SCOT clustered nearby though distinct from the CEU samples, and included a number of individuals diverging towards Nigerian, and Chinese and Japanese ancestries, the latter two clustering together (Figure 3.6).

I then performed PCA with both PoBI and SCOT together with 1000Genomes phase1v3 European populations to evaluate European ancestry populations more closely: British (GBR), CEU (as above), Finnish (FIN), Iberian Spanish-resident populations (IBS), and Toscani in Italy (TSI). Figure 3.7 shows PoBI and the majority of SCOT clustering with all of the European populations in the first two PCS, with two divergent groups again seen for SCOT.

I removed 102 individuals of non-European descent from SCOT (defined as eigenvector >0.015 on PC1 of the European 1000Genomes population PCA) from further analysis; no exclusions were made from PoBI based on PCA.

3.2.1.3 Per-SNP quality control

Following exclusion of individuals failing QC, I performed per-SNP QC, excluding SNPs with MAF<0.01, missingness >0.05, and not in Hardy-Weinberg equilibrium in controls ($P<1x10^{-6}$).

For the SCOT dataset, of 667,228 autosomal SNPs, 169,357 failed QC, resulting in 497,871 SNPs from which to impute. In the PoBI dataset, following exclusion of SNPs not included in the PoBI inclusion list and liftover of SNPs to build 37, 827,288 autosomal SNPs remained. A further 2,159 SNPs were removed following my own QC procedures, leaving 825,129 SNPs as the base for imputation.

3.2.1.4 Phasing and imputation

My planned approach to imputation was to impute from a commonly genotyped backbone. I therefore evaluated the overlap between the genotyped SCOT and PoBI datasets (including perfect proxies for SNPs in either dataset, i.e. SNPs with LD=1). Following QC procedures (including those from PoBI's own QC and exclusion list), there were only ~76K SNPs overlapping, insufficient for imputation, and so I imputed each dataset from its own QC'd SNP set.

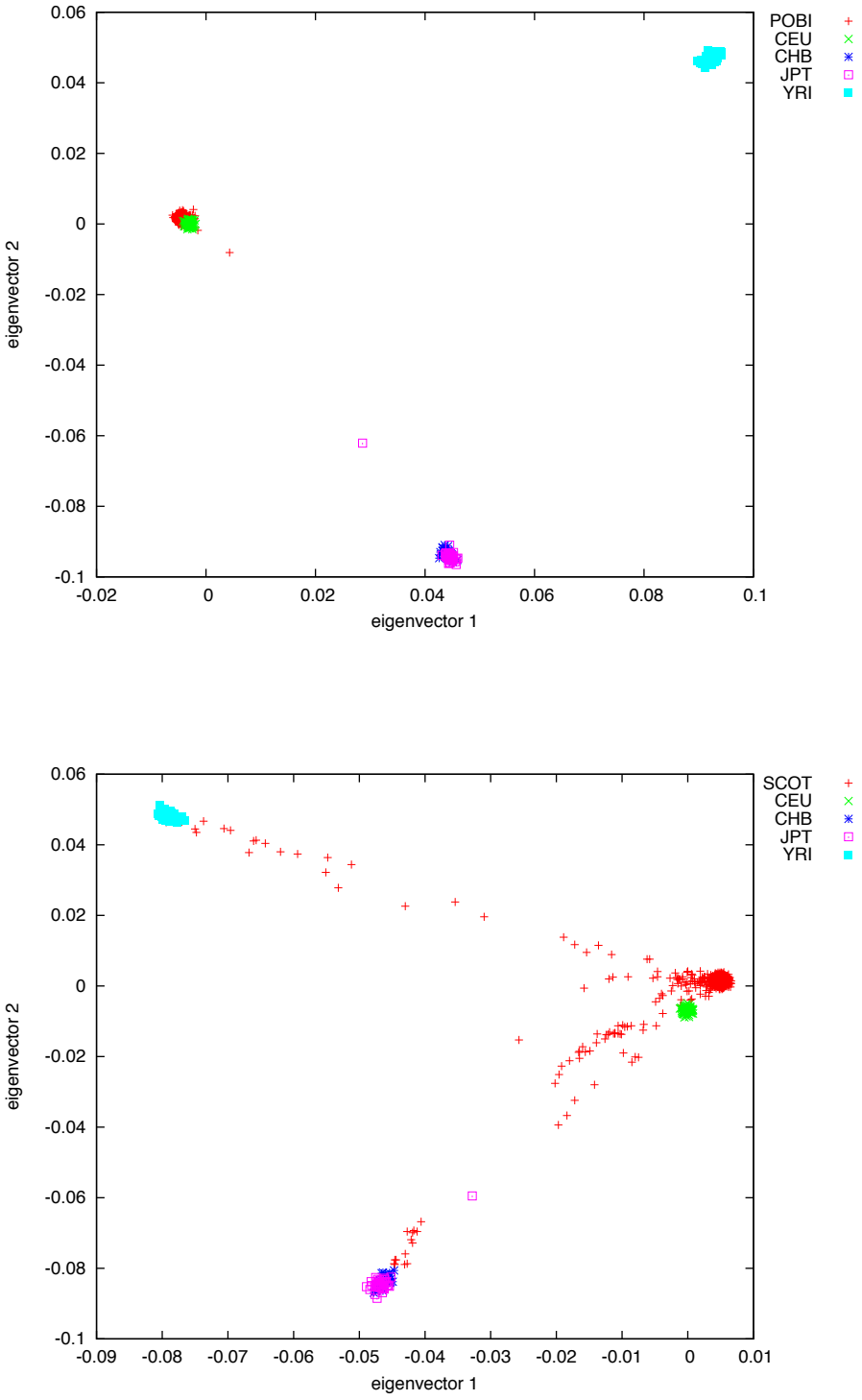

Figure 3.6: PCA of the POBI dataset (top) and SCOT dataset (bottom) with HapMap Samples, showing the first two principal components

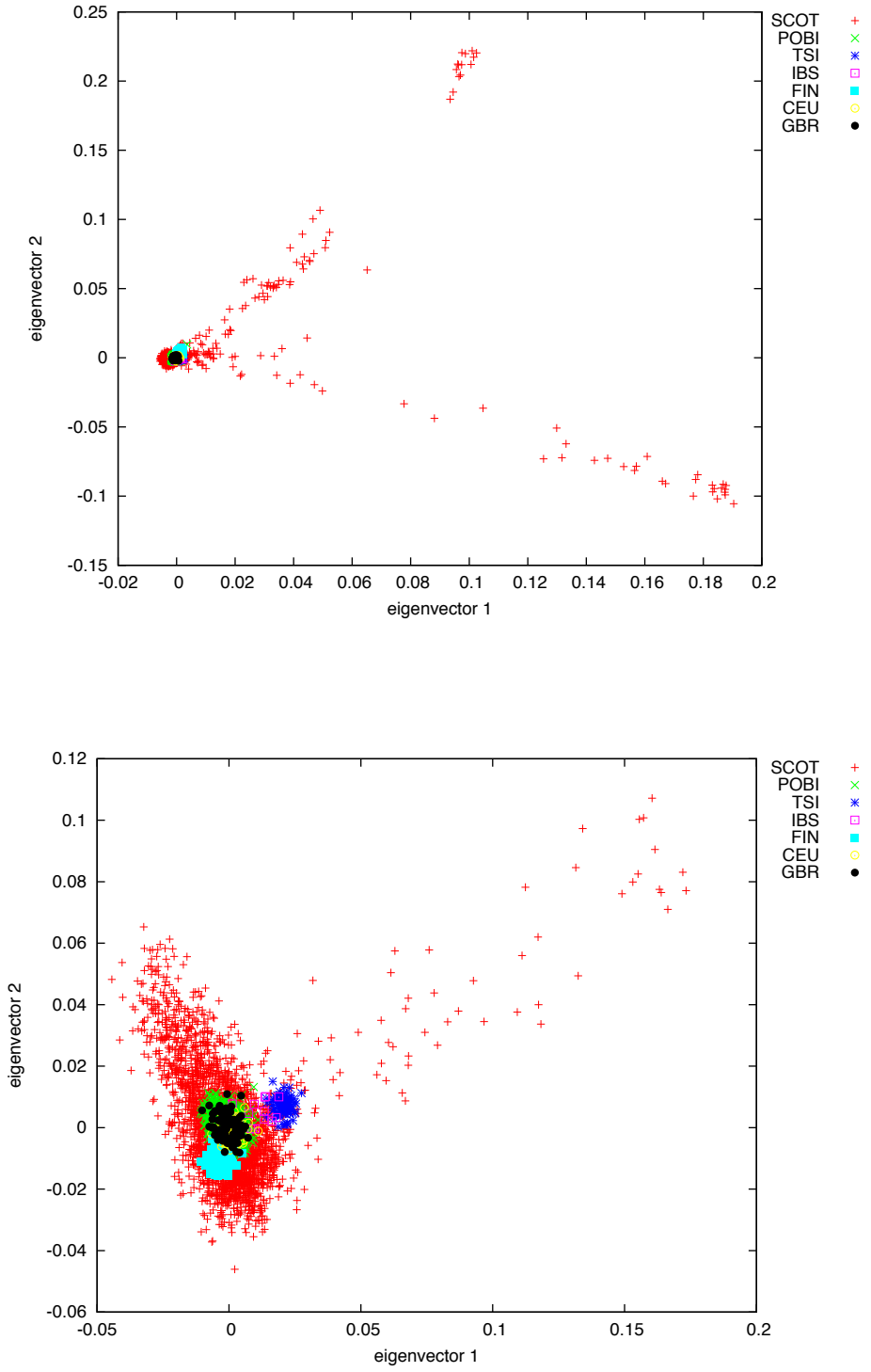

Figure 3.7: PCA of the SCOT and POBI datasets combined with 1000Genomes Phase1v3, showing the first two principal components (top), and repeated following exclusion of 102 SCOT participants with eigenvector >0.015 (bottom)

I used SHAPEITv2.790 [276] for phasing, running the programme in check mode, which checks for duplicate positions and strand alignment issues prior to phasing. Imputation of missing genotypes was performed to a combined reference panel of UK10K and 1000 Genomes Project reference panel [277, 278] using IMPUTE2 [274].

3.2.1.5 Association analysis

Following imputation, I removed SNPs with missingness >5%, MAF <1%, deviation from HWE ($P<1x10^{-6}$), and INFO score <0.4. I then performed association analysis using SNPTEST v2.5.6 [273], and subsequently in PLINK v1.9 for verification [272].

3.2.2 Genome-wide association study results

My initial analysis of SNP association with CRC in the SCOT-PoBI GWAS resulted in many thousands of significantly associated SNPs, as demonstrated in the QQ-plot (Figure 3.8 (top)), suggesting a fundamental problem with either the data or analysis.

In investigating this, I noted a large number of SNP exclusions from POBI, and discussed these with the POBI study investigators. They reported quite conservative QC procedures for the WTCCC2 analysis, but were not aware of any similar issues. I checked scripts with a colleague, confirmed QC measures had been evaluated correctly before and after imputation, and that imputed indels were not included in the association analysis. I excluded all AT/CG SNPs from the analysis, leaving 7,012,359 SNPs for association testing, and restricted the analysis to SNPs with an information score >0.8 in both datasets (5,070,088 SNPs), but neither measure improved the QQplot, with millions of SNPs still significantly associated.

One possibility considered was poor population matching, or failure to account for population diversity in the analysis. As discussed in Section 3.2.1, control populations for a GWAS study should ideally be matched as closely as possible by ancestry. My PCA analysis suggested that the populations were well matched (Figure 3.7). However, the PoBI study does contain some genetic outliers located in the Orkney Islands, and in my initial association analysis I had not adjusted

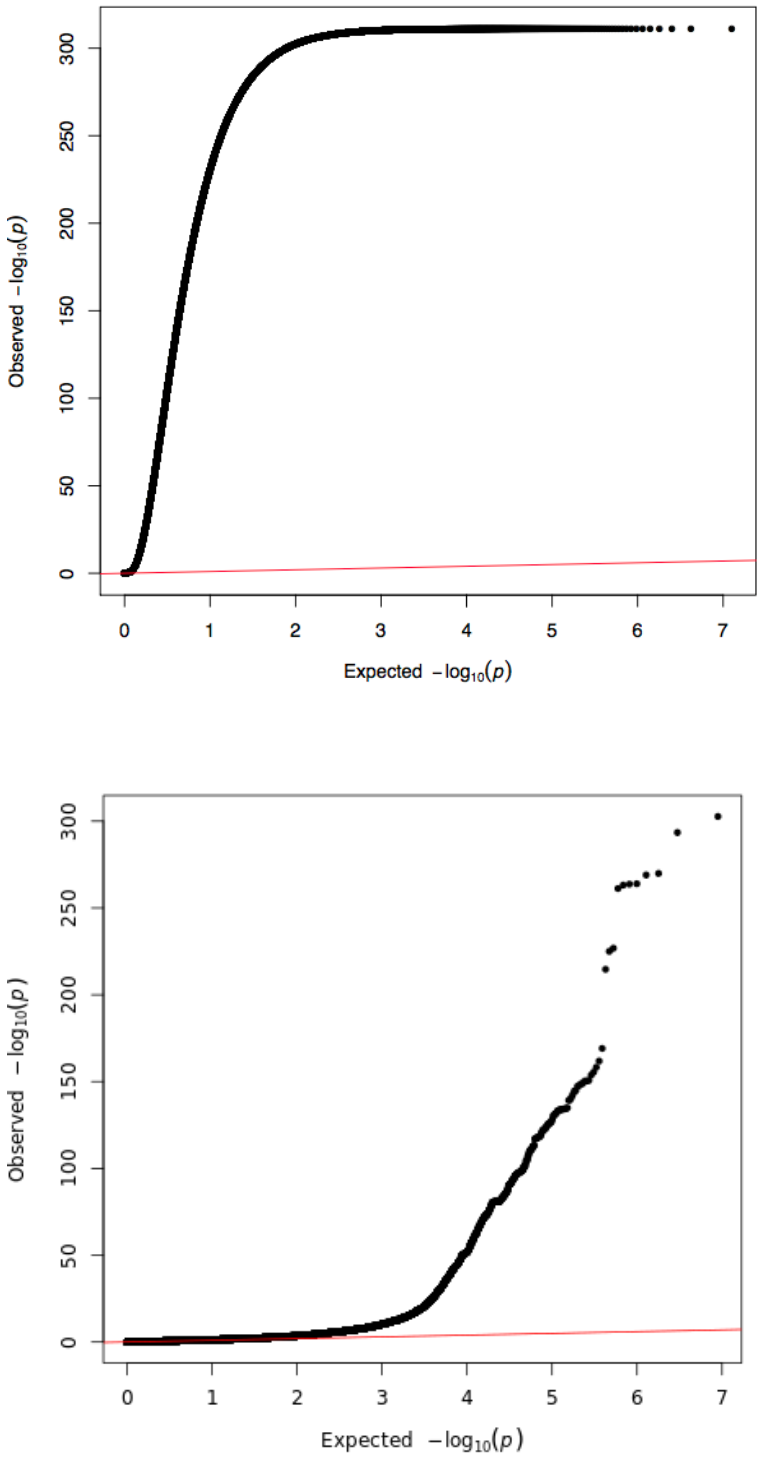

Figure 3.8: QQ plots for SCOT-PoBI association analysis, conducted using SNPtest (top), and PLINK (bottom), showing very large deviations from the expected distribution

for PCs. I therefore conducted the SNPtest analysis with 4 PCs included, with no improvement in the QQ plot. I subsequently reran the association analysis using PLINK rather than SNPtest, to see if this were an issue with my SNPtest analysis. Though the results were slightly improved the QQ plot (Figure 3.8) continued to show marked deviation.

Association analysis using only SNPs genotyped in both datasets created more plausible association results (Figure 3.9). I undertook a post-hoc exploration of SNPs with the most significant associations, presumed to be false positives (annotated as P=0 in SNPtest output, n = 164,817), using Chromosome 1 as a test case. Comparing MAFs in the SNPtest output with UK10K MAFs (including only SNPs with rsIDs present in both to enable matching, n = 1788), I found that PoBI MAFs deviated considerably more than for those for the SCOT data, with a mean deviation of 53.2% compared to 3.91%. Notably, 291 of these SNPs had been genotyped in both datasets, rather than imputed. As these deviations were so extreme, I re-checked the MAFs as calculated on the imputed PoBI dataset using QCTool; the MAFs were far closer to those of UK10K - with a mean deviation of 4.3%. This suggested an issue introduced on combining the datasets for analysis. Two different approaches were taken to combining the datasets (performed automatically by SNPtest software, and then manually using PLINK), but the issue was seen in both analyses (though to a lesser extent in PLINK-based analyses). However, following inspection of the data, reviewing of my scripts with colleagues, and trials of different SNPtest settings, I was unable to resolve the issue.

As a result of the problems with using the PoBI dataset, 4349 controls from the German Heinz-Nixdorf dataset (genotyped on the same array as SCOT) were used by Dr Philip Law in GWAS using SNPtest and subsequent meta-analysis study. This GWAS showed an expected distribution on QQplot (Figure 3.11), with a small amount of genomic inflation secondary to the geographic differences in populations.

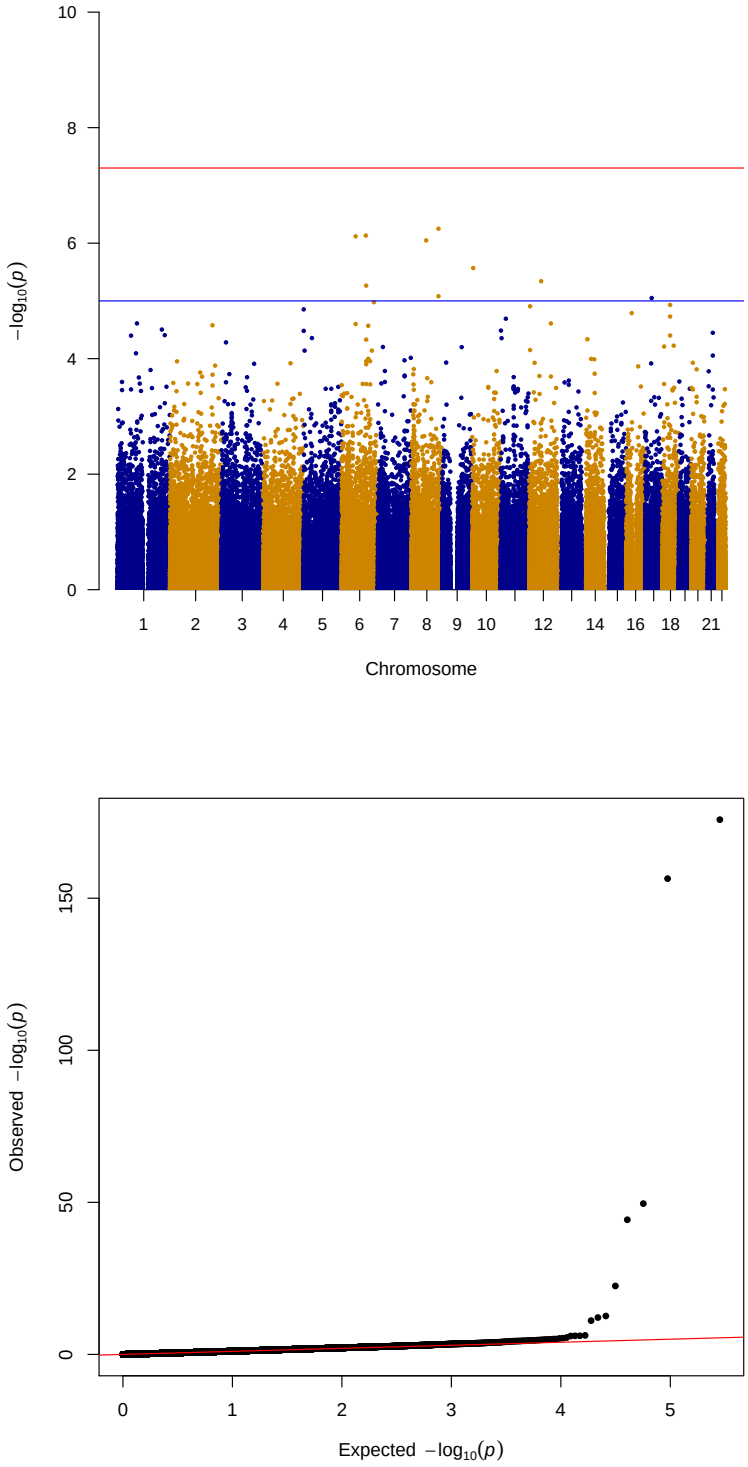

Figure 3.9: Manhatten plot (top) and QQ plots (bottom) of association analysis for common genotyped SNPs only

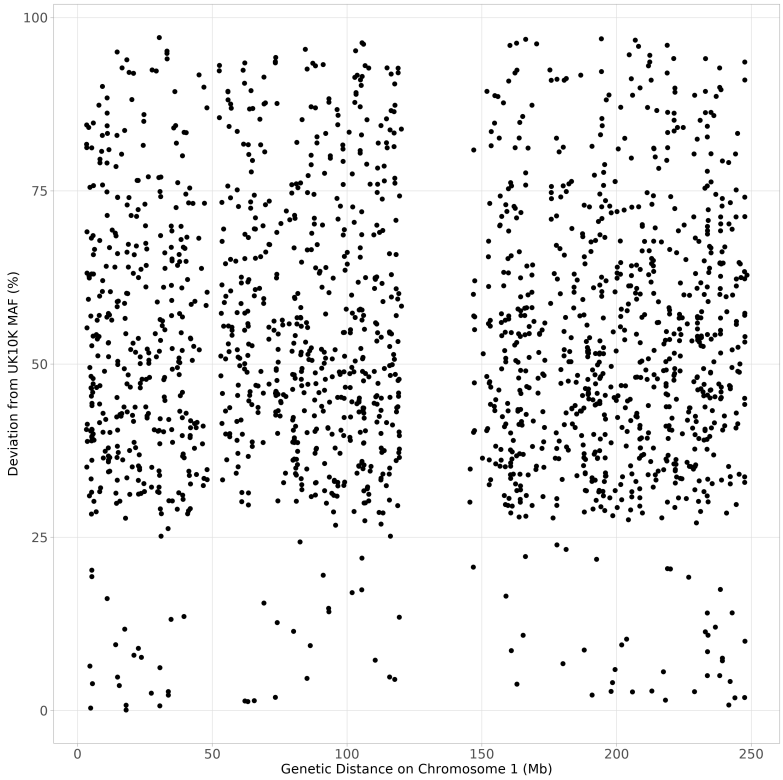

Figure 3.10: Distribution of SNPs with MAFs in SNPtest output highly divergent from UK10K MAF

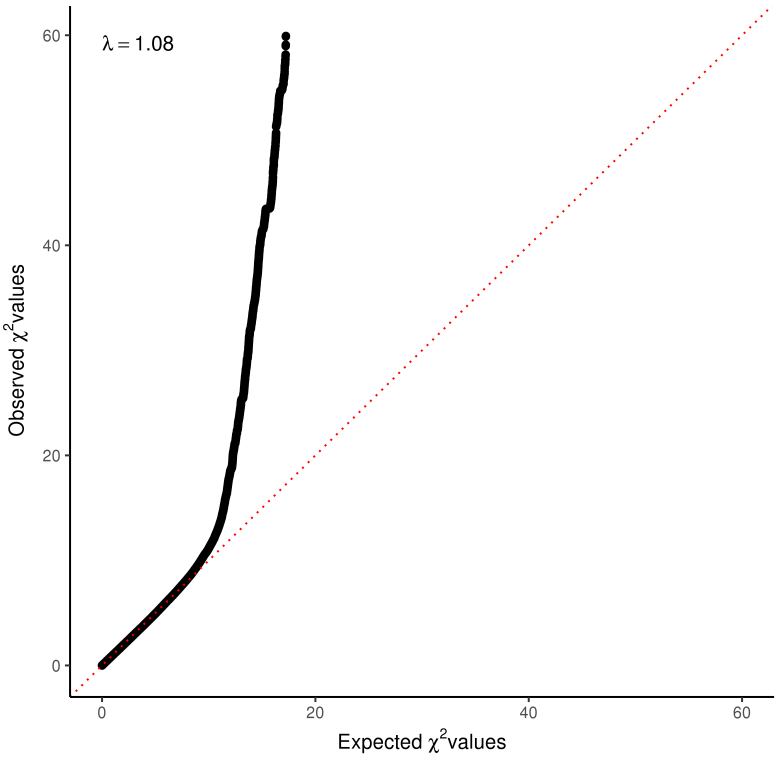

Figure 3.11: QQ plot of assocation analysis results of SCOT and Heinz-Nixdorf datasets

3.3 Genome-wide association study meta-analysis

3.3.1 Meta-analysis methods

Association testing of the 5 new GWAS cohorts and meta-analysis with 10 previously published studies (described in Section 2.2) was performed by Dr Philip Law and colleagues. To summarise, for new GWAS cohorts, individuals of non-European ancestry (based on PCA with HapMap Version 2 SNPs), and with missingness >5% were excluded. For related pairs, either the control was deleted (for case-control pairs), or the individual with the lowest call rate. In per-SNP QC, SNPs with missingness >5%, MAF < 0.5%, and deviation from Hardy-Weinberg Equilibrium (P< 10^{-5}) were excluded.

Most datasets were imputed to the merged 1000G/UK10K reference panel. Exceptions were:

- UKB, imputed to data from 1000 Genomes (Phase 3), UK10K and the Haplotype Reference Consortium (and performed by UKB)
- FIN and DACHS, imputed to a combined reference of 1000 Genomes Project supplemented with population-appropriate samples: 3882 Finnish haplotypes from the Sequencing Initiative Suomi (SISu), and 3000 sequenced individuals with CRC respectively.

HLA regions were imputed using a reference panel from the Type 1 Diabetes Genetics Consortium (T1DGC) using SNP2HLA.

Following imputation, quality thresholds of information scores >0.8 and MAF >0.5% were applied prior to analysis. SNPTEST v2.5.2 was used for association testing, assuming an additive genetic model, with adjustment for population stratification for SCOT, UKB, FIN, DACHS and NSCCG-OncoArray. Q-Q plots confirmed the absence of differential genotyping and adequate matching of case/controls. Genomic inflation was assessed for each GWAS study by calculating the genomic inflation factor, λ_{GC}, (the ratio of the median observed distribution of the test statistic to the expected median) using GenABEL [316].

Meta-analysis was performed with META v1.7, using the inverse-variance method based on a fixed-effects model [286]. Associations were evaluated for all previously reported SNPs, as identified through a Pubmed literature search (January 2018) and associated citations. Heterogeneity was evaluated with Cochran's Q-statistic and the I^2 statistic. New associations were defined as SNPs reaching $P < 5 \times 10^{-8}$ which were not previously reported. One SNP per 500kb region was included. The possibility of a false positive results was assessed for each SNP using the Bayesian False Discovery probability (BFDP) [317]. This calculates the approximate Bayes factor based on a prior probability of association (here 10^{-5}) given a plausible OR (95th percentile OR in the meta-analysis = 1.2).

In conditional analysis, regression was performed conditioning on the known and newly discovered SNPs. This analysis used Genome-wide Complex Trait Analysis (GCTA) [318]. SNPs with a conditional $P < 5 \times 10^{-8}$ and r^2>0.1 were clumped using PLINK v1.9.

Imputation accuracy was evaluated for all non-genotyped newly discovered SNPs by evaluating a subset of 201 samples included in the CORGI and NSCCG datasets for whom whole genome sequencing data was also available. There was over 98% concordance between imputed and sequenced SNPs in these individuals.

3.3.2 Meta-analysis results

Five new GWAS studies were conducted and meta-analysed alongside 10 previously published studies as part of the COGENT (COlorectal cancer GENeTics) consortium [319]. The resulting meta-analysis included 34,627 CRC cases and 71,349 controls. Over 10 million SNPs were imputed in each dataset. There was minimal genomic inflation in any dataset (λ_{GC} 1.01-1.11).

Of previously discovered loci for CRC risk in European populations (Table 3.1), 28 were replicated with associations at GWAS-significance level. An additional four loci (3q26.2, 12p13.32, 16q22.1 and 16q24.1) had associations significant at $P < 5 \times 10^{-6}$, while seven showed little evidence of association and had BFDP values >0.99, suggesting false-positive associations.

Following exclusion of 500 kb regions flanking known CRC risk SNPs, 623 were associated with CRC at $P < 5 \times 10^{-8}$, which after stepwise model selection identified risk SNPs at 31 previously unreported loci (Table 3.2). In addition, 9 SNPs previously validated in Asian populations were validated in Europeans (Table 3.2). A further eight independent SNPs at known or newly discovered European risk loci were also identified in conditional analysis (Table 3.3).

Table 3.1: Associations for previously reported CRC risk loci in Europeans. Adapted from "Association analyses identify 31 new risk loci for colorectal cancer susceptibility", Law et al., DOI: 10.1038/s41467-019-09775-w, CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`).

Locus	rsID	Position	Risk/Alternate Allele	P	Reported OR (95% CI)	OR (95% CI) in meta	P in meta	BFDP	Reference	Alternate Lead SNP, Risk/Alt	LD (R^2)	OR (95% CI)	P	BFDP
1p36.12	rs72647484	22,587,728	T/C	0	1.24 (1.15-1.33)	1.13 (1.09-1.17)	0.00e+00	2.02e-03	Al-Tassan NA, Sci Rep, 2015	-	-	-	-	-
1q25.3	rs10911251	183,081,194	A/C	0	1.09 (1.06-1.12)	1.08 (1.06-1.10)	0.00e+00	1.17e-05	Whiffin N, Hum Mol Genet, 2014	rs4546885, G/C	0.87; 0.99	1.09 (1.07-1.11)	0.00e+00	2.00e-07
1q41	rs6691170	222,045,446	T/G	0	1.06 (1.03-1.09)	1.08 (1.06-1.10)	0.00e+00	5.55e-05	Houlston RS, Nat Genet, 2010	rs6658977, T/G	0.97; 0.99	1.08 (1.06-1.11)	0.00e+00	1.20e-05
2q35	rs992157	219,154,781	A/G	0	1.10 (1.06-1.13)	1.08 (1.05-1.10)	0.00e+00	1.07e-04	Orlando O, Hum Mol Genet, 2016	rs13020391, C/T	0.58; 0.90	1.09 (1.06-1.11)	0.00e+00	2.80e-06
3p22.1	rs35360328	40,924,962	A/T	0	1.14 (1.09-1.19)	1.09 (1.06-1.12)	0.00e+00	6.00e-02	Schumacher FR, Nat Commun, 2015	rs35470271, G/A	0.92; 0.96	1.09 (1.06-1.13)	0.00e+00	1.00e-02
3p14.1 **	rs812481	66,442,435	G/C	0	1.09 (1.05-1.11)	1.01 (0.99-1.03)	4.38e-01	1.00e+00	Schumacher FR, Nat Commun, 2015	rs2279290, G/T	0.22; 0.97	1.05 (1.03-1.08)	5.40e-05	1.00e+00
3q26.2 *	rs10936599	169,492,101	C/T	0	1.08 (1.04-1.11)	1.07 (1.04-1.10)	1.00e-07	4.80e-01	Houlston RS, Nat Genet, 2010	rs35446936, G/A	0.99; 1.00	1.07 (1.05-1.10)	1.00e-07	2.90e-01
4q22.2 **	rs1370821	94,943,383	T/C	0	1.07 (1.04-1.1)	1.05 (1.02-1.07)	3.41e-05	9.90e-01	Schmit SL, JNCI, 2018	-	-	-	-	-
4q26 **	rs3987	118,759,055	G/A	0	1.36 (1.22-1.52)	1.02 (1.00-1.04)	9.74e-02	1.00e+00	Real LM, PLoS One, 2014	-	-	-	-	-
4q32.2 **	rs35509282	163,333,405	A/T	0	1.53 (1.33-1.75)	1.02 (0.99-1.06)	2.43e-01	1.00e+00	Schmit SL, Carcinogenesis, 2014	rs186722897, T/A	0.14; 0.81	1.10 (1.03-1.16)	2.01e-03	1.00e+00
5p15.33	rs2735940	1,296,486	A/G	0	1.09 (1.05-1.12)	1.08 (1.06-1.10)	0.00e+00	1.45e-04	Schmit SL, JNCI, 2018	-	-	-	-	-
5p13.1	rs58791712	40,281,797	G/T;G	0	1.10 (1.07-1.12)	1.10 (1.07-1.13)	0.00e+00	1.23e-05	Schmit SL, JNCI, 2018	rs1445011, C/T	1.00; 1.00	1.11 (1.08-1.13)	0.00e+00	0.00e+00
6p21.31	rs6906359	35,528,378	C/T	0	1.11 (1.06-1.16)	1.11 (1.07-1.16)	0.00e+00	2.20e-01	Schmit SL, JNCI, 2018	rs16878812, A/G	1.00; 1.00	1.11 (1.07-1.15)	0.00e+00	1.30e-02
6p21.2	rs1321311	36,622,900	A/C	0	1.10 (1.07-1.13)	1.09 (1.06-1.11)	0.00e+00	7.90e-04	Dunlop MG, Nat Genet, 2012	rs1321310, C/T	0.97; 0.99	1.09 (1.06-1.11)	0.00e+00	5.34e-04
6p12.1	rs62404968	55,714,314	C/T	0	1.09 (1.05-1.12)	1.07 (1.05-1.1)	0.00e+00	1.20e-01	Schmit SL, JNCI, 2018	rs62404966, C/T	0.97; 1.00	1.08 (1.05-1.10)	0.00e+00	2.10e-02
8q23.3	rs16892766	117,630,683	C/A	0	1.27 (1.20-1.34)	1.26 (1.22-1.31)	0.00e+00	0.00e+00	Tomlinson IP, Nat Genet, 2008	-	-	-	-	-
8q24.21	rs6983267	128,413,305	G/T	0	1.27 (1.16-1.39)	1.19 (1.16-1.21)	0.00e+00	0.00e+00	Tomlinson I, Nat Genet, 2007	-	-	-	-	-
10p14	rs10795668	8,701,219	G/A	0	1.15 (1.10-1.20)	1.11 (1.09-1.14)	0.00e+00	0.00e+00	Tomlinson IP, Nat Genet, 2008	rs7894531, G/A	0.86; 0.94	1.13 (1.10-1.15)	0.00e+00	0.00e+00
10q11.23 **	rs10994860	52,645,424	C/T	0	1.09 (1.05-1.12)	1.05 (1.02-1.08)	3.65e-04	1.00e+00	Schmit SL, JNCI, 2018	-	-	-	-	-
10q24.2	rs1035209	101,345,366	T/C	0	1.12 (1.08-1.16)	1.11 (1.08-1.14)	0.00e+00	0.00e+00	Whiffin N, Hum Mol Genet, 2014	rs2193352, G/A	1.00; 1.00	1.11 (1.08-1.14)	0.00e+00	0.00e+00
11q13.4	rs3824999	74,345,550	G/T	0	1.08 (1.05-1.10)	1.09 (1.07-1.11)	0.00e+00	0.00e+00	Dunlop MG, Nat Genet, 2012	rs57796856, T/A	0.98; 0.99	1.09 (1.07-1.12)	0.00e+00	0.00e+00
11q23.1	rs3802842	111,171,709	C/A	0	1.11 (1.08-1.15)	1.15 (1.12-1.17)	0.00e+00	0.00e+00	Tenesa A, Nat Genet, 2008	rs3087967, T/C	0.97; 0.99	1.15 (1.12-1.18)	0.00e+00	0.00e+00
12p13.32 *	rs3217810	4,388,271	T/C	0	1.19 (1.13-1.25)	1.11 (1.06-1.15)	1.10e-06	8.50e-01	Whiffin N, Hum Mol Genet, 2014	-	-	-	-	-
12q13.13	rs11169552	51,155,663	C/T	0	1.09 (1.05-1.12)	1.06 (1.03-1.08)	5.30e-06	9.70e-01	Houlston RS, Nat Genet, 2010	rs11169572, C/T	0.21; 0.93	1.09 (1.06-1.11)	0.00e+00	7.00e-07
12q24.12	rs3184504	111,884,608	C/T	0	1.09 (1.06-1.12)	1.09 (1.06-1.11)	0.00e+00	1.00e-07	Schumacher FR, Nat Commun, 2015	rs597808, G/A	0.98; 0.99	1.09 (1.07-1.11)	0.00e+00	1.00e-07
12q24.21	rs72013726	115,890,835	C;CACA	0	1.08 (1.04-1.11)	1.08 (1.06-1.11)	0.00e+00	1.79e-03	Schmit SL, JNCI, 2018	rs7315438, T/C	0.78; 1.00	1.08 (1.05-1.10)	0.00e+00	1.50e-04
12q24.22 **	rs73208120	117,747,590	G/T	0	1.16 (1.11-1.23)	1.04 (1.00-1.08)	2.67e-02	1.00e+00	Schumacher FR, Nat Commun, 2015	-	-	-	-	-
13q13.2	rs10161980	34,093,518	C/G	0	1.08 (1.05-1.11)	1.06 (1.04-1.09)	0.00e+00	1.30e-01	Schmit SL, JNCI, 2018	rs9537521, G/A	0.84; 0.92	1.08 (1.05-1.11)	0.00e+00	4.70e-02
14q22.2	rs4444235	54,410,919	C/T	0	1.11 (1.08-1.15)	1.08 (1.05-1.10)	0.00e+00	1.44e-04	COGENT, Nat Genet, 2008	rs35107139, C/A	0.67; 0.89	1.09 (1.07-1.12)	0.00e+00	1.60e-06
15q13.3	rs11632715	33,004,247	A/G	0	1.11 (1.08-1.16)	1.07 (1.05-1.10)	0.00e+00	1.93e-04	Tomlinson IP, PLoS Genet, 2011	rs73376930, G/A	0.19; 0.79	1.18 (1.15-1.21)	0.00e+00	0.00e+00
16q22.1 *	rs9929218	68,820,946	G/A	0	1.10 (1.06-1.14)	1.06 (1.04-1.09)	5.00e-07	7.60e-01	COGENT, Nat Genet, 2008	rs9939049, A/T	0.99; 0.99	1.06 (1.04-1.09)	2.00e-07	5.80e-01
16q24.1 *	rs2696839	86,340,448	G/C	0	1.06 (1.04-1.09)	1.05 (1.03-1.08)	1.30e-06	8.90e-01	Schmit SL, JNCI, 2018	-	-	-	-	-
18q21.1	rs4939827	46,453,463	T/C	0	1.18 (1.12-1.23)	1.20 (1.18-1.23)	0.00e+00	0.00e+00	Broderick P, Nat Genet, 2007	rs7226855, A/G	1.00; 1.00	1.21 (1.19-1.24)	0.00e+00	0.00e+00
19q13.11	rs10411210	33,532,300	C/T	0	1.15 (1.10-1.20)	1.14 (1.10-1.19)	0.00e+00	2.92e-05	COGENT, Nat Genet, 2008	rs73039434, T/G	0.35; 0.93	1.30 (1.22-1.38)	0.00e+00	0.00e+00
20p12.3	rs961253	6,404,281	A/C	0	1.12 (1.08-1.16)	1.11 (1.08-1.13)	0.00e+00	0.00e+00	COGENT, Nat Genet, 2008	-	-	-	-	-
20q11.22 **	rs2295444	33,173,883	C/T	0	1.08 (1.05-1.1)	1.02 (1-1.05)	2.57e-02	1.00e+00	Schmit SL, JNCI, 2018	-	-	-	-	-
20q13.13	rs1810502	49,057,488	C/T	0	1.08 (1.05-1.1)	1.07 (1.05-1.1)	0.00e+00	1.46e-03	Schmit SL, JNCI, 2018	-	-	-	-	-
20q13.33	rs4925386	60,921,044	C/T	0	1.08 (1.05-1.10)	1.10 (1.08-1.13)	0.00e+00	0.00e+00	Houlston RS, Nat Genet, 2010	rs1741640, C/T	0.52; 0.88	1.16 (1.13-1.20)	0.00e+00	0.00e+00
Xp22.2	rs5934683	9,751,474	T/C	0	1.07 (1.04-1.10)	1.08 (1.03-1.12)	4.06e-04	1.00e+00	Dunlop MG, Nat Genet, 2012	rs2732875, C/G	0.43; 1.00	1.18 (1.13-1.24)	0.00e+00	1.32e-04

* does not formally replicate in meta-analysis but $P < 5times10^{-6}$; ** does not replicate in meta-analysis, BFDP > 0.99

Table 3.2: CRC risk loci newly discovered in Europeans. Adapted from "Association analyses identify 31 new risk loci for colorectal cancer susceptibility", Law et al., DOI: 10.1038/s41467-019-09775-w, CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`).

Locus	rsID	Position	Genes	Risk/Alternate Allele	RAF	OR (95% CI)	P	BFDP	I^2	P_{het}
Newly discovered										
1p34.3	rs61776719	38461319	SF3A3, FHL3	C/A	0.45	1.07 (1.05-1.10)	2.19x10-10	1.98x10-3	1	0.44
1p32.3	rs12143541	55247852	PARS2, TTC22	G/A	0.15	1.1 (1.06-1.13)	9.44x10-10	7.44x10-3	16	0.28
2q11.2	rs11692435	98275354	ACTR1B	G/A	0.90	1.12 (1.07-1.16)	1.22x10-8	0.079	29	0.14
2q33.1	rs11893063	199601925	PLCL1, SATB2	A/G	0.47	1.07 (1.04-1.09)	9.34x10-9	0.069	43	0.04
2q33.1	rs7593422	200131695	PLCL1, SATB2	T/A	0.55	1.07 (1.05-1.10)	3.56x10-11	3.50x10-4	15	0.28
3p21.1	rs9831861	53088285	SFMBT1, RFT1	G/T	0.59	1.07 (1.05-1.09)	4.17x10-10	3.72x10-3	0	0.87
3q13.2	rs12635946	112916918	C3ORF17, BOC	C/T	0.62	1.08 (1.06-1.10)	1.02x10-11	1.03x10-4	11	0.33
4q24	rs17035289	106048291	CXXC4, TET2	T/C	0.83	1.1 (1.07-1.13)	2.73x10-10	2.30x10-3	0	0.95
4q31.21	rs75686861	145621328	HHIP	A/G	0.10	1.12 (1.08-1.16)	1.76x10-9	0.014	0	0.49
6p24.1	rs2070699	12292772	EDN1	T/G	0.48	1.07 (1.04-1.09)	3.88x10-9	0.031	29	0.14
6p21.33	rs3131043	30758466	IER3, DDR1	G/A	0.43	1.07 (1.05-1.1)	2.67x10-8	0.159	60	0.01
6p21.32	rs9271770	32594248	HLA-DQA1	A/G	0.81	1.08 (1.05-1.11)	3.60x10-8	0.192	0	0.91
6q21	rs6928864	105966894	PREP, PRDM1	C/A	0.91	1.13 (1.09-1.19)	1.37x10-8	0.094	0	0.73
7p12.3	rs10951878	46926695	TNS3	C/T	0.49	1.06 (1.04-1.09)	1.10x10-8	0.08	0	0.65
7p12.3	rs3801081	47511161	TNS3	G/A	0.68	1.08 (1.06-1.11)	2.00x10-11	1.96x10-4	50	0.01
9p21.3	rs1412834	22110131	CDKN2B	T/C	0.50	1.08 (1.06-1.11)	4.13x10-14	5.05x10-7	14	0.30
11p15.4	rs4450168	10286755	SBF2	C/A	0.17	1.1 (1.06-1.13)	1.24x10-8	0.079	0	0.81
12q13.3	rs7398375	57540848	LRP1	C/G	0.72	1.09 (1.06-1.13)	3.91x10-10	3.23x10-3	0	0.93
13q13.3	rs12427600	37460648	SMAD9	C/T	0.24	1.09 (1.06-1.11)	5.43x10-11	5.01x10-4	0	0.81
13q22.1	rs45597035	73649152	KLF5	A/G	0.64	1.08 (1.05-1.10)	2.16x10-10	1.94x10-3	0	0.53
13q22.3	rs1330889	78609615	EDNRB, PU4F1, RNF219	C/T	0.87	1.11 (1.07-1.14)	6.50x10-10	5.25x10-3	0	0.59
13q34	rs7993934	111074915	COL4A2	T/C	0.65	1.08 (1.05-1.10)	3.03x10-11	2.94x10-4	0	0.55
15q22.31	rs4776316	67007813	SMAD6	A/G	0.73	1.08 (1.05-1.10)	1.11x10-8	0.076	22	0.21
15q23	rs10152518	68177162	SKOR1, PIAS1	G/A	0.19	1.08 (1.05-1.11)	3.24x10-8	0.18	0	0.84
15q26.1	rs7495132	91172901	CRTC3	T/C	0.12	1.11 (1.07-1.14)	7.92x10-10	6.34x10-3	29	0.14
16q23.2	rs61336918	80007266	MAF, DYNLRB2	A/T	0.29	1.09 (1.06-1.12)	2.04x10-12	2.14x10-5	0	0.90
17p12	rs1078643	10707241	FOXL1, FOXC2	A/G	0.77	1.09 (1.06-1.12)	4.14x10-11	3.81x10-4	0	0.56
19p13.11	rs285245	16420817	AP1M1, KLF2	T/C	0.11	1.11 (1.07-1.15)	3.71x10-8	0.195	2	0.42
19q13.33	rs12979278	49218602	MAMSTR, FUT2	T/C	0.53	1.07 (1.05-1.09)	6.11x10-10	5.35x10-3	15	0.28
20q13.13	rs6066825	47340117	PREX1	A/G	0.65	1.1 (1.08-1.13)	3.82x10-17	5.67x10-10	0	0.49
20q13.33	rs3787089	62316630	RTEL1	C/T	0.32	1.07 (1.05-1.10)	5.80x10-9	0.043	0	0.80
Previously identified in Asian GWAS										
5q31.1	rs639933	134467751	PITX1	C/A	0.38	1.07 (1.05-1.10)	1.14x10-9	9.50x10-3	0	0.73
6p21.1	rs6933790	41672769	TFEB	T/C	0.83	1.1 (1.07-1.14)	3.65x10-10	3.03x10-3	21	0.23
10q22.3	rs704017	80819132	ZMIZ1-AS1	G/A	0.60	1.1 (1.08-1.13)	2.96x10-16	4.15x10-9	23	0.21
10q25.2	rs12255141	114294892	VTI1A, TCF7L2	G/A	0.10	1.11 (1.07-1.15)	2.97x10-9	0.022	0	0.81
12p13.31	rs10849438	6412036	CD9	G/T	0.12	1.12 (1.08-1.16)	1.04x10-10	9.49x10-4	21	0.23
17p13.3	rs73975588	816741	NXN, TIMM22	A/C	0.87	1.1 (1.06-1.13)	8.71x10-9	0.058	33	0.11
19q13.2	rs9797885	41873001	B9D2	G/A	0.71	1.08 (1.05-1.10)	2.77x10-10	2.43x10-3	0	0.70
20p12.3	rs6055286	7718045	HAO1	A/G	0.15	1.11 (1.07-1.14)	9.69x10-11	8.61x10-4	50	0.02
20q13.12	rs2179593	42660286	TOX2	A/C	0.72	1.07 (1.05-1.10)	4.62x10-9	0.035	0	0.67

Table 3.3: CRC risk variants discovered in analysis conditioning on the sentiel SNP and risk loci in Europeans. Adapted from "Association analyses identify 31 new risk loci for colorectal cancer susceptibility", Law et al., DOI: 10.1038/s41467-019-09775-w, CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`).

Sentinel SNP	Conditioning SNP	Cytoband	Chromosome	Position	Top Gene	Risk/Alternate Allele	RAF	Conditional OR (95% CI)	Conditional P value	BFDP	I^2	P_{het}	Reference
rs77776598	rs2735940	5p15.33	5	1,240,998	TERT	C/T	0.06	1.16	2.84x10-10	0.003	0	0.93	-
rs4944940	rs3824999	11q13.4	11	74,415,252	POLD3, CHRDL2	G/A	0.96	1.28	3.21x10-17	2.73x10-9	6	0.38	-
rs12818766	rs3217810	12p13.32	12	4,376,091	PARP11, CCND2	A/G	0.18	1.10	5.29x10-9	0.037	30	0.16	Wang (2014)
rs1570405	rs4444235	14q22.2	14	54,554,234	BMP4	G/A	0.31	1.07	1.91x10-7	0.125	0	0.46	Tomlinson (2008)
rs16969681	rs73376930	15q13.3	15	32,993,111	GREM1	T/C	0.09	1.21	2.85x10-24	1.33x10-16	442	0.04	Tomlinson (2008)
rs16959063	rs73376930	15q13.3	15	33,105,730	GREM1	A/G	0.01	1.32	5.40x10-9	0.23	30	0.13	-
rs17816465	rs73376930	15q13.3	15	33,156,386	GREM1	A/G	0.20	1.12	8.36x10-15	1.07x10-7	44	0.04	-
rs899244	rs2696839	16q24.1	16	86,700,030	FOXL1, FOXC2	T/C	0.21	1.09	1.13x10-10	4.06x10-3	14	0.29	-
rs6091213	rs1810502	20q13.13	20	49,384,745	PARD6B, BCAS4	C/T	0.26	1.08	5.68x10-10	3.88x10-8	0	0.96	-
rs4811050	rs1810502	20q13.13	20	48,980,670	PARD6B, BCAS4	A/G	0.18	1.09	2.43x10-11	4.06x10-3	20	0.23	-
rs6085661	rs961253	20p12.3	20	6,693,128	BMP2	T/C	0.39	1.09	1.63x10-14	4.76x10-3	6	0.39	Tomlinson (2008)

RAF = risk allele fequency in Europeans, OR = Odds ratio, CI = confidence interval, BFDP = Bayesian False Discovery Probability, P_{het} = Probability for heterogenety

Of the candidate genes proximal to newly discovered loci, some are of particular interest. At the 1p34.3 locus, rs61776719 lies between *SF3A3* and *FHL3*. *FHL3* is part of the FHL3/TGFβ/Smad signalling pathway [320], acting as both a tumour suppressor and oncogene in a number of malignancies [321], while *SF3A3* encodes subunit 3 of the splicing factor 3a protein complex involved in pre-mRNA processing, and inhibits p53 activity [322]. Two other *SMAD* genes were also identified in this analysis - *SMAD9* at q13.13 and *SMAD6* at 15q22.31 - joining other BMP/TGF-β signalling genes (*SMAD7*, *GREM1*, *BMP2*, *BMP4*) identified in previous GWAS, and highlighting the importance of this pathway [115]. Several other newly highlighted loci also have roles in key signalling pathways. At 10q25.2, *TCF7L2* is has a role in *MYC* signalling, and has been associated with multiple cancers [323], whilst *RTEL1* at 20q33.3 is a DNA helicase involved in the maintenance of chromosomal integrity [324].

Other candidate genes are known to have roles in carcinogenesis. The 9p21.3 locus is commonly deleted in a number of cancers, with *CDKN2A/B* deletions in up to 13% of tumours [325]; downregulation confers resistance to immune checkpoint inhibitors [326]. *CXXC4* at 4q24 is a known tumour suppressor, which negatively regulates Wnt signalling, in a number of malignancies including CRC [327, 328], whilst *HHIP*, the candidate gene at 4q31.21, is a regulator of Hedgehog signalling which has been implicated in a large number of cancers [329]. *EDN1* at 6p24.1 encodes the preprotein of a potent vasoconstrictor. The endothelin signalling network promotes epithelial-to-mesenchymal transition, cell proliferation, and neovascularisation, and thus may have a number of roles in carcinogenesis [330]. At 6q21, *PRDM1* is a known tumour suppressor gene, which has been shown to act in response to p53, silencing stem-cell related genes [331]. The 7p12.3 locus, and *TNS3* gene, is significantly associated with pancreatic cancer in GWAS [332]. *TNS3* is a focal adhesion protein which is involved in a molecular phosphorylation switch, triggered by activation of MAPK1/2 by epithelial growth factor, and is essential for initiating and perpetuating cell migration, and implicated in epithelial cell tumorigenesis [333].

3.4 Linkage

3.4.1 Linkage methods

3.4.1.1 Family inclusion criteria

I identified families for linkage from the CORGI research study. Families for whom we had blood samples or array or sequencing data available for at least two family members with CRC or multiple polyps were included (ideally with sequencing data for at least one family member); 183 families fulfilled these criteria.

These families will ultimately be analysed in phenotypic groups:

- early onset CRC (<60 years at diagnosis) with at least one additional relative with CRC under 70, dominant inheritance
- early onset CRC (<60 years at diagnosis) with at least one additional relative with CRC under 70, recessive inheritance
- late onset CRC - families with two or more cancer with age of onset not meeting the above criteria
- multiple-adenoma - one family member with ≥ 10 adenomas, with an additional family member with ≥ 10 polyps of any kind
- dominant inheritance, any phenotype
- recessive inheritance, any phenotype

In addition per-family analysis of families with exceptional phenotypes who may not fulfil the criteria will be analysed.

3.4.1.2 Data sources

Individuals included in linkage families were genotyped on one of four different arrays: Illumina Hap550, Illumina HumanCoreExome, Affymetrix Axiom Biobank, and Illumina OncoArray. Initial QC, phasing and imputation to the 1000G Phase 3 reference was performed by Dr Claire Palles.

Sequencing data was from individuals included in several projects: the Oxford-Illumina WGS500 project [263], the Illumina-215 project, and Complete Genomics 196 (Section 2.3).

3.4.1.3 Post-imputation QC and identification of SNP union

I extracted a set of SNPs passing QC in all datasets and sequenced to 10x read depth. Per-SNP QC following imputation was more stringent than that used for GWAS, owing to the higher cost of incorrect genotyping in linkage analysis. I excluded SNPs with missingness >0.05, MAF<0.1, INFO score <0.8, and those not in Hardy-Weinberg equilibrium (P $<1x10^{-3}$), using GTOOLS, QCTOOLS, PLINK and BCFtools (Section 2.9).

Unfortunately the sequencing quality of the Complete Genomics (CG) dataset was relatively low, resulting in very low overlap (~35,000 SNPs) between all of these datasets. As a result, CG-sequenced individuals are excluded from the current analyses, and they plus a number of recently recruited family members have subsequently been genotyped on Illumina's CoreExome-24 v1.2 array. I extracted DNA from blood samples where required as described in 2.6, and genotyping was carried out at Genomics Birmingham, University of Birmingham. Further analysis of this extended data is yet to be completed. As a result, ~1.2 million genotypes were included in the linkage analysis.

3.4.1.4 Linkage analysis

I used MERLIN [279] to perform non-parametric linkage analysis, calculating Kong and Cox LOD scores [334] in centimorgan (cM) grids. I examined sequencing data (previously annotated by Dr Luke Freeman-Mills) in regions within ±0.75 centimorgan from the peak regions of nominal K+C significance (P≤ 0.05) in exponential analysis.

I prioritised variants initially based on the presence of alleles in controls in the WGS500 dataset, retaining those with fewer than 2 alleles, and allele count <100 in Exome Aggregation Consortium (ExAC) exomes [335]. I then filtered these further based on functional consequence as predicted by the Ensembl Variant Effect Predictor (VEP, [336]), retaining variants encoded as "transcript ablation", "splice acceptor variant", "splice donor variant", "stop gained", "frameshift variant", "stop lost", "start lost", "transcript amplification", "inframe insertion", "inframe

deletion", "missense variant", "protein altering variant", "splice region variant", "incomplete terminal codon variant", "stop retained variant", "5 prime UTR variant", or "3 prime UTR variant".

I evaluated allele frequency in other reference datasets including UK10K, 1000 Genomes, ExAC, TopMed and gnomAD. I examined the functional impact of coding variants using SIFT [337] and PolyPhen [338] scores, and of non-coding variants using HaploReg v4.1 [339], conservation using PhyloP, and mRNA expression in GTEx.

For a shortlist of variants of particular interest, I checked frequency of the specific variant, and of truncating mutations in the candidate gene, in sequencing data from 3,441 CRC cases (2,636 from the 100,000 Genomes Project and 805 from our own sequencing projects), and 14,426 Northern European controls (as defined by PCA). Examination of the 100,000 Genomes Project data was undertaken by Professor Ian Tomlinson.

3.4.2 Linkage results

I present here the results of the analysis of two families of interest, and of the three families with a multiple-adenoma phenotype.

3.4.2.1 Family 2733

Family 2733 (also known in CORGI as StM3023) consists of two brothers, one with CRC at the age of 41, and the other 2 adenomas and 3 hyperplastic polyps at 41. Their father, for whom we do not have genetic data, had CRC at the age of 65.

Linkage analysis highlighted linkage peaks on most chromosomes. Figure 3.12 shows sample peaks on chromosomes 1 and 8. After filtering sequenced variants covered by the LOD peaks by frequency in control populations and functional annotation in VEP, 195 variants of potential interest remained; restricting these further to missense variants left 27 (Table 3.4). The majority were not predicted to be deleterious based on SIFT and PolyPhen scores.

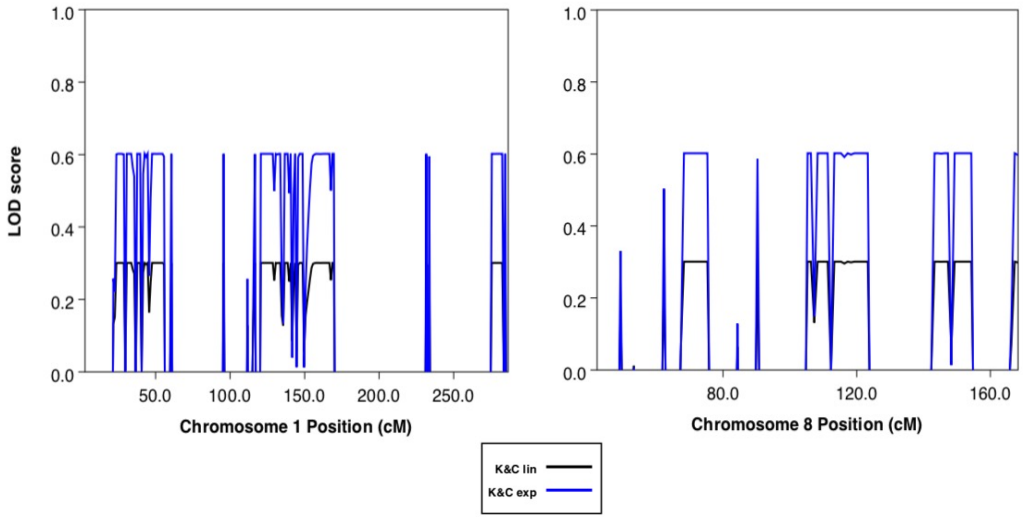

Figure 3.12: LOD peaks for Chromosomes 1 and 8 in family 2733

Table 3.4: Missense variants identified in linkage analysis of family 2733

Position	rsID	Gene	Amino Acid Change	1000 Genomes	ExAC	UK10K	SIFT/PolyPhen	PhyloP	Comments
1:11844641:C>T	rs780856228	C1orf167	P[249/306/523/587/1163]L	NA	0.0003	NA	1/0.003	0.238	Benign
1:12024350:C>T	rs11553676	PLOD1	R[441/488]W	0.0002	0.0001	NA	0/0.489-0.656	1.225	connective tissue disorders
1:16053832:A>G	rs376306853	PLEKHM2	D402G	NA	0.0003	0.0003	0/0.999	2.047	Candidate
1:17312707:G>T	rs773484997	ATP13A2	H[324/1084]N	NA	0.0002	NA	unknown	-0.824	Unknown impact, not conserved
1:17396617:T>C	rs763691887	PADI2	E[461/577]G	NA	0.0001	NA	0.22/0.099-0.465	0.879	Benign
1:22159016:C>T	rs534160301	HSPG2	G3727S	0.0002	0.0000	NA	0.14/0.933	2.402	Likely benign
1:23760749:G>A	rs780277335	ASAP3	T[154/641/650]I	NA	0.0000	NA	1/0.19	2.287	Benign
1:92941735:C>T	rs367740686	GFI1	G374S	NA	0.0000	NA	0.02/0.997	2.635	Candidate
1:114255919:A>T	rs529721504	PHTF1	D[11/202/210/155]E	0.0002	0.0001	NA	0.01-0.31/0.493-0.968	-0.128	Gene poorly characterised
1:155630058:A>C	-	YY1AP1	I[394/517/528/537/548/594/666/686]S	NA	NA	NA	0/0.956-0.964	1.165	Gene poorly characterised
1:156819165:C>G	rs148814291	INSRR	K439N	NA	0.0000	NA	0/1	0.614	NTRK on opposite strand
1:158687796:A>C	rs138237790	OR6K3	F[37/53]C	NA	0.0000	0.0003	0/0.992	1.819	olfactory gene
3:194063002:G>C	rs199846219	CPN2	L[144/520]V	NA	0.0001	NA	0.13/0.02	-0.995	Benign
5:79950727:G>C	rs1574197	MSH3	A61P	0.0417	NA	NA	0.29/0	-0.103	Benign
8:135614791:G>C	rs760117294	ZFAT	L[329/379/391]V	NA	NA	NA	0.08/0.528-0.999	2.717	Candidate
8:146157646:G>A	rs144660752	ZNF16	T176I	NA	0.0000	NA	0.14/0.013	-1.329	Benign
10:29812417:C>G	rs373969255	SVIL	E[616/1042]D	NA	0.0000	NA	0.77/0.001	-1.968	Benign
11:66029645:G>A	rs143312355	KLC2	E[31/94/171]K	0.0008	0.0008	0.0017	0.06/0.491-0.757	2.081	Low colon exp
11:29812417:C>G	rs373969255	SVIL	E[616/1042]D	NA	0.0000	NA	0.77/0.01	-1.968	Benign
11:76371565:G>A	rs150096749	LRRC32	R358C	0.0006	0.0004	0.0011	0.19/0.462	0.984	Benign
11:86663244:C>A	rs771911050	FZD4	G185V	NA	0.0000	0.0003	0.49/0.041	2.760	Benign
11:3723942:G>A	rs146515319	NUP98	P1088L	NA	0.0002	0.0004	0.12/0.003-0.013	-0.243	Benign
11:5717709:G>A	rs200668710	TRIM22	E83K	0.0002	0.0012	0.0004	0.07/0.134	0.590	Benign
11:6519875:G>A	rs149433314	DNHD1	A144T	NA	0.0000	0.0001	0.38/0.009	-0.247	Benign
11:6651470:A>G	rs199544459	DCHS1	P1519S	NA	0.0008	NA	0.06/0.215	2.689	Benign
11:129742942:A>G	rs377366500	NFRKB	V[867/877/892]A	NA	NA	NA	0.21-0.34/0.261	0.853	Benign
15:59323283:A>G	-	RNF111	S[88/540]G	NA	NA	NA	0.64-0.83/0	0.984	Benign
18:51800428:C>G	-	POLI	T[206/244/257/402/517/599/707]S]	NA	NA	NA	0.09-0.23/0.087-0.351	1.442	Benign

3. Genetic susceptibility

I identified several strong candidates among potentially deleterious missense variants. Firstly, rs367740686 (Chr1:92941735:C>T, G374S), is found in exon 7 of *GFI1* at the 1p22.1 locus. *GFI1* encodes a nuclear zinc finger protein, which acts as a transcription repressor. It binds directly to a consensus DNA recognition motif (TAAATCAC(A/T)GCA) in target gene promoters [340], controlling histone modification by recruiting histone deacetylase complexes, and ultimately silencing target gene promoters [341]. Gfi1 regulates a large number of genes and is involved in many biological processes, and in CRC has been reported to act as a tumour suppressor gene [342].

The variant is present at a frequency of 0.00002 in ExAC, and absent from WGS500. It is a conserved site (PhyloP score 2.635), and is expressed in transverse colon in GTEx. The Gfi1 protein consists of an N-terminal SNAG domain, and intermediate region, and 6 C-terminal zinc finger domains [340]. The G374S substitution is predicted to be deleterious in each transcript (SIFT 0.02; PolyPhen 0.977); this residue is located on zinc finger 5, which binds to the DNA recognition motif. G374S is classified as a variant of uncertain significance for severe congenital neutropoenia 2 in ClinVar, but not previously reported in association with other phenotypes.

An additional potential site of interest is rs148814291 (Chr1:156819165:C>G, 1q23). This encodes a K439N change in *INSRR*, an insulin receptor protein which is minimally expressed in colonic tissue, however on the reverse strand this represents an intronic change in *NTRK1* (neurotrophic tyrosine kinase, type 1; c.122+7180C>A), predicted to result in nonsense mediated decay or retained introns. This variant has a frequency of 0.00003 in ExAC, is not present in WGS500, and is conserved. Examination of the variant in Haploreg v.4.1 shows an active enhancer mark (H3K4me1) in colonic and rectal mucosa at this site. *NTRK1* encodes a membrane-bound receptor, TRKA receptor tyrosine kinase, which phosphorylates itself and MAPK pathway proteins, regulating cell differentiation. Somatic *NTRK1* fusions result in constitutive activity of the kinase, and are seen in lung adenocarcinomas, CRC, ductal breast cancers, and melanoma, as well as multiple other cancers [343, 344]. *NTRK1* fusions are actionable with targeted

therapies, entrectinib and larotrectinib, the former of which was the first drug to receive tumour-agnostic approval by NICE in the UK.

Another potentially interesting change, rs760117294, was found in *ZFAT* on 8q24.22 (Chr8:135614791:G>C). *ZFAT* is a highly conserved transcriptional regulator which was originally identified in individuals with autoimmune thyroid disease, with a role in haematopoiesis and angiogenesis [345]. *ZFAT* knockdown induced apoptosis in blood cell lines [346], and the gene is identified as a region of high copy number gain in ovarian cancer [347]. There are 9 transcripts for this gene as a result of alternative splicing, resulting in L>V changes at position 329, 379 or 391. The predicted impact of the change is variable, tolerated by SIFT (0.08) but by PolyPhen scores is predicted to be possibly/probably damaging. This variant is not present in WGS500 controls, is not seen in our core adenoma or HPPS samples, and is present once in the core early onset cancer samples.

rs376306853 (Chr1:16053832:A>G) results in a D402G amino acid change in Exon 9 of *PLEKHM2* at 1p36.21. It is rare (frequency of 0.0003 in ExAC and UK10K), deleterious (SIFT 0, PolyPhen 0.999), and well conserved (PhyloP score 2.047). *PLEKHM2* is required for lysosomal movement away from the microtubule-organising centre towards the periphery of the cell, acting through a protein chain which recruits the lysosomal transport protein kinesin 1 to the lysosome [348, 349].

PLEKHM2 has recently been described as a fusion partner in an *ALK*-fusion mutation in small-cell lung cancer [350]. Mechanistically, outward movement of lysosomes is seen in response to acidic microenvironments [351] such as are found in tumours, and knockdown of *PLEKHM2* inhibits this response to acidification [348]. Disruption of this process in prostate cancer cells reduces tumour invasion [352]. However, the D402G residue is not located in key functional protein domains. The remaining deleterious variants were less likely candidates based on gene function (Table 3.4).

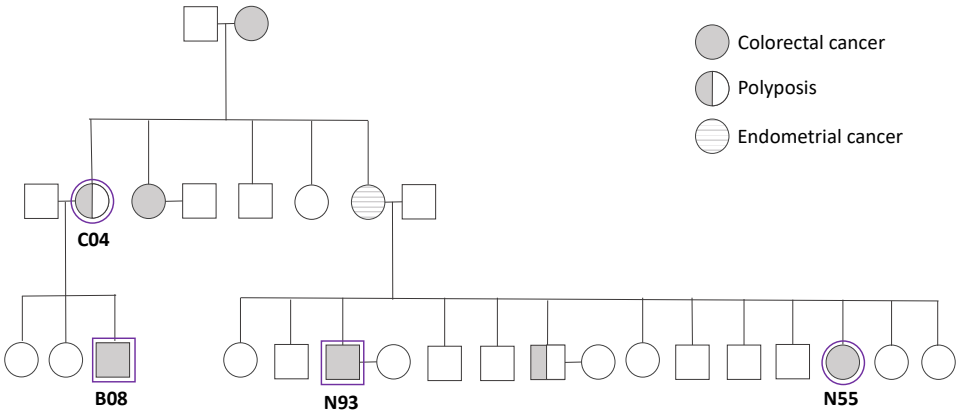

Figure 3.13: Pedigree for family 3491. Individuals with genetic data are outlined in purple. C04 (sequenced) had 9 adenomas at 63 years; B08 had CRC at 54; N93 had CRC at 64; N55 had CRC at 46.

3.4.2.2 Family 3491

Family 3491 (known in CORGI as BelN99/28951) has an early onset CRC phenotype, as seen in the pedigree in Figure 3.13. Narrow LOD peaks at or nearing the maximum possible LOD for the pedigree (1.183) were seen at 1p36.21, 3q26.2, 4p15.1, 8q24.3, and 9p13.3 (Figure 3.14). Filtering variants as with the previous family, there was 1 missense variant localising to these regions; including any functionally annotated variant produced 3 variants.

1p36.21 has not previously been associated with CRC, and the genomic region underlying this peak (Chr1:14,200,000-14,630,000) contained no functional variants. The region lies just upstream of *PRDM2* which I examined separately but contained no potentially pathogenic variants.

The 3q26.2 locus is frequently implicated in carcinogenesis and is a known CRC risk locus in GWAS [248]. I identified one sequenced variant with functional impact, rs557433728 (Chr3:170606856delTTTGT) at this locus. This is a 3' UTR variant downstream of *EIF5A2*, a widely expressed gene with several biological roles including in cell cycle progression and regulating apoptosis. The variant

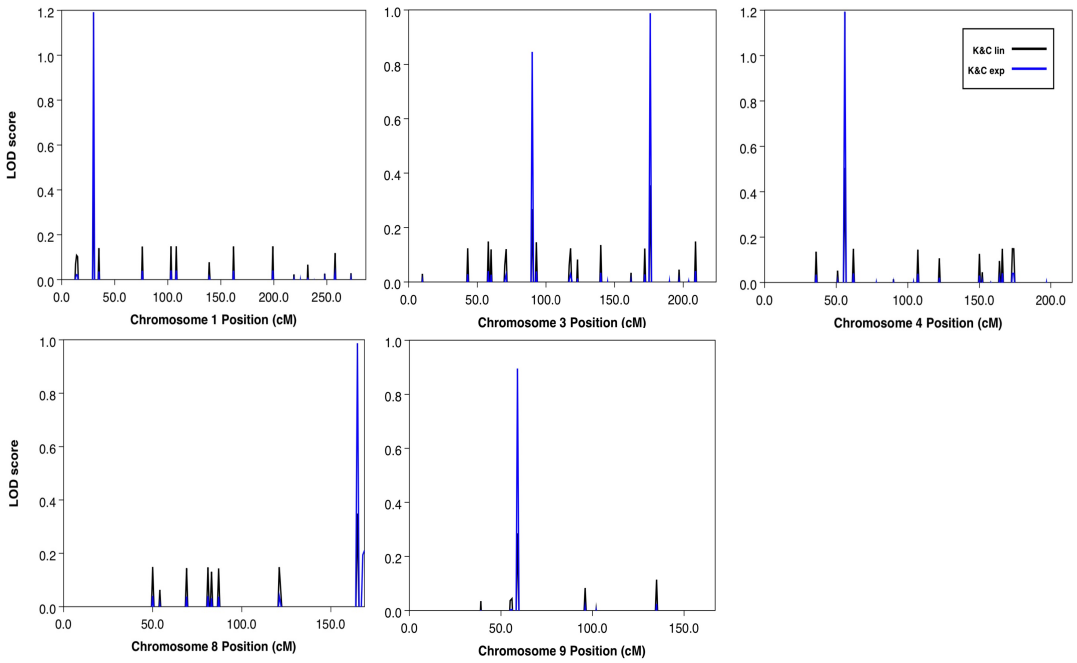

Figure 3.14: LOD peaks for family 3491

is present once in the WGS500 dataset, and has an allele frequency of ≤ 0.002 in gnomAD and TOPmed genomes. *EIF5A2* acts an oncogene in CRC, with over-expression correlated to tumour stage and survival [353]. This is effected through the upregulation of *MTA1*, which induces epithelial-mesenchymal transition [354]. Additional potential genes of interest at this locus include *MECOM*, *TERC*, *TRAF2*, *TNIK*, and *SKIL*.

The genomic region underlying the linkage peak at 4p15.1 (Chr4:33,387,000-35,412,000) contained no functional variants. This region is intergenic, adjacent to the genes *ARAP2* which acts downstream of *RhoA* to control focal adhesion dynamics, and *DTHD1* which is involved in the apoptosis pathway, but neither has been strongly associated with carcinogenesis. Deletions in the region of 4p15.1 have previously been reported in colorectal adenocarcinomas and adenomas [355], with LOH found at 4p15.1-4p15.31 in sessile serrated adenomas [356].

At 8q24.3, I identified a splice site variant, rs754431097, under the LOD peak (Chr8:143,190,000-143,904,000) in *ADGRB1*. However the gene is poorly expressed

in gastro-intestinal tissue, being largely brain specific, thus this is unlikely to be pathogenic. Copy number variation at 8q24.3 is common in a number of cancers in TCGA data, and is associated with prognosis - deletion of 8q24.3 conferred an increase in 5-year mortality in CRC patients (OR 4.12, 95% confidence interval 1.15-14.82), as did gain of a copy (OR 1.75, 95% CI 1.32-2.31) [357]. Amplification of this locus is also associated with MSI-high colorectal tumours [358].

The single variant under the peak at 9p13.3 (Chr9:33,416,000-36,034,000), rs113475471 (Chr9:36003344:G>T), locates to a polymorphic pseudogene, *OR13C7*, encoding an olfactory receptor, and so is likely irrelevant. The 9p13.3 locus has not previously been associated with CRC, but is a lung adenocarcinoma susceptibility locus in TWAS [359], with copy number gains in prostate cancer, olfactory neuroblastoma, and gastric cancer cell lines [360–362].

3.4.2.3 Multiple adenoma families

My multiple-adenoma linkage analysis included 3 families fulfilling our pre-specified multiple adenoma phenotype.

Family 3702 (StM2349) includes two brothers, one of whom (sequencing ID IT16) was diagnosed with 9 adenomas and 6 hyperplastic polyps (HPPs) at the age of 29, and the other CRC at 63 with 4 adenomas and 16 HPPs. Their paternal grandfather and paternal uncle had both died from CRC, and two cousins have colorectal polyps.

Family 6863 (StM3708) also includes two brothers, one with 16 adenomas and 2 HPPs at 63 (sequencing ID IT10), and his brother 10 polyps (unspecified) at the age of 66.

Family 17471 (SGH171498) includes a father and two sons. The father had over 150 TVAs and serrated adenomas by the age of 43. His sons subsequently underwent colonoscopy; the elder son 15 adenomas and 2 HPPs at the age of 21, and the younger had 39 adenomas and 1 HPPs by 20 years of age. Of note, these individuals have all been sequenced in the GEL project in addition to being genotyped.

3. Genetic susceptibility

Linkage analysis identified narrow peaks at 4q13.1, 5p15.33, 7q32.1, 11q23.3, 14q21.1 and 18q.1-12.2. As with Family 3491 I examined all functionally annotated variants (listed in Table 3.5), as there were just two missense variants.

Table 3.5: Functionally annotated variants identified in linkage analysis of multiple adenoma families

Locus	Position	rsID	Gene	Impact	1000 Genomes	ExAC	UK10K	SIFT/PolyPhen	PhyloP
4q13.1	4:65147217:A>G	rs146610448	TECRL	V[13/298]A	0.0002	0.0009	0.0021	0/0.294-0.996	2.112
	4:66196519:G>A	rs551436483	EPHA5	3' UTR variant	-	-	0.0029	-	-0.215
7q32.1	7:127225458:G>A	rs548961200	ARF5	upstream gene variant	-	-	0.0028	-	0.202
	7:127672115:C>G	rs142220343	LRRC4	5' UTR variant, upstream gene variant	-	-	0.0284	-	1.408
	7:127945084:G>A	-	RBM28	3' UTR variant	-	-	-	-	-0.054
11q23.3	11:118097941:T>C	-	MPZL3	3' UTR variant, downstream gene variant	-	-	-	-	0.230
		-	AMICA1	upstream gene variant	-	-	-	-	0.230
	11:118098148:G>C	-	MPZL3	3' UTR variant	-	-	-	-	0.415
		-	AMICA1	upstream gene variant	-	-	-	-	0.415
	11:118415381:C>G	rs546689601	IFT46	3' UTR variant, downstream gene variant	0.0008	-	0.0021	-	1.359
		rs546689601	TMEM25	intron variant	0.0008	-	0.0021	-	1.359
	11:118533570:CA>C	-	TREH	P27X	-	-	-	unknown	-0.319
	11:118533572:G>GTGATCA	-	TREH	intron variant	-	-	-	-	-0.765
	11:118533574:C>CACAA	-	TREH	intron variant	-	-	-	-	0.944
	11:118829210:A>T	-	UPK2	3' UTR variant	-	-	-	-	1.829
	11:118987186:T>C	rs752017704	C2CD2L	3' UTR variant	-	-	-	-	0.006
	11:119186771:C>T	rs192316821	MCAM	intron variant, splice acceptor variant	-	-	0.0005	-	-0.280
18q12.1	18:31763514:G>C	rs72961222	NOL4	3' UTR variant, intron variant	0.0008	0.0004	-	-	-0.584
18q12.2	18:32831349:C>CA	rs528812569	ZNF397	intron variant	-	-	0.0021	-	-0.122
		rs528812569	ZSCAN30	3' UTR variant, downstream gene variant	-	-	0.0021	-	-0.122
	18:32842714:T>C	rs181377293	ZNF397	downstream gene variant	-	-	0.0007	-	0.427
		rs181377293	ZSCAN30	3' UTR variant, downstream gene variant	-	-	0.0007	-	0.427

The 4q13.1 locus has been identified previously in a linkage study of familial colorectal cancer type X (in which Lynch-defining criteria are fulfilled without a variant identifies in a MMR gene) [105]. Under the peak at the 4q13.1 locus (Chr4:64,176,000-67,035,000), IT16 harbored a missense variant, rs146610448 (c.1004+1010A>G; V298A), in *TECRL*. Though the variant is rare and predicted to be damaging, this gene is not expressed in gastrointestinal tissue, and with no prior reports of association with malignancy this seems an unlikely candidate. IT10 harbored a variant, rs551436483, which locates to the 3' UTR repeat of a tyrosine kinase, *EPHA5*. The variant has a frequency of 0.002 in UK10K, but is not conserved and was not evaluable in Haploreg. Two VUS in *EPHA5* (c.242A>C) were recently identified in 2 cases of Barret's Oesophagus with a family history of CRC [363].

I could not identify any variants of interest at the 5p15.33 locus below the peak (Chr5:3,682,000-4,376,000), which is intergenic. The closest gene to the peak is *IRX1*, a known tumour suppressor gene in gastric and lung cancers which lies approximately 80kb upstream [364, 365]. The 5p15.33 locus has previously been highlighted in CRC GWAS [312].

On chromosome 7 (Chr7:126,127,000-128,381,000, 7q32.1), I identified 3 variants. Rs548961200, identified in IT16, is a 5' UTR variant of *ARF5*. The allele frequency is 0.0027 in UK10K; PhyloP score is 0.202. *ARF5* is one of 6 *ARF* genes, which are involved in endosomal trafficking, and is widely expressed. Knockdown of Arf5 has been shown to increase tumour growth and cell migration in murine models of glioma [366]. This variant was not evaluable in HaploReg v4.1, and so its functional impact is unclear. The second, rs142220343 (Chr7:127672115:C>G), identified in IT10, is intronic to *LRRC4*. This is a known to be a tumour suppressor in glioma [367], and was recently included as a methylation marker for pancreatic cancer detection [368]. However rs142220343 is common in UK10K, and so an unlikely candidate. The final variant, chr7:127945084:G>A, is a 3' UTR variant of *RBM28*. Upregulation of *RBM28* is seen in a number of different cancers, including colon cancer, and is associated with poorer prognosis in TCGA data [369]. *RBM28* promotes cell proliferation by impairing p53 transcriptional activity [369].

I identified 9 variants within the peak region at 11q23.3 (chr11:117,800,000-119,217,000). This locus shows LOH is some somatic CRC [370, 371], was previously identified as a suggestive locus in CRC linkage [111], and was recently identified as a CRC risk locus in GWAS [114].

IT16 harbored two different 3'-UTR variants of *MPZL3* (Chr11:118097941:T>C and Chr11:118098148:G>C). Neither was present in any reference dataset, but the sites are moderately conserved. *MPZL3* has previously been associated with lung cancer risk [372], and RNA expression was significantly increased in radio-resistant rectal cancer cell lines [373]. It is expressed in transverse colon at low levels.

A 3'-UTR variant in *IFT46*, rs546689601, was found in IT10. This has frequency of 0.002 in UK10K and is conserved (PhyloP 1.359). This gene is widely expressed, and functions in ciliary protein trafficking, but has no reported association with cancer, and was not evaluable in Haploreg. An additional variant was identified in IT16 in *UPK2*, but this urothelium-specific protein is unlikely to be relevant here.

Three variants in *TREH* were identified in IT16, one of which was predicted to be a frameshift variant, 11:118533570:CA>C, P27X. There are multiple transcripts of this gene, and in other transcripts this change was reported as an intron variant causing feature truncation. The other variants are intronic, and all cluster within 5bp. Examination the surrounding sequence shows that this is not a region of short repeats, and these may therefore represent sequencing or mapping error.

rs752017704 is a 3'-UTR variant, or in some transcripts a non-coding variant causing a retained intron, in *C2CD2L*. This variant has a frequency of <0.001 in gnomAD genomes and TOPMed, and PhyloP score of 0.006, but was not evaluable in Haploreg. This is a recently characterised endoplasmic reticulum protein, which transports phosphatidylinositol from the ER to the plasma membrane where is converted to phosphatidylinositol 4,5-bisphosphate (PI(4,5)P2)[374]. This activity replenishes PI(4,5)P2, which is phosphorylated to phosphatidylinositol-3,4,5-trisphosphate (PI(3,4,5)P3) by phosphoinositide 3-kinase (PI3K), and which then activates the AKT-mTOR pathway, promoting cell growth and proliferation [375]. The PI3K-AKT-mTOR pathway is implicated in the pathogenesis of a wide

range of cancers, and is well recognised as being overactivated in CRC [376]. Thus modification of *C2CD2L* expression could play a role in cancer pathogenesis.

An additional variant in IT10 is rs192316821, an intronic variant within *MCAM*. The variant has a frequency of 0.0005 in UK10K, but the site is not strongly conserved. High levels of *MCAM* expression in cancer-associated fibroblasts in CRC has recently been shown to be inversely associated with survival, whilst in mice, Mcam knockout in the stroma reduced tumeroid growth [377]. On examination in Haploreg this variant is in an ChromHMM inactive state (state 10 or 11) in colorectal samples in the 15-state model, but in a promoter state (3) in the 25-state model [339].

Below the 14q21.1 peak (Chr14:40,258,000-43,140,000) there were no variants of interest noted. The only gene under this peak is LRFN5, a calcium-dependent cell-adhesion molecule which regulates synapse development [378], previously associated with autism and major depressive disorder [123, 379], though it has been reported as a possible methylation-based marker in pancreatic cancer [380], and is expressed in colonic tissues. The 14q21.1 locus has not previously been reported in CRC though 14q22.2 was an early GWAS locus [247].

In the region of interest on chromosome 18 (18:31,062,000-33,320,000), which localises to a region spanning parts of 18q12.1-12.2, both individuals harbored 3' UTR variants in ZSCAN30. In IT10, rs528812569 is an insertion with a frequency of 0.002 in UK10K, and IT16 harbored rs181377293 (UK10K allele frequency 0.0006). There is no data on either variant in HaploReg v1.4. On the opposite strand is ZNF397, a ubiquitously expressed transcription activator [381], in which rs528812569 is an intronic or downstream variant. However neither gene is previously reported to be associated with cancer.

3.4.2.4 Follow-up of candidate variants

On examination of the top candidate variants and genes in sequenced cases and controls (from the Genomics England dataset and our own in-house sequencing data, Table 3.6), several genes were deprioritised. Candidate variants and truncating mutations were identified in multiple controls for *PLEKHM2*, *ZFAT*, *EIF5A2*,

Table 3.6: Replication of top linkage analysis candidate variants and genes in sequenced cases and controls

			Variant (N)		Truncating mutation (N)	
Locus	Variant	Gene	Cases	Controls	Cases	Controls
Family 2733						
1p36.21	rs376306853	PLEKHM2, D402G	1	4	0	3
1p22.1	rs367740686	GFI1, G374S	0	0	1	9
1q23	rs148814291	NTRK1, intronic	1	0	0	0
8q24.22	rs760117294	ZFAT, L[329/379/391]V	0	3	3	3
Family 3491						
3q26.2	rs557433728	EIF5A2, 3' UTR	0	0	0	4
Multiple Adenoma Families						
4q13.1	rs551436483	EPHA5, V298A	0	0	0	1
7q32.1	rs548961200	ARF5, 3' UTR	0	0	0	0
	rs551436483	LRRC4, intronic	0	0	0	1
	7:127945084:G>A	RBM28, 3' UTR	0	0	3	16
11q23.3	rs752017704	C2CD2L, 3'-UTR	0	0	0	1
	rs192316821	MCAM, intronic	0	0	0	6

RBM28, and *MCAM*. Truncating mutations in *GFI1* were also seen in one case and 9 controls, however on closer inspection in 7 of the controls the variants clustered within a 12bp region, and may therefore have been secondary to quality or mapping issues.

3.5 Discussion

In this chapter I have reported the discovery of new CRC loci through genome-wide association studies, and propose several new potential loci and putatively pathogenic variants identified in linkage analysis.

In my GWAS analysis of SCOT and PoBI datasets, the very high number of false positive results (i.e. type 1 errors), ultimately appeared to arise from an error introduced when combining the two datasets. My investigation of the issues in this GWAS highlight some key concerns around imputation-based GWAS analysis, and the importance of choice of control dataset.

I was initially concerned that imputation may have introduced bias (or error) to my analysis. Imputation quality for downstream analysis is enhanced by ensuring an appropriate reference panel is used, and removing SNPs with poor imputation scores. Imputation techniques have also evolved and improved considerably over

the last 15 years. The imputation reference panel used in my SCOT-PoBI GWAS had been used by our research group and collaborators for multiple other studies [115], and ought to have been appropriate to my European ancestry datasets.

Despite selection of appropriate reference data, inherent uncertainty in calling genotypes can introduce bias in imputation and subsequent GWAS analysis. This can be a particular issue when combining datasets typed on different arrays, when overlap in SNP content can be low (as I found here). In this case, SNPs may be imputed from a common SNP set, or separately from available SNPs from each array. When imputed from differing SNP sets, uncertainty will vary as a result of some SNPs being genotyped in one dataset, and imputed in another. Imputing from a common SNP set has been shown to reduce type 1 error [178], however the shared content in my analysis was insufficient to use this approach without significantly affecting imputation quality. Additionally, if the MAF in the case or control datasets differs significantly from the MAF in the imputation panel, spuriously divergent MAFs between the cases and controls can increase the type 1 error rate [382].

An alternative source of bias is confounding due to population stratification [315]. Ideally cases and controls should be from the same population, and representative of the population in which cases arise [383]. The effect of population structure can be minimised by matching controls to cases based on confounders which act as proxies for population structure, or by adjusting for population structure during analysis [383].

Restricting analyses to one population reduces the impact of population structure, but the issue may persist if there is significant variation in allele frequency or disease incidence within the population [383]. Differences in allele frequencies between populations can lead to false-positive associations in GWAS if the population substructure is unbalanced between cases and controls [271]. The impact of population stratification is positively correlated with sample size [384], and thus it is common in the most recent large-scale GWAS to adjust for population structure, using PCs, in the association analysis.

In their paper using haplotype-based methods to investigate population substructure within PoBI, Leslie et al. [242] were the first to demonstrate rich fine-scale genetic variation within a country. In studying Caucasian individuals from across the UK whose grandparents had all resided within 80Km of one another, they effectively studied DNA from the grandparents in these regions, prior to the population shifts of the 20th century (average year of grandparental birth was 1885). The observed population substructure coincided with geography across the UK [242]. However, PCA separated only Welsh and Orkney-based samples from the rest of the UK, whilst more refined analysis using hierarchical clustering (fineSTRUCTURE) defined 17 clusters. It is possible that fine-scale population structure observed in PoBI could be insufficiently accounted for by PCA, which might bias GWAS results. I was unable to find any other paper that had used PoBI as a GWAS control study (per Web of Science search for citations 20th June 2022), which was unexpected given that one of the stated purposes of the project was to provide a well-characterised control population for future disease studies.

Given time constraints and the identification of an alternative control dataset, I did not pursue my investigations of this further. The subsequent European CRC GWAS meta-analysis identified 31 new risk loci, 8 additional independent risk SNPs at these new or previously reported loci, and an additional 9 SNPs previously found in Asian GWAS [115]. Concurrently with publication of this paper, a large US-led consortium also published a CRC meta-analysis, which replicated 18 of the novel associations identified in our study, identified a further 5 independent signals at loci identified in our study, and reported a further 17 loci or conditionally independent signals beyond those seen in our analysis [222].

SNPs identified in the CRC meta-analysis presented mapped to regions with enhancer marks in colonic tissue, and 87% of SNPs were predicted to disrupt transcription factor binding motifs [115]. SNPs may also affect chromatin-looping interactions, which impact cis-regulatory transcriptional networks. These chromatin interactions confirmed the relationship between rs6983267 and *MYC*, and support rs1412834 action on *CDKN2B*, and rs12255141 on *TCF7L2*. Incorporation of

expression quantitative trait loci (eQTLs, loci associated with disease status based on RNA expression levels) with GWAS data supports candidate genes for a number of other CRC loci, for example rs4546885 and *LAMC1*, and rs12427600 and *SMAD9* [115].

Subsequently the data from this meta-analysis has been pooled with almost all existing CRC GWAS data, including 100,000 cancer cases and 150,000 controls of European and Asian descent. This has identified a further 50 independent risk associations, bringing the total to 205 [114]. Further multi-omic analysis with transcriptome- and methylome-wide association studies (TWAS and MWAS) identified a further 53 risk associations. Functional and pathway enrichment analysis in these studies highlights several key effector pathways in CRC, including Wnt, TGF-β/BMP, and Hippo. Fernandez-Rozadilla *et al.* identify over 150 plausible effector genes, and functional annotation of these highlights genes in relevant biological processes including RhoA/ROCK signalling genes linked to cell migration, Ras/RAF pathway genes controlling cell proliferation, genes associated with prostaglandin metabolism, and with roles in the cytoskeleton and ion transport.

My linkage analysis identified potential new CRC loci, and replicated two GWAS loci (3q26.2, 5p15.33), and the 4q13.1 and 11q23 loci identified previously in Lynch-like families [105, 111]. 1q23 overlaps with the 1q22–q24.2 locus identified in extended CRC families [110].

The top priority variant for further investigation is the substitution found in *GFI1*. Gfi1 was initially discovered in an experiment aimed to find pathways which made T-cells IL-2 independent [385]. It was subsequently found to be a common viral insertion site in retroviral tagging experiments which sought to identify novel oncogenes collaborating with Myc and Pim-1 genes in T-cell lymphomas in transgenic mice [386–388], and was identified as a transcriptional repressor.

Gfi1 is involved in multiple processes which could potentially drive carcinogenesis. In the TGF-β pathway it is recruited to *CDKN1A/p21* and *CDKN1B* promotors via MIZ-1, repressing transcription [389]. Notably *CDKN1A* was also identified as a candidate gene by Fernandez-Rozadilla *et al.*, whilst a *CDKN1B* mutation

identified through exome sequencing of a family with colorectal advanced adenoma, CRC, and stomach cancer, segregated with disease [390]. *GFI1* also regulates Toll-like receptor signalling pathways [391], and thus could also affect carcinogenesis via inflammatory pathways.

GFI1 has been identified as a somatic tumour suppressor gene in CRC. In murine models, loss of GFI1 increased adenoma count and resulted in larger adenomas in APC mutant ($Apc^{Min/+}$; $GFI1^{F/F}$; CDX2-cre) mice compared to mice with only *APC* knockout [342]. Downregulation of GFI1 has also been shown to promote metastatic spread of CRC [392]. Given its action in multiple pathways, it is perhaps unsurprising that *GFI1* appears to act as either an oncogene and or as tumour suppressor in other solid tumours including breast and prostate, medulloblastoma, oesophageal squamous cell carcinomas, and gastric cancers [393–396].

The location of the G374S mutation supports its potential role in disease. Germline mutations in the *GFI1* gene have been implicated in families with severe congenital neutropoenia [397]. In one family, a 4 month old boy with severe neutropoenia and his affected brother were found to have a heterozygous N382S substitution, which like the G374S mutation affects the 5th zinc finger. In vitro, this mutation abolished transcriptional repressor activity in a dominant negative manner, and recognition of the consensus binding site was abolished [397]. Follow-up of the G374S variant is now needed in the wider family, confirming segregation with disease.

Unsurprisingly, several regions strongly implicated in cancer pathogenesis were highlighted. The 7q32.1 locus was recognised in early studies of chromosomal alterations in a range of cancers as a fragile site, with frequent abnormalities [e.g. 398, 399, 400]. The variants highlighted in my multiple adenoma families were in genes plausibly involved in other malignancies. This locus contains multiple additional genes implicated in carcinogenesis. *GRM8* is a G-protein coupled receptor commonly mutated in squamous lung cancers [401], and identified as a potential tumour suppressor in endometrial cancer [402]. *PAX4* is a PAX-domain containing transcription factor, an oncogene upregulated in breast cancer cells and head and neck squamous cell carnimona tissue, which promotes migration and invasion in

murine models by reducing miR-144/451 levels, which act on ADAM protein family members [403]. *SND1* has also long been associated with CRC pathogenesis, is highly expressed in CRC, and is upregulated in abnormal crypts, suggesting an early role in pathogenesis [404]. It has varied roles including as a transcriptional co-activator, in pre-mRNA splicing, and translation [405].

A number of potentially interesting genes are found at the 8q24.3, a region of frequent copy number alteration in CRC [357, 358], however I did not find a potential variant underlying in the linkage peak in Family 3491. *HSF1* regulates the cellular stress response [357], and drives extracellular matrix remodelling in chronic intestinal inflammation, which has a central role in the transition to colon cancer [406]. Amplification of *PRL-3*, a tyrosine phosphatase, in this region has been associated with metastatic potential of CRC [407]. *WISP1*, a beta-catenin regulated growth factor, is an oncogene which is overexpressed in colonic tumours [408, 409].

Although the variants identified under the linkage peak at 18q12.1-12.2 were not strong candidates, this locus contains a number of other potentially interesting genes. Deletions of 18q have long been recognised in CRC [10], and have been associated with poor prognosis [410]. Another ubiquitously expressed zinc finger protein, *ZNF24*, represses *VEGF* transcription [411], and acts as a tumour suppressor gene in-vitro and in xenograft models in breast cancer, gastric cancer, non-small cell lung cancer, and thyroid cancer [411–414]. *DTNA* mRNA is significantly downregulated in early CRC compared to CIN [415]. *ASXL3* is one of 3 ASXL family proteins with functions epigenetic regulation through histone modification. The family is frequently mutated across a range of cancers [416]. *MAPRE2*, also known as *EB2*, encodes the RP1 protein [417]. It is one of three proteins in the EB1 family (the other two being *EB1* and *EB3*) which all have microtubule end-binding motifs, and bind to the APC (adenomatous polyposis coli) protein. *EB2* appears to be essential for apico-basal differentiation of epithelial cells [418], and to have roles in focal adhesion dynamics [419, 420] and mitotic progression and genome stability [421]. Upregulation of *MAPRE2* expression has been associated with poor outcomes in pancreatic cancer and HCC [422, 423]. In vitro, over-expression of *MAPRE2* in

HCC cell lines promoted cell proliferation, migration and invasion; *EB2* appeared to promote microtubule destabilisation, with subsequent induction of Src kinase, promoting ERK activation, which fed back to promote *EB2* expression [423].

One key issue with my linkage analysis is the paucity of overlap in the QC'd datasets on chromosome 12, which likely reflects the complexity of imputation around the major histocompatibility complex (MHC) region. This precluded analysis of linkage on this chromosome. In future analysis, this chromosome may need to be imputed separately and re-analysed. Candidate loci under suggestive linkage peaks will need to be followed up in independent datasets, with subsequent functional analysis if still considered likely candidates. Further linkage analysis of the planned phenotypic groups will increase power to identify causative loci.

4

The UK Biobank Cohort

In this chapter I describe the UK Biobank dataset, focusing particularly on colorectal cancer, and the cohorts used for modelling work presented in Chapters 5 and 6. I also describe my evaluation of data quality, and coding of modelling predictors used in subsequent chapters.

4.1 Background

In developing or validating a prediction model, a thorough understanding of the dataset used is essential [284, pp. 191-211]. A number of key factors and decisions around the type of dataset and how it is handled can introduce bias to a study and affect interpretation of results [424]. These include:

- Appropriate choice of study design (for example case control or cohort study data)
- Selection of participants, and exclusions from the dataset
- Size of the dataset, and the number of cases
- Availability and coding of outcome and predictors
- Quantity and handling of missing data

UK Biobank (UKB) is a prospective cohort study of over 500,000 participants across the UK, conceived to allow the prospective study of the determinants of a wide range of health outcomes [425]. It is a deeply phenotyped resource with extensive genetic information, biological measurements, imaging data, and linkage to external datasets such as cancer and death registry data.

Individuals aged between 40 and 69 years old and living within 20 miles of 22 recruitment centres were invited to participate between 2006 and 2010. The age range was chosen to provide a cohort in which baseline assessment could take place before the onset of many health conditions of interest, and in which there would be sufficient incident cases relatively early in follow-up to facilitate study. The size of the cohort was powered to nested case-control studies to detect odds ratios of 1.3-1.5 for studies of 5-10,000 cases. The population approached was intended to cover a rural-urban, socioeconomic and ethnic mix [425].

Participants attended recruitment centres for baseline assessment, completing touch screen questionnaires, brief interviews with a health professional, baseline biometric measurements and donating samples of blood, urine and saliva. Several repeated assessment rounds have been conducted in a subset of participants, to address regression-dilution bias (the tendency to underestimate associations between exposures and outcomes due to random errors in baseline measurements [426]).

UK Biobank participants gave permission for ongoing follow-up through linkage to national health records, including hospital inpatient data, primary care data, and death and cancer registry data. Notably, cancer and death registries are a separate entity in Scotland, compared to England and Wales, resulting in different censoring dates for Scottish participants when considering incident cancers.

The genetic resource includes imputed genotyped data on 97% of participants, whole genome sequencing data on 200,000 participants available since late 2021, and anticipated for the whole cohort in 2023.

As of August 2021, colorectal cancer (CRC) was the second most common incident cancer in UKB, based on linked registry data up to July 2019. Approximately 2% of these have occurred in participants under the age of 50, and 18% in the under

60's [427]. In both male and female UKB participants, CRC is the fourth leading cause of death (by sex), after ischaemic heart disease, lung cancer, and prostate, and breast cancer respectively (censored at 28th February 2021) [428].

4.1.1 Chapter outline

In this chapter I address these considerations and define and describe the modelling datasets used in Chapters 5 and 6. Differences between the modelling dataset and the population in which the model was initially derived (for validation studies), or the intended population of model use, have implications for how results may be interpreted, which will also be discussed at the end of this chapter.

4.2 Methods

4.2.1 Colorectal cancer case-finding

CRC cases were identified for this thesis in cancer registry, death registry and hospital inpatient records using ICD-9 and ICD-10 codes, and through self-reporting (see Table 4.1). Self-reported dates of cancer were recorded as fractional years, which I recoded to mid-year (30th June) for simplicity, whilst "Date uncertain or unknown" or "Preferred not to answer" were recorded as missing (resulting in one individual with no date recorded). Where an individual had more than one CRC (or the same cancer recorded in multiple data sources), the first instance of cancer was used. ICD-10 codes for anal cancers (C21*) were not included, as these have a different aetiology to CRC, and detection of these lesions is not the purpose of bowel cancer screening, although these were used in the development of QCancer-10.

4.2.2 Colorectal cancer incidence

Cancer incidence rates were calculated in UKB as a whole, by region (Table 4.2), and in my modelling datasets, and compared with data from the Office for National Statistics 2013 cancer registry data for England (equating to the approximate mid-point of available UKB follow-up).

Table 4.1: Case-finding sources in UKB

Source	UKB Field	Date Field	Encoding
Self Report	20001	20006	1020, 1022, 1023
Cancer Registry	40013, 40006	40005	ICD9: 153, 154.0, 154.1 ICD10: C18-C20
Death Registry	40001, 40002	40000	ICD9: 153, 154.0, 154.1 ICD10: C18-C20
Hospital Inpatient Data	41270	41280	ICD9: 153, 154.0, 154.1 ICD10: C18-C20

Table 4.2: UK Biobank region coding

Region	UK Biobank assessment centre (field 54)
North West	Bury, Cheadle, Liverpool, Manchester, Stockport (pilot)
North East	Middlesborough, Newcastle
Yorkshire and Humber	Leeds, Sheffield
East Midlands	Nottingham
West Midlands	Birmingham, Stoke
South West	Bristol
South East	Oxford, Reading
London	Barts, Croydon, Hounslow
Wales	Cardiff, Swansea, Wrexham
Scotland	Edinburgh, Glasgow

When cancer incidence is described as a crude rate (e.g. cases per 100,000 of the population) differences in underlying population structure make comparisons across different cohorts or populations near-meaningless. More accurate comparisons can be made with age-specific incidence rates (ASRs) and directly standardised incidence rates (DSRs), both of which I calculated here. In calculating ASRs, the population is divided into age bands, and rate for age band k is

$$(Case_k/Population_k) \times 100000$$

Age-specific rates were calculated between 40 and 80 years in 5 year age bands in males and females separately.

Fry et al. [266] conducted a commonly cited analysis of ASRs in UKB over a shorter duration of follow-up than was available for this thesis (to September 30, 2014), which they compared to ONS data for 2012. They excluded prevalent cases (aside from non-melanomatous skin cancer, ICD-10 code C44, which is also excluded

from ONS data), and used first cancer of any cancer type as their outcome, prior to analysing data by cancer type. As a result, an individual with, for example, incident breast cancer prior to CRC is not included in the CRC outcomes. In contrast the ONS reference population is the mid-year population estimate for England, which does not exclude individuals with prior cancers. Additionally, the numerator for each cancer type is all incident cases of a given cancer in the year under consideration. Thus this is not necessarily an equitable comparison.

I recreated the ASRs as calculated by Fry et al. [266] (details of which I confirmed with the paper's corresponding author) for the whole UKB cohort, with follow-up right censored at the end of available complete registry follow-up (31^{st} October 2015 for Scottish participants, 31^{st} March 2016 for English and Welsh participants). This was then compared with an analysis which was perhaps more reflective of cancer registration statistics, in which prevalent cases were not removed, and the outcome of first incident cancer registry CRC was used.

In DSRs, which I calculated for UKB overall and my Integrated Modelling Cohort, age- and sex-specific rates are multiplied by the number of individuals in a standard population. The Office for National Statistics uses the European Standard Population 2013, in which the reference age bands range from 0 to >90. In order to compare rates for individuals of screening (and UKB) age, I used European Standard Population 2013 age bands 40-80 as my standard population (Table 4.3). For all rate calculations, I calculated follow-up in person years as the period from recruitment to the date of either first cancer diagnosis (for the analysis following Fry et al. [266], excluding non-melanomatous skin cancers and carcinoma-in-situ), or first CRC diagnosis, death, or end of follow-up (whichever came first).

4.2.3 Coding of QCancer-10 predictors

I matched coding of QCancer-10 predictors in UKB as closely as possible. Data were gathered in UKB through touchscreen questionnaire and subsequent interviews with a study nurse. A proportion of UKB participants have returned for longitudinal

Table 4.3: European Standard Population age 40-80 years

Age Band	Population
40+ thru 45	7000
45+ thru 50	7000
50+ thru 55	7000
55+ thru 60	6500
60+ thru 65	6000
65+ thru 70	5500
70+ thru 75	5000
75+ thru 80	4000

Table 4.4: UK Biobank ethnicity coding mapped to QCancer-10 variable categories

Q Cancer	UK Biobank
White/not recorded	White, British, Irish, Any other white background, Prefer not to answer, Do not know, Missing
Indian	Indian
Pakistani	Pakistani
Bangladeshi	Bangladeshi
Other Asian	Asian or Asian British, Any other Asian background
Caribbean	Caribbean
Black African	African
Chinese	Chinese
Other	Black or Black British, Any other Black background, Mixed, White and Black Caribbean, White and Black African, White and Asian, Any other mixed background, Other ethnic group

follow-up and have multiple instances of this data; I used information gathered at baseline only.

4.2.3.1 Ethnicity

Ethnic background from UKB was recoded to align with Qcancer-10 coding (Table 4.4).

4.2.3.2 Smoking status

Data on smoking status in UKB was obtained through a series of touchscreen questions. The summary 'Smoking status' variable in UKB (field 20116) is coded as 'Never', 'Previous', 'Current', and 'Prefer not to answer'. Participants were also asked if they currently smoked tobacco ('Yes, on most days', 'Only occasionally', 'No', 'Prefer not to answer'), and then led through a set of more detailed questions

Table 4.5: Coding of smoking status in UKB to match QCancer-10

QCancer-10	Smoking frequency	Additional coding
Non-smoker	Never	NA
Ex-smoker	Previous	NA
Light smoker	Current	Frequency = 'Only occasionally' OR cigarettes/day <10
Moderate smoker	Current	cigarettes/day = 10-19
Heavy smoker	Current	cigarettes/day >= 20
Missing	Missing, Prefer not to answer	NA

NA - not applicable

depending on their response, including smoking frequency and number of cigarettes smoked. I used a combination of these fields (Table 4.5) to map to QCancer-10 smoking categories. Previous UKB validation of QCancer-10 by Usher-Smith et al. [172] used the summary coding of smoking status rather than this more granular coding.

4.2.3.3 Alcohol intake

Alcohol data were collected in UKB using the screening question 'About how often do you drink alcohol?' (field 1558), with 7 potential answers from 'Daily or almost daily' to 'Never', and 'Prefer not to answer'. Those drinking more than once a week were then asked to give weekly intakes (by the glass, pint or measure) of different alcohol types; those drinking less frequently were asked to give monthly intakes. Participants were asked to average intake over the year if intake was highly variable.

I converted alcohol intake by the glass/pint by alcohol type to units, using NHS Choices Live Well Alcohol Units (NHS [429], as in Usher-Smith et al. [172], see Table 4.6), assuming a standard wine glass size of 175ml (2.1 units), and taking the average of low and higher strength beers (2.5 units). Fortified wine was taken to be 20% ABV and therefore to have 1 unit in a standard 50ml glass. The 'Other alcoholic drinks' question in UKB touchscreen information gives the example of alcopops (1.5 units/drink) which was used to convert to units. I then summed total monthly or weekly intake, and dividing by 30 or 7 (for monthly or weekly intake respectively).

Table 4.6: Units per glass of different alcohol types used to calculate daily alcohol intake

Alcohol Type	Glass size (ml)	ABV (%)	Units per glass
Wine	175	12	2.1
Beer	568	4.4	2.5
Spirits	25	40	1
Fortified wine	50	20	1
Other (e.g. alcopops)	275	5.5	1

In QCancer-10, alcohol intake is coded as a categorical variable with levels defined by units/day intake, and I therefore converted UKB levels to this categorisation.

4.2.3.4 Family history

Family history for a range of conditions was queried for father, mother and siblings (specifically for blood relations only). Conditions are arbitrarily divided into two groups, with each group queried in one question. Bowel cancer is queried alongside Parkinson's disease, severe depression, lung cancer, prostate cancer, and breast cancer. Participants could select more than one of these options, or 'Do not know' or 'Prefer not to answer'. In QCancer-10, a positive family history was recorded if any family member had a history of bowel cancer; participants were assumed not to have any family history if this was not coded. As this is a fairly broad inclusion category, I coded family history in UKB as positive if any of father, mother or sibling had a history of bowel cancer, as negative if there was data available for any one of father mother or sibling for category 2 illnesses, and missing only if information was unavailable for all relatives. I conditioned sibling family history on whether the participants had answered yes to having brothers or sisters (i.e. missing data for family history in siblings was discounted if they did not have any siblings).

4.2.3.5 Previous medical history

I obtained prior medical history from self-report at enrolment interview with a UKB nurse (as in Usher-Smith et al. [172], see Table 4.7). Diabetes included participants coded as either 'diabetes' or 'type 2 diabetes' because the majority of cases were coded simply as 'diabetes' rather than by type; 'type 1 diabetes' and 'gestational diabetes' were not included.

Table 4.7: UK Biobank codes for self-reported previous medical history

Condition	UK Biobank Code
Diabetes	1223, 1220
Ulcerative colitis	1463
Bowel polyps	1460
Breast cancer	1002
Uterine cancer	1040
Ovarian cancer	1039
Cervical cancer	1041
Lung cancer	1001
Blood cancer	1047, 1048, 1050, 1051, 1052, 1053, 1055, 1056, 1058
Oral cancer	1004, 1005, 1006, 1010, 1011, 1012, 1015, 1077, 1078, 1079

4.2.3.6 Age, body mass index, and Townsend score

For continuous variables, I used age at enrolment to UKB (equal to the study entry time point, field 21003), body mass index (BMI, field 21001), Townsend deprivation score (field 189) as recorded at UKB enrolment visits. The Townsend deprivation score is a postcode level assessment of social deprivation which incorporates unemployment, non-car ownership, non-house ownership, and household overcrowding [269], where high values indicate increasing levels of deprivation.

4.2.4 Genetic quality control

The genotyped dataset in UKB and its quality control is described in detail in Bycroft et al. [267]. Samples were genotyped on two arrays: 49,950 on the Applied Biosystems UK BiLEVE Axiom Array (807,411 markers; Affymetrix, now Thermo Fischer Scientific), and 438,427 on the Applied Biosystems UK Biobank Axiom Array (825,927 markers). Over 95% content of the Biobank Array is shared with the BiLEVE Array. Genotyping was performed in 106 sequential batches, with protocols in place to minimise batch effects, with a custom genotype calling pipeline and QC optimised for biobank genotyping (Affymetrix).

Following QC, phasing was carried out using SHAPEIT3 using 1000 Genomes Phase 3 as a reference panel. Imputation was then carried out using IMPUTE4 firstly to the Haplotype Reference Consortium (HRC) dataset as the main reference panel, and secondarily with merged UK10K and 1000 Genomes phase 3 reference

panels. The two imputed datasets were then combined. This resulted in an imputed dataset of 487,442 individuals with 93,095,623 autosomal SNPs, short insertion-deletions (indels), and other structural variants. SNP annotation was based on GRCh37 assembly of the human genome [267].

I followed standard per-person QC measures [271] on the whole imputed dataset, excluding individuals with sex-chromosome aneuploidy, and mismatch between self-reported and genotyped sex (which can indicate sample handling errors). I excluded individuals with high levels of kinship by including only those included in the UKB PCA analysis in the derivation dataset, as this included a computed maximal dataset in which those with 3rd degree kinship or greater were removed [267]. Outliers for heterozygosity and missingness were not present in the imputed UKB data.

Section 5.2 describes further quality control measures specific to PRS modelling datasets.

4.2.5 Outliers and missingness

4.2.5.1 Handling of missing data

I evaluated per-variable and per-person missingness. Levels and patterns of missingness were examined by recording the proportion of missing data per variable, plotting cluster plots of association between missingness for different predictors and the outcome CRC, and plotting paired plots of missing and observed values across variables.

I considered possible mechanisms of missingness, focusing on predictors with higher levels of missingness and those likely to have a particularly strong effect in the models. Missingness is categorised into three mechanisms: missing completely at random (MCAR), in which there is no systematic difference between missing and observed data, and the individuals are a random subsample of the full dataset; missing at random (MAR) in which there are systematic differences which can be explained by the observed information; and missing not at random (MNAR) in which systematic differences cannot be explained by the data collected or are dependent on the missing values [284, pp.117-118].

4.2.5.2 Handling of outliers

Outliers for continuous variables were defined as values outside 3 times the IQR above the third quartile or below the first quartile [430], and I manually inspected these. I retained biologically plausible values, whilst implausible values were set to missing. Further discussion of handling of outliers is given in Chapters 5 and 6.

4.2.6 Sample size

4.2.6.1 Polygenic risk score

The accuracy of a polygenic risk score in predicting individual risk is driven by the size of the base GWAS cohort [116], which is finite and defined by availability of data. The precision of estimation of SNP effect sizes and the proportion of genetic variation explained by the score (both of which are improved by increasing GWAS sample size) influence PRS performance. I explore this further in Chapter 5.

4.2.6.2 Integrated modelling

In developing and validating a prediction model, the number of predictors potentially included in the model is key to sample size requirements. Additionally, it is the number of outcome events, rather than total sample, which decides the effective sample size [168].

The maximum number of predictors included in the integrated ((i.e. combined genetic and non-genetic) model for each sex was calculated as follows for the QCancer-10 predictors: 1 for each degree of freedom of each categorical variable (alcohol intake = 5; ethnicity = 8; smoking = 4); 1 parameter for each fractional polynomial term for age, and 2 parameters for each interaction term calculated (confirmed with the R-package's author, Joie Ensor); 1 for each continuous variables (BMI, Townsend Deprivation Score); 1 for each boolean predictor; 1 for the QCancer-10 score; and 1 for the PRS. This totalled 34 parameters for models for males, and 33 parameters for females.

Historically, 10-20 events per predictor (EPP) was recommended as the minimum number of events in model development [431], which would require 680 cases for

males and 660 for females. I had intended to use a Geographic validation cohort for all modelling on this basis, and as seen in the PRS datasets, I would have been well-powered for model development.

However, estimation of required sample size for risk model development is an area of active research and during the course of my research, more nuanced criteria were proposed by Riley et al. [432] to calculate the required sample size for prediction model development. This derives a minimum number of participants and events relative to the number of parameters included in the model to fulfil three criteria. These are:

1. Small optimism in predictor effect estimates as defined by a global shrinkage factor of >0.9
2. Small absolute difference of <0.05 in the model's apparent and adjusted Nagelkerke's R^2
3. Precise estimation of the overall risk or rate in the population.

The first two criteria reduce overfitting of the model, whilst the third criteria ensures that the model will reliably predict mean risk in the population studied, which is essential to predicting individual risk.

For a time-dependent model, the sample size calculation uses estimates of expected Cox-Snell R^2 (R^2_{CSapp}), the number of predictors included in model development (including all candidate predictors prior to any selection process), the shrinkage factor applied (recommended at 0.9), the estimated event rate in the population, the time point of interest, and the mean follow-up in the cohort studied. The calculation is available as R package 'pmsampsize' [433] which I used to calculate sample size requirements for male and female models across 5-8 years of follow-up.

Where unavailable in the literature, Riley et al. [432] describe a series of steps to obtain R^2_{CSapp} from reported statistics. I used the C-statistics from open cohort QCancer-10 validation in UKB performed by Usher-Smith et al. [172], which were 0.7 for the male model and 0.65 in the female model.

R^2_{CSapp} is derived as follows: firstly Royston's D statistic is calculated from the C-statistic using the following formula from Jinks, Royston, and Parmar [434]

$$D = 5.5(C - 0.5) + 10.26(C - 0.5)^3$$

From the D statistic, R^2_D is calculated as:

$$R^2_D = \frac{\frac{\pi}{8}D^2}{\frac{\pi^2}{6} + \frac{\pi}{8}D^2}$$

This is then used as a proxy for Royston's measure of explained variation, $R^2_{Royston\_app}$, (as the two are sufficiently similar [435]) which is used to calculate $R^2_{O'Quigley\_app}$

$$R^2_{O'Quigley\_app} = \frac{-\frac{\pi^2}{6}R^2_{Royston\_app}}{(1 - \frac{\pi^2}{6})R^2_{Royston\_app} - 1}$$

This is then used to calculate the likelihood ratio, using the cohort size and number of events (E, here taken from Usher-Smith et al. [172] open cohort)

$$LR = -E\ln(1 - R^2_{O'Quigley\_app})$$

.

From this the Cox-Snell generalised definition of the apparent R2 is calculated as

$$R^2_{CS\_app} = 1 - exp\left(\frac{-LR}{n}\right)$$

If the measure of performance were taken from a model development study the $R^2_{CS\_app}$ would then need shrinking to adjust for optimism, however the C-statistic used here was from a validation study so is not required.

This corresponds to an R^2_{CSapp} of 0.0032154 in men and 0.0011695 in women. I used the follow-up and crude incidence rate for individuals available for integrated model development (i.e. with QC'd genetic data and complete QCancer-10 data, excluding the 30,000 individuals used to develop the PRS): follow-up of 7.08 and

Table 4.8: Sample size calculations for integrated models for men and women

	Males	Females
Sample size calculation inputs		
Predictors	34	33
R^2_{CSapp}	0.003215	0.00117
Follow-up (years)	7.08	7.10
Crude event rate (per person year)	0.001386823	0.0008704994
Sample size requirements		
Required event count	933	1569
Required sample size	94996	253780
EPP	27.43	47.53

7.10 years for men and women respectively, and incidence rate of 0.001386823 and 0.0008704994 respectively.

Table 4.8 shows the sample sizes required for the integrated model. While for men the EPP needed is close to the previously recommended 20 EPP, that required for women is far higher. Sample size requirements are met for men, but the sample size and cases available in the Integrated Modelling Cohort for women (n = 238,496, including 1458 cases) do not reach calculated requirements. The implications of this are discussed at the end of this chapter.

4.2.7 Definition of modelling cohorts

For PRS modelling, I divided the UKB dataset into a Derivation Cohort, a Geographical Validation Cohort, and a Minority Ethnic Validation Cohort. In selecting the Geographic cohort I considered guidance for the size of datasets in validation studies. This is an active area of research, however Collins, Ogundimu, and Altman [436] recommend a minimum of 100, and ideally 200 cases. In addition I considered which region would best demonstrate the models' portability to other settings.

For integrated modelling, I used all individuals with complete data, excluding 30,000 individuals used to train the PRS.

4.2.8 Descriptive statistics

I reported the demographics of the modelling cohorts by sex and by CRC status, and compared the demographics of the integrated modelling cohort with that of the QResearch cohort used to derive the QCancer-10 model. I compared differences between categorical variables in the two datasets using χ^2 tests, and continuous predictors using two-sample t-tests.

4.3 Colorectal cancer in UK Biobank

I removed 29,124 prevalent cancers present in the UKB dataset from the analysis of CRC incidence following Fry et al.'s methodology, leaving 473,403 individuals contributing follow-up time, and 30,076 incident cancers, of which 3,204 were CRC. Age-standardised CRC incidence for in the whole UKB cohort, based on registry data, was 108.3 and 73.9 cases per 100,000 years at risk for men and women respectively, compared to 127.8 and 80.7 cases per 100,000 years at risk in ONS data.

Figure 4.1 shows the age-specific incidence rate (ASR) of CRC in UKB biobank, using the approach used by Fry et al. [266], and without removing prevalent cases, compared to ONS data for 2013. The incidence rates by this method are broadly as described by Fry et al., and both datasets show similar rates to ONS up to 70 years of age, after which ASRs are lower in UKB. In women aged 75-80 the incidence is lower using Fry et al.'s method, potentially due to the exclusion of a number of CRC cases who may previously have had breast cancer. Table 4.9 shows DSRs by region of UKB, based on registry cases, demonstrating that the crude rates and DSRs can differ significantly.

4.4 Demographics of UKB

Table 4.10 shows the demographics of the whole potential prospective modelling cohort (n = 499455, prevalent CRC cases (n = 3033) excluded).

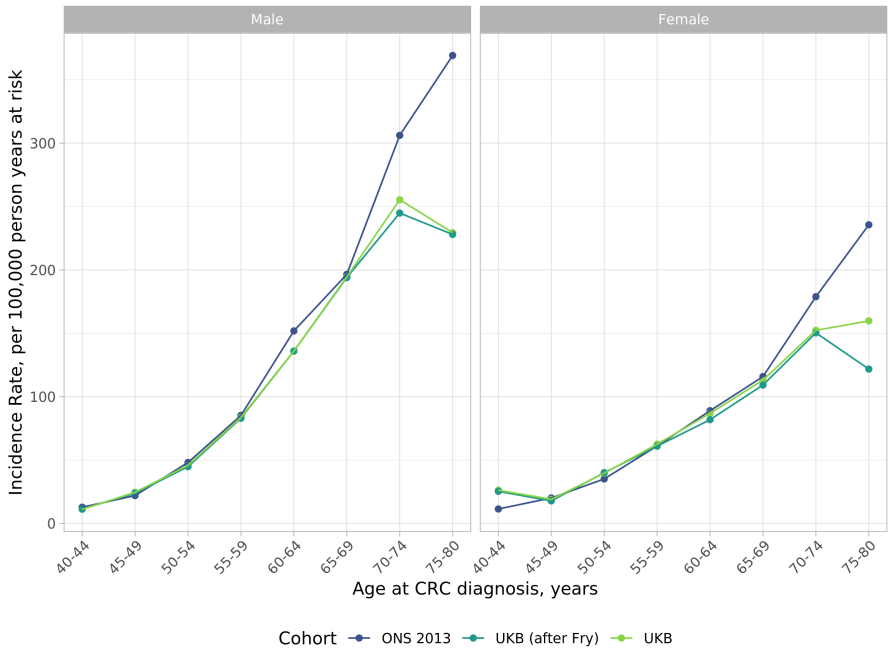

Figure 4.1: Age-specific CRC incidence rates in UKB compared to ONS data

Table 4.9: Directly standardised CRC rates by region, per 100,000 person years at risk, based on linked cancer registry data

Region	Crude Rate	DSR (95% CI)
Males		
West Midlands	115	91 (73-118)
South East	118	95 (78-118)
London	119	89 (77-121)
South West	122	116 (89-156)
North East	127	107 (88-133)
Scotland	132	102 (84-128)
East Midlands	132	108 (81-147)
Yorkshire and Humber	138	131 (107-162)
Wales	138	107 (81-151)
North West	144	117 (102-135)
Females		
London	66	66 (45-104)
Yorkshire and Humber	73	69 (54-91)
East Midlands	76	62 (46-90)
North West	77	71 (59-88)
Wales	79	64 (44-105)
North East	82	71 (58-90)
South West	85	92 (67-128)
West Midlands	88	78 (61-107)
Scotland	93	84 (67-107)
South East	100	84 (69-105)

DSR - Directly standardised rate, CI - confidence interval

Table 4.10: Demographics of the UKB cohort by sex

		Males	Females
Region	East Midlands	15298 (6.7)	18371 (6.8)
	London	30396 (13.4)	38060 (14)
	North East	26371 (11.6)	31526 (11.6)
	North West	36554 (16.1)	41830 (15.4)
	Scotland	15760 (6.9)	19824 (7.3)
	South East	19178 (8.4)	24052 (8.8)
	South West	18868 (8.3)	23912 (8.8)
	Wales	9443 (4.2)	11243 (4.1)
	West Midlands	21723 (9.6)	22901 (8.4)
	Yorkshire and Humber	33832 (14.9)	40313 (14.8)
Ethnicity	White/not recorded	215121 (94.6)	257402 (94.6)
	Indian	3003 (1.3)	2933 (1.1)
	Pakistani	1118 (0.5)	716 (0.3)
	Bangladeshi	159 (0.1)	74 (0.0)
	Other Asian	996 (0.4)	857 (0.3)
	Caribbean	1637 (0.7)	2855 (1.0)
	Black African	1701 (0.7)	1677 (0.6)
	Chinese	581 (0.3)	989 (0.4)
	Other	3107 (1.4)	4529 (1.7)
Follow-up (years)	-	7.08 (1.34)	7.09 (1.31)
Alcohol	Non-drinker	14472 (6.4)	25889 (9.5)
	Trivial drinker	48955 (21.5)	109497 (40.3)
	Light drinker	66332 (29.2)	87095 (32.0)
	Moderate drinker	69467 (30.5)	42795 (15.7)
	Heavy drinker	17115 (7.5)	4374 (1.6)
	Very heavy drinker	10323 (4.5)	1646 (0.6)
	Missing	759 (0.3)	736 (0.3)
Smoking	Non-smoker	110840 (48.7)	161336 (59.3)
	Ex-smoker	86721 (38.1)	84928 (31.2)
	Light smoker	11009 (4.8)	10158 (3.7)
	Moderate smoker	6958 (3.1)	8402 (3.1)
	Heavy smoker	10474 (4.6)	5708 (2.1)
	Missing	1421 (0.6)	1500 (0.6)
Age	-	58.0 (14.0)	57.0 (13.0)
	Missing	0 (0.0)	0 (0.0)
Townsend	-	-2.1 (4.3)	-2.1 (4.1)
	Missing	293 (0.1)	325 (0.1)

Table 4.10: Demographics of the UKB cohort by sex *(continued)*

		Males	Females
BMI	-	27.3 (5.1)	26.1 (6.3)
	Missing	1639 (0.7)	1448 (0.5)
Height	-	176.0 (9.0)	162.0 (9.0)
	Missing	1380 (0.6)	1146 (0.4)
Weight	-	84.3 (17.6)	69.1 (17.0)
	Missing	1408 (0.6)	1351 (0.5)
Family history of CRC	No	196133 (86.2)	239997 (88.2)
	Yes	21638 (9.5)	24773 (9.1)
	Missing	9652 (4.2)	7262 (2.7)
Diabetes	No	211552 (93)	262440 (96.5)
	Yes	15513 (6.8)	9294 (3.4)
	Missing	358 (0.2)	298 (0.1)
Colorectal polyps	No	226334 (99.5)	271014 (99.6)
	Yes	711 (0.3)	708 (0.3)
	Missing	378 (0.2)	310 (0.1)
Ulcerative colitis	No	225858 (99.3)	270343 (99.4)
	Yes	1187 (0.5)	1379 (0.5)
	Missing	378 (0.2)	310 (0.1)
Breast cancer	No	-	260218 (95.7)
	Yes	-	11165 (4.1)
	Missing	-	649 (0.2)
Uterine cancer	No	-	270189 (99.3)
	Yes	-	1194 (0.4)
	Missing	-	649 (0.2)
Ovarian cancer	No	-	270572 (99.5)
	Yes	-	811 (0.3)
	Missing	-	649 (0.2)
Cervical cancer	No	-	269398 (99)
	Yes	-	1985 (0.7)
	Missing	-	649 (0.2)
Lung cancer	No	226643 (99.7)	-
	Yes	149 (0.1)	-
	Missing	631 (0.3)	-
Blood cancer	No	225436 (99.1)	-
	Yes	1356 (0.6)	-
	Missing	631 (0.3)	-
	No	226216 (99.5)	-

Table 4.10: Demographics of the UKB cohort by sex *(continued)*

		Males	Females
Oral cancer	Yes	576 (0.3)	-
	Missing	631 (0.3)	-
Imputed genetic data	Yes	220923 (97.1)	262403 (96.5)
	Missing	6500 (2.9)	9629 (3.5)

4.5 Handling of missingness and outliers

4.5.1 Missing data

Table 4.10 gives the number and proportion of missing values for each QCancer-10 predictor. There was no missing data for region or ethnicity (as coded in QCancer-10). Missingness was <1% for most predictors aside from family history and availability of imputed genetic data, both of which were <5% missing.

Figure 4.2 shows associations between missingness for QCancer-10 predictors in UK Biobank. Predictor missingness is not associated with outcome, and overall the fraction of missing data in common between the predictors is extremely low. Where missing values cluster (for example between previous medical history of cancers), these are generally where these values are asked in a single question, or coded in one field, in the UKB touchscreen questionnaire. When availability of imputed genetic data was also considered, association between missing variables remained low. The clustering reflects answers given in response to the same question or stem of questions in the UKB assessment centre.

This is supported by matrix plots of missingness (Figures 4.3 and 4.4) in which the values for predictors are plotted across the x-axis, against missing predictors across the y-axis. Shaded parts of the bar represent the proportion of individuals with a given predictor status on the x-axis with a missing value for the y-predictor. For both sexes, missingness is correlated across previous medical history predictors (i.e. of those with NA values across the x-axis, a high proportion also have missing values in other medical history predictors). This is to be expected to a degree as medical

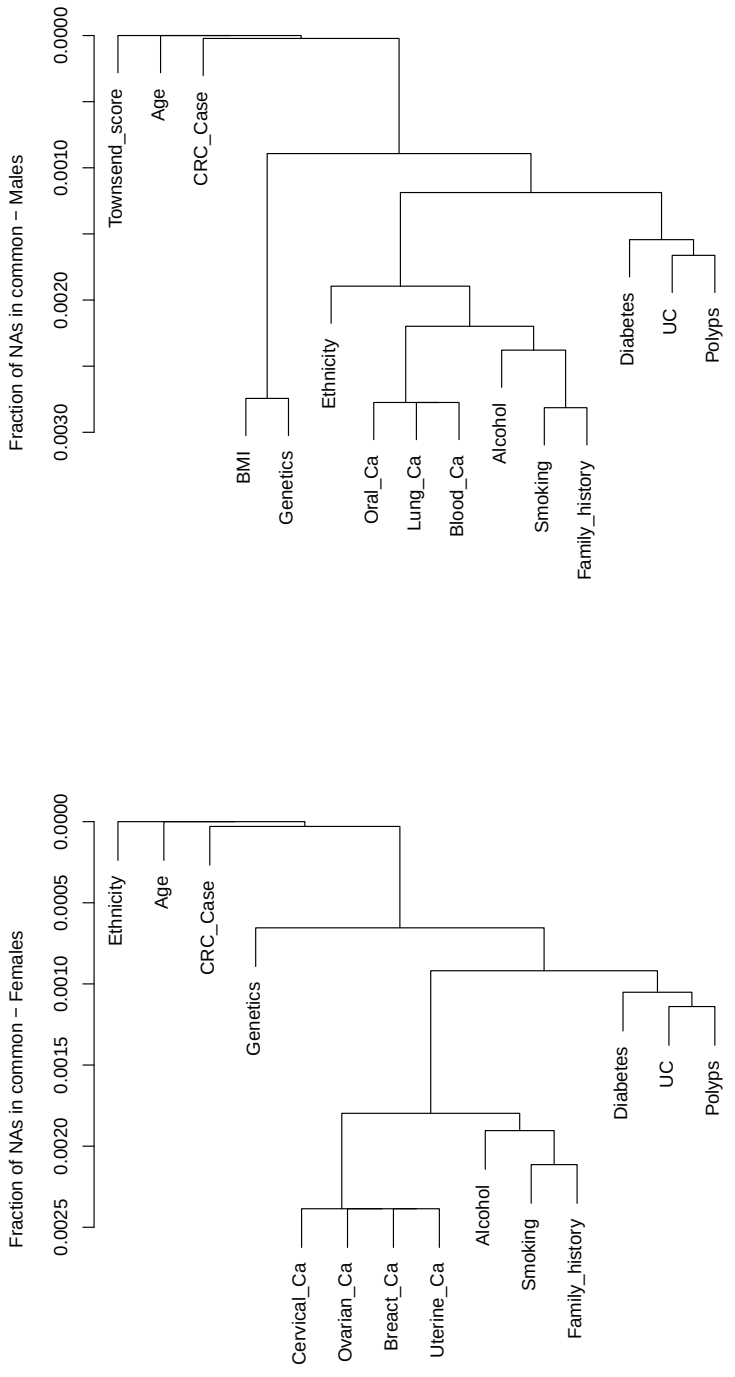

Figure 4.2: Cluster plots of missingness in UK Biobank for males (above) and females (below)

history questioning is grouped, but there is also a high proportion of missingness for smoking and alcohol history for these individuals, which are queried separately.

The matrix plots show that in men and women, missingness for family history (row 8) is greater in those from minority ethnicities, as is missingness for genetic data in women. In men, Townsend deprivation score tended to be higher in those with missing data (indicating higher levels of deprivation) for most other predictors. Focusing on family history and imputed genetic data, which had the highest levels of missingness, statistical tests of relationships between family history and other predictors (Table 4.11 supports a significant correlation with ethnicity and Townsend score, and indicate a relationship to smoking status (higher in non-smokers) and alcohol intake (higher in moderate-heavy drinkers). Individuals with missing family history tended to be older. Missingness for family history for females showed similar patterns to males, though there was no significant difference in age (Table 4.13).

Table 4.12 shows similar patterns of relatedness of missing genetic data to ethnicity and socioeconomic status in men, but not with age. Current smokers, non-drinkers, and diabetics were more likely to have missing genetic data. The latter may be related to higher missingness in minority ethnicities, in whom prevalence of diabetes is higher, and regular alcohol consumption lower. The same patterns are seen for women (Table 4.14).

The prospective nature of case identification here means that predictors should not be directly correlated with missingness of outcome (individuals with less than 5 years of available follow-up, marked as NA in "Case", n = 900). As expected, we see that there is no correlation with other variables (across row 1), though those with less than 5 years of follow-up tended to be younger.

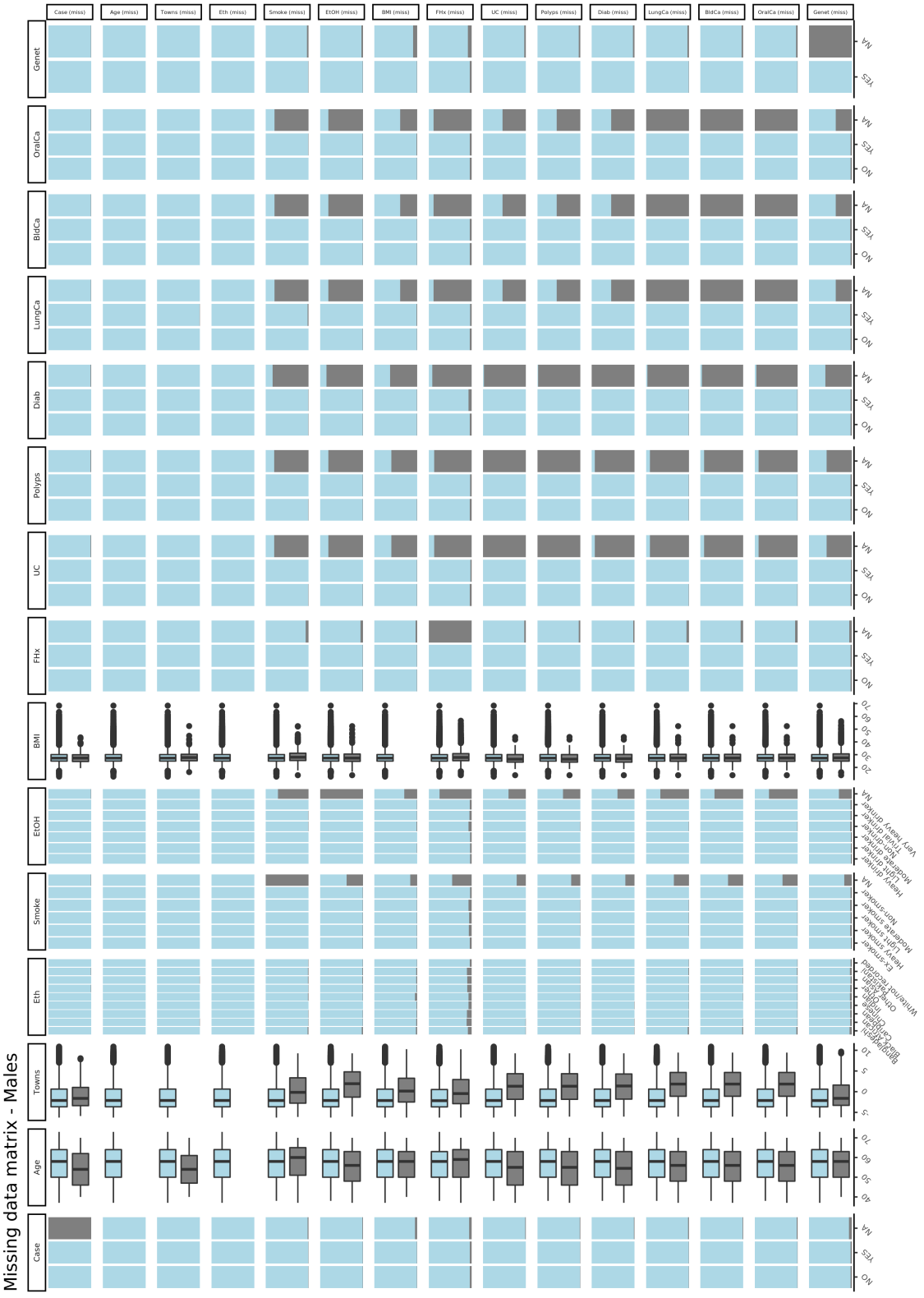

Figure 4.3: Missingness matrix for males in UKB

Table 4.11: Missing data analysis for family history in males, n = 224,827

Missing data analysis: Family history		Not missing	Missing	p
Case	NO	215289 (95.8)	9538 (4.2)	0.963
	YES	2099 (95.8)	92 (4.2)	
	(Missing)	383 (94.6)	22 (5.4)	
Age	Mean (SD)	56.7 (8.2)	57.1 (8.6)	<0.001
Townsend	Mean (SD)	-1.3 (3.1)	0.2 (3.6)	<0.001
Ethnicity	Bangladeshi	145 (91.2)	14 (8.8)	<0.001
	Black African	1510 (88.8)	191 (11.2)	
	Caribbean	1463 (89.4)	174 (10.6)	
	Chinese	542 (93.3)	39 (6.7)	
	Indian	2788 (92.8)	215 (7.2)	
	Other	2799 (90.1)	308 (9.9)	
	Other Asian	893 (89.7)	103 (10.3)	
	Pakistani	995 (89.0)	123 (11.0)	
	White/not recorded	206636 (96.1)	8485 (3.9)	
Smoking status	Ex-smoker	83150 (95.9)	3571 (4.1)	<0.001
	Heavy smoker	9782 (93.4)	692 (6.6)	
	Light smoker	10396 (94.4)	613 (5.6)	
	Moderate smoker	6477 (93.1)	481 (6.9)	
	Non-smoker	107185 (96.7)	3655 (3.3)	
	(Missing)	781 (55.0)	640 (45.0)	
Alcohol intake	Heavy drinker	16539 (96.6)	576 (3.4)	<0.001
	Light drinker	63987 (96.5)	2345 (3.5)	
	Moderate drinker	67276 (96.8)	2191 (3.2)	
	Non-drinker	13401 (92.6)	1071 (7.4)	
	Trivial drinker	46504 (95.0)	2451 (5.0)	
	Very heavy drinker	9873 (95.6)	450 (4.4)	
	(Missing)	191 (25.2)	568 (74.8)	
BMI	Mean (SD)	27.8 (4.2)	28.4 (4.6)	<0.001
UC	NO	216570 (95.9)	9288 (4.1)	0.037
	YES	1153 (97.1)	34 (2.9)	
	(Missing)	48 (12.7)	330 (87.3)	
Polyps	NO	217035 (95.9)	9299 (4.1)	0.281
	YES	688 (96.8)	23 (3.2)	
	(Missing)	48 (12.7)	330 (87.3)	
Diabetes	NO	203294 (96.1)	8258 (3.9)	<0.001
	YES	14446 (93.1)	1067 (6.9)	
	(Missing)	31 (8.7)	327 (91.3)	
Lung cancer	NO	217555 (96.0)	9088 (4.0)	0.843
	YES	144 (96.6)	5 (3.4)	
	(Missing)	72 (11.4)	559 (88.6)	
Blood cancer	NO	216402 (96.0)	9034 (4.0)	0.566
	YES	1297 (95.6)	59 (4.4)	
	(Missing)	72 (11.4)	559 (88.6)	
Oral cancer	NO	217148 (96.0)	9068 (4.0)	0.765
	YES	551 (95.7)	25 (4.3)	
	(Missing)	72 (11.4)	559 (88.6)	
Genetics	YES	211843 (95.9)	9080 (4.1)	<0.001
	NO	5928 (91.2)	572 (8.8)	

BMI - body mass index; UC - ulcerative colitis

Table 4.12: Missing data analysis for imputed genetic data in males, n = 224,827

Missing data analysis: Genetics		Not missing	Missing	p
Case	NO	218416 (97.1)	6411 (2.9)	0.901
	YES	2130 (97.2)	61 (2.8)	
	(Missing)	377 (93.1)	28 (6.9)	
Age	Mean (SD)	56.7 (8.2)	56.5 (8.3)	0.014
Townsend	Mean (SD)	-1.3 (3.1)	-0.7 (3.3)	<0.001
Ethnicity	Bangladeshi	151 (95.0)	8 (5.0)	<0.001
	Black African	1642 (96.5)	59 (3.5)	
	Caribbean	1571 (96.0)	66 (4.0)	
	Chinese	561 (96.6)	20 (3.4)	
	Indian	2876 (95.8)	127 (4.2)	
	Other	2987 (96.1)	120 (3.9)	
	Other Asian	968 (97.2)	28 (2.8)	
	Pakistani	1065 (95.3)	53 (4.7)	
	White/not recorded	209102 (97.2)	6019 (2.8)	
Smoking status	Ex-smoker	84414 (97.3)	2307 (2.7)	<0.001
	Heavy smoker	10087 (96.3)	387 (3.7)	
	Light smoker	10662 (96.8)	347 (3.2)	
	Moderate smoker	6719 (96.6)	239 (3.4)	
	Non-smoker	107869 (97.3)	2971 (2.7)	
	(Missing)	1172 (82.5)	249 (17.5)	
Alcohol intake	Heavy drinker	16668 (97.4)	447 (2.6)	<0.001
	Light drinker	64624 (97.4)	1708 (2.6)	
	Moderate drinker	67726 (97.5)	1741 (2.5)	
	Non-drinker	13935 (96.3)	537 (3.7)	
	Trivial drinker	47429 (96.9)	1526 (3.1)	
	Very heavy drinker	10009 (97.0)	314 (3.0)	
	(Missing)	532 (70.1)	227 (29.9)	
BMI	Mean (SD)	27.8 (4.2)	28.1 (4.7)	<0.001
Family History	NO	190799 (97.3)	5334 (2.7)	0.843
	YES	21044 (97.3)	594 (2.7)	
	(Missing)	9080 (94.1)	572 (5.9)	
UC	NO	219606 (97.2)	6252 (2.8)	0.140
	YES	1163 (98.0)	24 (2.0)	
	(Missing)	154 (40.7)	224 (59.3)	
Polyps	NO	220078 (97.2)	6256 (2.8)	1.000
	YES	691 (97.2)	20 (2.8)	
	(Missing)	154 (40.7)	224 (59.3)	
Diabetes	NO	205798 (97.3)	5754 (2.7)	<0.001
	YES	14988 (96.6)	525 (3.4)	
	(Missing)	137 (38.3)	221 (61.7)	
Lung cancer	NO	220386 (97.2)	6257 (2.8)	0.488
	YES	143 (96.0)	6 (4.0)	
	(Missing)	394 (62.4)	237 (37.6)	
Blood cancer	NO	219229 (97.2)	6207 (2.8)	0.003
	YES	1300 (95.9)	56 (4.1)	
	(Missing)	394 (62.4)	237 (37.6)	
Oral cancer	NO	219970 (97.2)	6246 (2.8)	0.880
	YES	559 (97.0)	17 (3.0)	
	(Missing)	394 (62.4)	237 (37.6)	

BMI - body mass index; UC - ulcerative colitis

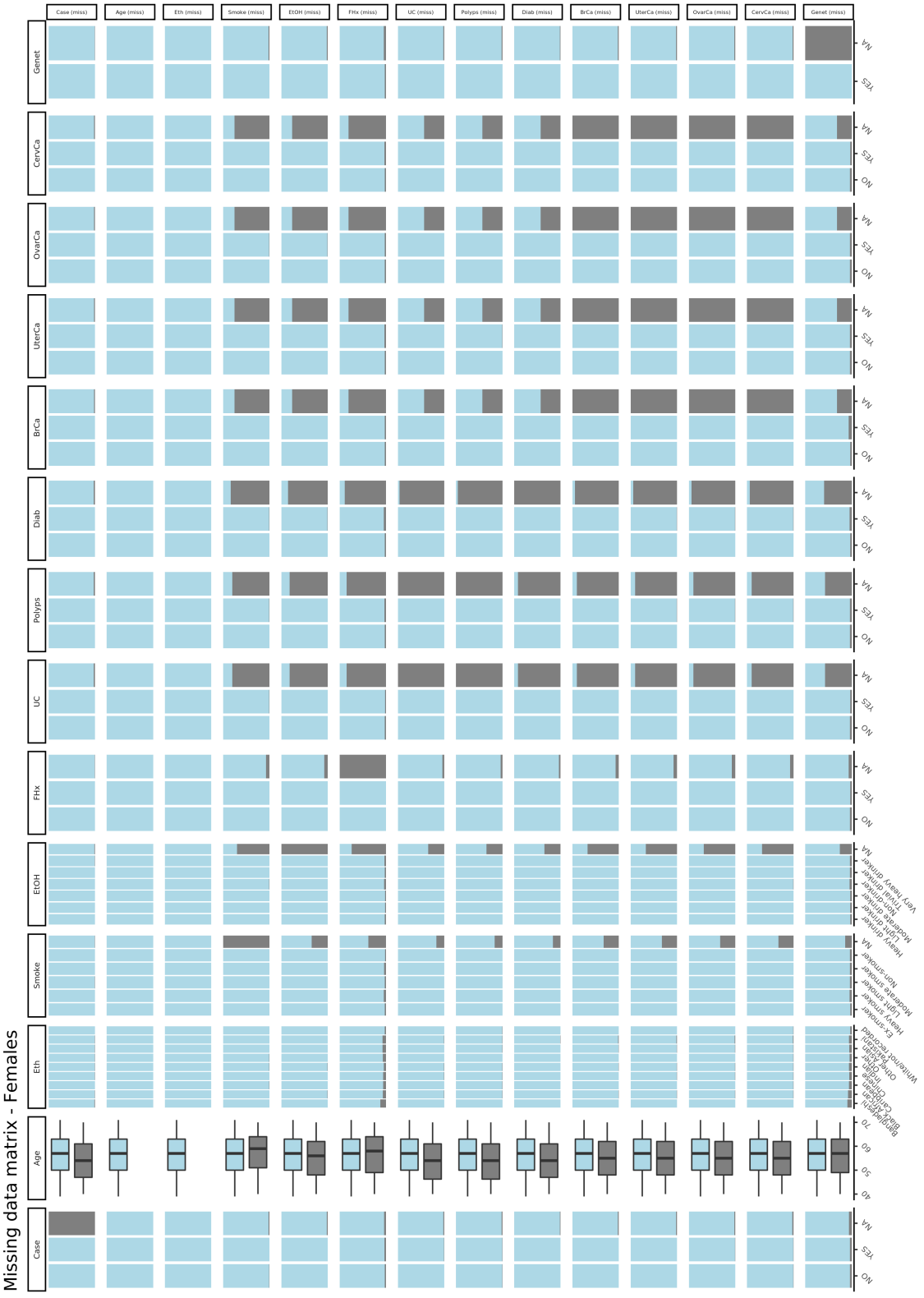

Figure 4.4: Missingness matrix for females in UKB

Table 4.13: Missing data analysis for family history in females, n = 269,867

Missing data analysis: Family history		Not missing	Missing	p
Case	NO	262678 (97.3)	7189 (2.7)	0.234
	YES	1622 (96.8)	53 (3.2)	
	(Missing)	470 (95.9)	20 (4.1)	
Age	Mean (SD)	56.3 (8.0)	56.5 (8.5)	0.040
Ethnicity	Bangladeshi	65 (87.8)	9 (12.2)	<0.001
	Black African	1563 (93.2)	114 (6.8)	
	Caribbean	2689 (94.2)	166 (5.8)	
	Chinese	926 (93.6)	63 (6.4)	
	Indian	2804 (95.6)	129 (4.4)	
	Other	4232 (93.4)	297 (6.6)	
	Other Asian	794 (92.6)	63 (7.4)	
	Pakistani	664 (92.7)	52 (7.3)	
	White/not recorded	251033 (97.5)	6369 (2.5)	
Smoking status	Ex-smoker	82794 (97.5)	2134 (2.5)	<0.001
	Heavy smoker	5431 (95.1)	277 (4.9)	
	Light smoker	9864 (97.1)	294 (2.9)	
	Moderate smoker	8041 (95.7)	361 (4.3)	
	Non-smoker	157715 (97.8)	3621 (2.2)	
	(Missing)	925 (61.7)	575 (38.3)	
Alcohol intake	Heavy drinker	4252 (97.2)	122 (2.8)	<0.001
	Light drinker	85438 (98.1)	1657 (1.9)	
	Moderate drinker	41902 (97.9)	893 (2.1)	
	Non-drinker	24802 (95.8)	1087 (4.2)	
	Trivial drinker	106593 (97.3)	2904 (2.7)	
	Very heavy drinker	1593 (96.8)	53 (3.2)	
	(Missing)	190 (25.8)	546 (74.2)	
UC	NO	263376 (97.4)	6967 (2.6)	0.607
	YES	1347 (97.7)	32 (2.3)	
	(Missing)	47 (15.2)	263 (84.8)	
Polyps	NO	264037 (97.4)	6977 (2.6)	0.438
	YES	686 (96.9)	22 (3.1)	
	(Missing)	47 (15.2)	263 (84.8)	
Diabetes	NO	255901 (97.5)	6539 (2.5)	<0.001
	YES	8835 (95.1)	459 (4.9)	
	(Missing)	34 (11.4)	264 (88.6)	
Breast cancer	NO	253755 (97.5)	6463 (2.5)	0.963
	YES	10889 (97.5)	276 (2.5)	
	(Missing)	126 (19.4)	523 (80.6)	
Uterine cancer	NO	263490 (97.5)	6699 (2.5)	0.066
	YES	1154 (96.6)	40 (3.4)	
	(Missing)	126 (19.4)	523 (80.6)	
Ovarian cancer	NO	263853 (97.5)	6719 (2.5)	1.000
	YES	791 (97.5)	20 (2.5)	
	(Missing)	126 (19.4)	523 (80.6)	
Cervical cancer	NO	262732 (97.5)	6666 (2.5)	0.001
	YES	1912 (96.3)	73 (3.7)	
	(Missing)	126 (19.4)	523 (80.6)	
Genetics	YES	255610 (97.4)	6793 (2.6)	<0.001
	NO	9160 (95.1)	469 (4.9)	

UC - ulcerative colitis

Table 4.14: Missing data analysis for imputed genetic data in females, n = 269,867

Missing data analysis: Genetics		Not missing	Missing	p
Case	NO	260326 (96.5)	9541 (3.5)	0.866
	YES	1614 (96.4)	61 (3.6)	
	(Missing)	463 (94.5)	27 (5.5)	
Age	Mean (SD)	56.3 (8.0)	55.9 (8.1)	<0.001
Ethnicity	Bangladeshi	67 (90.5)	7 (9.5)	<0.001
	Black African	1545 (92.1)	132 (7.9)	
	Caribbean	2692 (94.3)	163 (5.7)	
	Chinese	937 (94.7)	52 (5.3)	
	Indian	2765 (94.3)	168 (5.7)	
	Other	4314 (95.3)	215 (4.7)	
	Other Asian	815 (95.1)	42 (4.9)	
	Pakistani	674 (94.1)	42 (5.9)	
	White/not recorded	248594 (96.6)	8808 (3.4)	
Smoking status	Ex-smoker	82120 (96.7)	2808 (3.3)	<0.001
	Heavy smoker	5461 (95.7)	247 (4.3)	
	Light smoker	9758 (96.1)	400 (3.9)	
	Moderate smoker	8089 (96.3)	313 (3.7)	
	Non-smoker	155683 (96.5)	5653 (3.5)	
	(Missing)	1292 (86.1)	208 (13.9)	
Alcohol intake	Heavy drinker	4224 (96.6)	150 (3.4)	<0.001
	Light drinker	84262 (96.7)	2833 (3.3)	
	Moderate drinker	41451 (96.9)	1344 (3.1)	
	Non-drinker	24807 (95.8)	1082 (4.2)	
	Trivial drinker	105533 (96.4)	3964 (3.6)	
	Very heavy drinker	1580 (96.0)	66 (4.0)	
	(Missing)	546 (74.2)	190 (25.8)	
UC	NO	260937 (96.5)	9406 (3.5)	0.717
	YES	1334 (96.7)	45 (3.3)	
	(Missing)	132 (42.6)	178 (57.4)	
Polyps	NO	261590 (96.5)	9424 (3.5)	0.700
	YES	681 (96.2)	27 (3.8)	
	(Missing)	132 (42.6)	178 (57.4)	
Diabetes	NO	253413 (96.6)	9027 (3.4)	<0.001
	YES	8870 (95.4)	424 (4.6)	
	(Missing)	120 (40.3)	178 (59.7)	
Breast cancer	NO	251518 (96.7)	8700 (3.3)	<0.001
	YES	10442 (93.5)	723 (6.5)	
	(Missing)	443 (68.3)	206 (31.7)	
Uterine cancer	NO	260810 (96.5)	9379 (3.5)	0.746
	YES	1150 (96.3)	44 (3.7)	
	(Missing)	443 (68.3)	206 (31.7)	
Ovarian cancer	NO	261180 (96.5)	9392 (3.5)	0.653
	YES	780 (96.2)	31 (3.8)	
	(Missing)	443 (68.3)	206 (31.7)	
Cervical cancer	NO	260040 (96.5)	9358 (3.5)	0.674
	YES	1920 (96.7)	65 (3.3)	
	(Missing)	443 (68.3)	206 (31.7)	

UC - ulcerative colitis

Table 4.15: Number of missing predictors per person in UK Biobank (n, %)

Missing predictors (N)	Males	Females
QCancer-10 predictors		
0	214829 (94.63)	263062 (96.88)
1	11201 (4.93)	7675 (2.83)
2	320 (0.14)	135 (0.05)
3	75 (0.03)	26 (0.01)
4	38 (0.02)	92 (0.03)
5	13 (0.01)	33 (0.01)
6	199 (0.09)	17 (0.01)
7	37 (0.02)	254 (0.09)
8	13 (0.01)	3 (0.00)
9	96 (0.04)	6 (0.00)
10	196 (0.09)	239 (0.09)
11	1 (0.00)	-
QCancer-10 predictors and imputed genetic data		
0	209338 (92.21)	253977 (93.53)
1	15994 (7.05)	16446 (6.06)
2	976 (0.43)	445 (0.16)
3	110 (0.05)	29 (0.01)
4	42 (0.02)	83 (0.03)
5	13 (0.01)	41 (0.02)
6	190 (0.08)	19 (0.01)
7	37 (0.02)	227 (0.08)
8	18 (0.01)	28 (0.01)
9	90 (0.04)	6 (0.00)
10	26 (0.01)	86 (0.03)
11	183 (0.08)	155 (0.06)
12	1 (0.00)	-

Taken in combination my evaluations of missingness suggest that predictors are likely to be MAR rather than MCAR. It is possible that some QCancer-10 predictors might be MNAR (i.e. influenced by the value that is missing), for example very heavy drinkers or smokers might omit these questions for fear of judgement, but testing for MNAR requires examination of the missing data. This is not possible here, and so MNAR cannot be completely excluded.

Evaluation of per-person missingness (Table 4.15) shows that for QCancer-10 predictors, ~95% of males and females had complete data, which fell slightly when genetic data was considered, with few individuals having more than one missing predictor. In light of these findings, I used complete case analysis for my modelling.

4.5.2 Outliers

On examination of the distribution of continuous predictors, there were no outlying values for age or Townsend deprivation score. UKB aimed to recruit individuals aged 40-69, however there are a small number of participants aged between 37-39 and 70-73, who are not outliers by the criterion chosen, and I retained these individuals in my modelling cohort. Figure 4.5 shows distributions of BMI in males and females. I identified participants with outlying values for BMI, height or weight, and visually inspected the values; one female was an extreme outlier for BMI and weight, and this datapoint was set to missing. Remaining values were retained.

4.6 Definition and description of modelling cohorts

4.6.1 PRS Cohorts

Figure 4.6 details the per-person QC exclusions (explained in further detail in Section 5.2) and formation of the PRS modelling cohorts used in Chapter 5. The Derivation dataset consisted of individuals of white-British ancestry, identified through self-reported ethnicity and PCA by UKB [267], with imputed genetic data passing QC measures, who were recruited in England and Wales.

I chose to use a cohort of UKB participants recruited in Scotland as my Geographic Validation Cohort. This met case number requirements for model validation (Table 4.16), and also represents a population with different demographics to the remaining majority English cohort. The NHS, and cancer screening provision, is devolved and thus healthcare also differs in this Validation Cohort. For example, prescriptions are free, and bowel screening has started at 50 years of age since 2009. This validation cohort comprised Scottish participants with European ancestry (UK Biobank self-reported ethnicities of “British”, “Irish”, “White”, and “Any other white background”) passing QC. Participants for the Minority Ethnic Validation Cohort included individuals recruited from any centre, of self-reported minority ethnicity and passing QC.

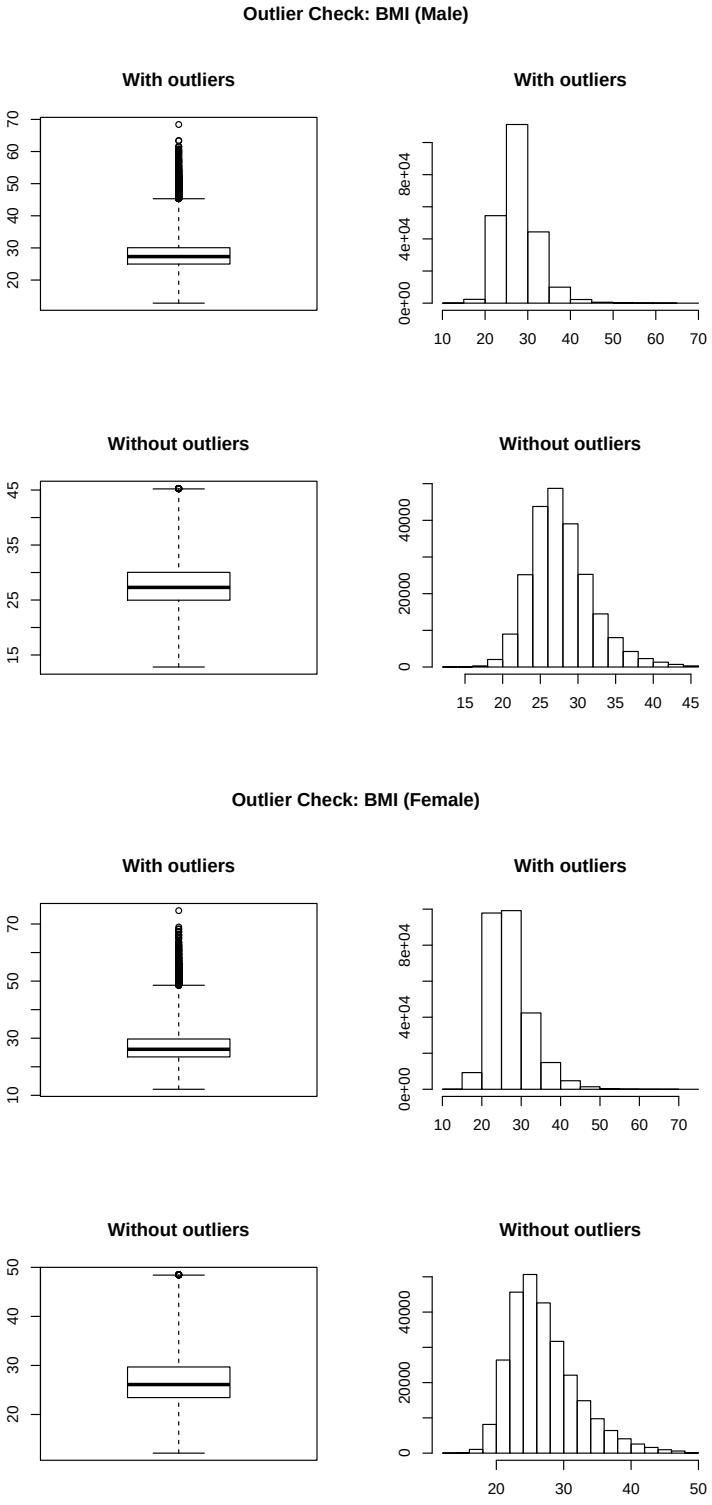

Figure 4.5: Boxplots and histograms of BMI in UKB

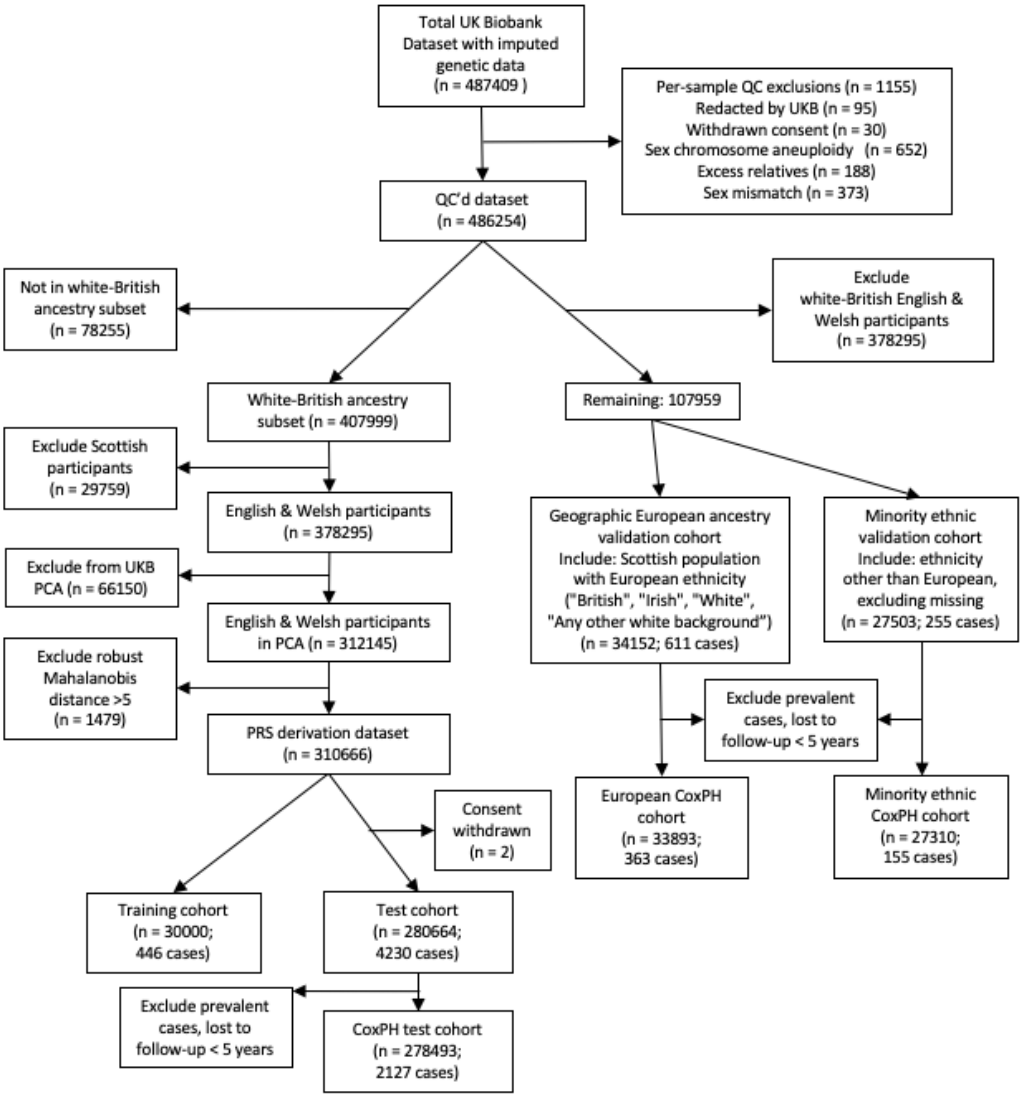

Figure 4.6: UKB participant flow diagram

Table 4.16: Cancer cases (n, %) by region in UK Biobank for PRS modelling (including prevalent and incident cases)

Region	Overall	Male	Female
Wales	301 (1.5)	184 (2)	117 (1.1)
East Midlands	484 (1.5)	260 (1.7)	224 (1.2)
South West	587 (1.4)	313 (1.7)	274 (1.2)
Scotland	611 (1.8)	330 (2.1)	281 (1.5)
West Midlands	651 (1.5)	376 (1.8)	275 (1.2)
South East	667 (1.6)	336 (1.8)	331 (1.4)
London	841 (1.3)	463 (1.6)	378 (1)
North East	902 (1.6)	525 (2)	377 (1.2)
Yorkshire and Humber	1052 (1.4)	611 (1.8)	441 (1.1)
North West	1165 (1.6)	693 (2)	472 (1.2)

4. UK Biobank

For PRS training and evaluation of performance I used a nested case-control dataset, in which prevalent and incident CRC cases were included. I also evaluated PRS performance in a prospective cohort, with prevalent CRC cases and individuals lost-to-follow-up within 5 years of enrolment excluded. Follow-up for all prospective models began at date of enrolment, and was censored at the earliest of date of incident CRC, loss to follow-up, death, or end of available registry follow-up (see Section 4.2.2). Table 4.17 show basic demographics for the different PRS modelling cohorts.

Table 4.17: Basic demographics of PRS modelling cohorts

	Training		Test		Geographic Validation		Minority Ethnic Validation	
	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases
Male (n, %)	13751 (46.5)	254 (57.0)	127823 (46.2)	2425 (57.3)	14851 (44.3)	330 (54.0)	12746 (46.8)	128 (50.2)
Female (n, %)	15803 (53.5)	192 (43.0)	148611 (53.8)	1805 (42.7)	18690 (55.7)	281 (46.0)	14502 (53.2)	127 (49.8)
Age (mean, SD)	56.82 (8.01)	61.64 (6.10)	56.84 (7.99)	61.41 (6.15)	56.31 (8.05)	61.00 (6.51)	52.75 (8.25)	58.25 (7.97)
Age (min-max)	40-70	40-70	39-72	40-70	40-70	40-70	39-72	40-70

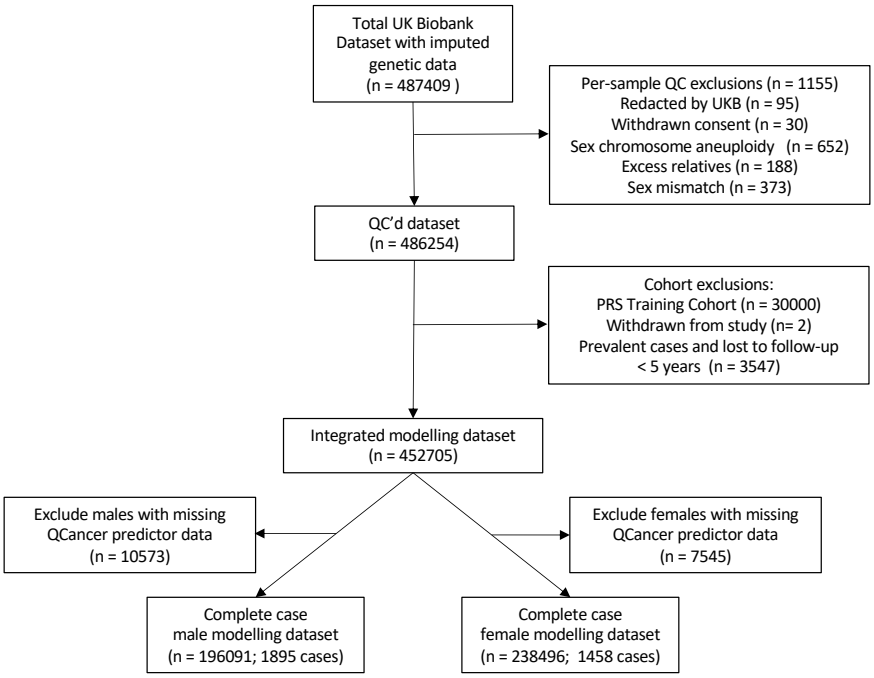

Figure 4.7: UKB participant flow diagram for Integrated Modelling Cohorts

4.6.2 Integrated modelling cohorts

Quality control exclusions and participant selection used to derive the Integrated Modelling Cohort are described in Figure 4.7. For validation of QCancer-10 and integrated modelling and I included all individuals with imputed genetic data passing basic QC (i.e. with high levels of kinship, sex-chromosome aneupoloidy, and mismatch between self-reported and genotyped sex), and excluded the 30,000 individuals used as the training PRS dataset (Section 5.2), prevalent CRC cases, and individuals lost to follow-up within 5 years of enrolment. Censoring dates were as defined for the PRS cohorts.

Table 4.18 describes the demographics of the Integrated Modelling Cohort by sex. CRC incidence was 118.0 CRCs per 100,000 years follow-up in men and 79.3 in women. Comparison of the demographics of the Integrated Modelling Cohort and the QCancer-10 derivation cohort (Table 4.19) demonstrates significant differences between the two datasets.

Table 4.18: Demographics of the Integrated Modelling Cohort. Brackets show percentages unless otherwise indicated

		Males		Females	
		Whole Cohort	Cases	Whole Cohort	Cases
Numbers		196091	1895	238946	1458
Median (IQR) Follow-up (years)		7.08 (1.34)	3.7 (3.49)	7.1 (1.29)	3.84 (3.39)
Median (IQR) age (years)		58 (13)	63 (8)	57 (13)	61 (10)
Region	East Midlands	13254 (6.8)	127 (6.7)	16175 (6.8)	88 (6.0)
	London	25843 (13.2)	203 (10.7)	33080 (13.9)	150 (10.3)
	North East	22789 (11.6)	221 (11.7)	27688 (11.6)	174 (11.9)
	North West	30259 (15.4)	325 (17.2)	35278 (14.8)	203 (13.9)
	Scotland	14690 (7.5)	173 (9.1)	18729 (7.9)	150 (10.3)
	South East	16812 (8.6)	165 (8.7)	21367 (9.0)	179 (12.3)
	South West	16467 (8.4)	150 (7.9)	21086 (8.8)	136 (9.3)
	Wales	8150 (4.2)	90 (4.7)	9942 (4.2)	63 (4.3)
	West Midlands	18530 (9.4)	154 (8.1)	19783 (8.3)	128 (8.8)
	Yorkshire and Humber	29297 (14.9)	287 (15.1)	35368 (14.8)	187 (12.8)
Ethnicity	White/not recorded	185813 (94.8)	1848 (97.5)	224316 (94.6)	1399 (96.0)
	Indian	2510 (1.3)	11 (0.6)	2601 (1.1)	12 (0.8)
	Pakistani	903 (0.5)	1 (0.1)	616 (0.3)	4 (0.3)
	Bangladeshi	132 (0.1)	0 (0.0)	61 (0.0)	0 (0.0)
	Other Asian	841 (0.4)	2 (0.1)	748 (0.3)	2 (0.1)
	Black African	1397 (0.7)	5 (0.3)	1412 (0.6)	6 (0.4)
	Caribbean	1363 (0.7)	8 (0.4)	2498 (1.0)	10 (0.7)
	Chinese	516 (0.3)	2 (0.1)	865 (0.4)	5 (0.3)
	Other ethnic group	2616 (1.3)	18 (0.9)	3980 (1.7)	20 (1.4)
Median (IQR) Townsend deprivation index		-2.18 (4.19)	-2.33 (4.19)	-2.17 (4.09)*	-2.38 (3.96)*
Median (IQR) BMI (kg/m2)		27.28 (5.04)	27.92 (5.11)	26.08 (6.23)*	26.42 (6.00)*
Smoking status	Non-smoker	97088 (49.5)	739 (39.0)	142569 (59.8)	820 (56.2)
	Ex-smoker	75100 (38.3)	935 (49.3)	74934 (31.4)	525 (36.0)
	Light smoker	9361 (4.8)	84 (4.4)	8885 (3.7)	43 (2.9)
	Moderate smoker	5816 (3.0)	43 (2.3)	7235 (3.0)	43 (2.9)
	Heavy smoker	8726 (4.4)	94 (5.0)	4873 (2.0)	27 (1.9)
Alcohol intake	Non-drinker	11985 (6.1)	89 (4.7)	22415 (9.4)	171 (11.7)
	Trivial drinker	41810 (21.3)	335 (17.7)	96085 (40.3)	591 (40.5)
	Light drinker	57817 (29.5)	521 (27.5)	76942 (32.3)	433 (29.7)
	Moderate drinker	60694 (31.0)	624 (32.9)	37830 (15.9)	234 (16.0)
	Heavy drinker	14960 (7.6)	205 (10.8)	3797 (1.6)	25 (1.7)
	Very heavy drinker	8825 (4.5)	121 (6.4)	1427 (0.6)	4 (0.3)
Previous medical history	Ulcerative colitis	1053 (0.5)	17 (0.9)	1211 (0.5)	12 (0.8)
	Colorectal polyps	616 (0.3)	11 (0.6)	612 (0.3)	6 (0.4)
	Diabetes	12893 (6.6)	184 (9.7)	7885 (3.3)	62 (4.3)
	Breast cancer	-	-	9448 (4.0)	71 (4.9)
	Uterine cancer	-	-	1030 (0.4)	16 (1.1)
	Ovarian cancer	-	-	724 (0.3)	11 (0.8)
	Cervical cancer	-	-	1711 (0.7)	10 (0.7)
	Lung cancer	125 (0.1)	1 (0.1)	-	-
	Blood cancers	1146 (0.6)	10 (0.5)	-	-
	Oral cancer	483 (0.2)	12 (0.6)	-	-
Family history of CRC		19505 (9.9)	266 (14.0)	22252 (9.3)	169 (11.6)

Table 4.19: Characteristics of the UKB Integrated Modelling Cohort compared with the QCancer-10 derivation cohort

		Males			Females		
		IMC	QCancer-10 derivation	P	IMC	QCancer-10 derivation	P
Number		196091	2447866		238946	2495899	
Age (years), mean (SD)		56.7 (8.2)	44.3 (14.8)		56.3 (8.0)	44.9 (15.9)	
Ethnicity	White/not recorded	185813 (94.8)	2 231 641 (91.2)	<0.001	224316 (94.6)	2 271 520 (91.0)	<0.001
	Indian	2510 (1.3)	42 771 (1.7)		2601 (1.1)	37 773 (1.5)	
	Pakistani	903 (0.5)	17 169 (0.7)		616 (0.3)	16 893 (0.7)	
	Bangladeshi	132 (0.1)	17 169 (0.7)		61 (0.0)	13 170 (0.5)	
	Other Asian	841 (0.4)	24 494 (1.0)		748 (0.3)	27 750 (1.1)	
	Caribbean	1397 (0.7)	37 003 (1.5)		1412 (0.6)	40 742 (1.6)	
	Black African	1363 (0.7)	18 553 (0.8)		2498 (1.0)	23 920 (1.0)	
	Chinese	516 (0.3)	12 493 (0.5)		865 (0.4)	17 702 (0.7)	
	Other	2616 (1.3)	41 738 (1.7)		3980 (1.7)	46 429 (1.9)	
Townsend deprivation index, mean (SD)		-1.3 (3.1)	0.3 (3.6)		-1.4 (3.0)	0.2 (3.6)	
BMI (kg/m2), mean (SD)		27.8 (4.2)	26.3 (4.2)		27.0 (5.2)	25.7 (5.0)	
Smoking status	Non-smoker	97088 (49.5)	1 081 822 (44.2)	<0.001	142569 (59.8)	1 433 446 (57.4)	<0.001
	Ex-smoker	75100 (38.3)	448 480 (18.3)		74934 (31.4)	392 870 (15.7)	
	Light smoker	9361 (4.8)	351 559 (14.4)		8885 (3.7)	284 482 (11.4)	
	Moderate smoker	5816 (3.0)	167 089 (6.8)		7235 (3.0)	152 115 (6.1)	
	Heavy smoker	8726 (4.4)	139 985 (5.7)		4873 (2.0)	86 114 (3.5)	
Alcohol intake	Non-drinker	11985 (6.1)	433 515 (17.7)	<0.001	22415 (9.4)	753 150 (30.2)	<0.001
	Trivial drinker	41810 (21.3)	585 589 (23.9)		96085 (40.3)	849 734 (34.0)	
	Light drinker	57817 (29.5)	358 713 (14.7)		76942 (32.3)	295 009 (11.8)	
	Moderate drinker	60694 (31.0)	486 003 (19.9)		37830 (15.9)	176 644 (7.1)	
	Heavy drinker	14960 (7.6)	41 223 (1.7)		3797 (1.6)	5332 (0.2)	
	Very heavy drinker	8825 (4.5)	18 473 (0.8)		1427 (0.6)	3743 (0.1)	
Medical history	Ulcerative colitis	1053 (0.5)	8956 (0.4)	<0.001	1211 (0.5)	8983 (0.4)	<0.001
	Colorectal polyps	616 (0.3)	3146 (0.1)	<0.001	612 (0.3)	2447 (0.1)	<0.001
	Diabetes	12893 (6.6)	68 727 (2.8)	<0.001	7885 (3.3)	53 070 (2.1)	<0.001
	Breast cancer				9448 (4.0)	25 108 (1.0)	<0.001
	Uterine cancer				1030 (0.4)	1987 (0.1)	<0.001
	Ovarian cancer				724 (0.3)	2242 (0.1)	<0.001
	Cervical cancer				1711 (0.7)	3582 (0.1)	<0.001
	Lung cancer	125 (0.1)	1488 (0.1)	0.643			
	Blood cancers	1146 (0.6)	5953 (0.2)	<0.001			
	Oral cancer	483 (0.2)	964 (0.0)	<0.001			
Family history of CRC		19505 (9.9)	29 877 (1.2)	<0.001	22252 (9.3)	43 741 (1.8)	<0.001

Values are numbers (%) unless indicated; proportions given for QCancer-10 derivation cohort are for those with recorded data

4.7 Discussion

UK Biobank is a hugely powerful resource, and one of few with a sufficiently large phenotyped and genotyped cohort to facilitate the evaluation of integrated risk prediction models for CRC. However, my examination of the dataset raises several key issues in CRC prediction modelling.

There are notable differences between the UKB dataset and the general population. For example, I confirmed that registry-based CRC rates were lower in UKB than in ONS data, particularly at older ages, though in my derived Integrated Modelling Cohort the directly standardised rate for women was close to that of the

general population. Lower CRC rates in UKB reflect the known 'healthy volunteer' effect within UKB [266]. This is a form of selection bias commonly seen in cohort studies relying on voluntary participation [437]. Research volunteers tend to be healthier, with higher health-consciousness, higher educational attainment and higher socioeconomic status than non-volunteers [438, 439].

The extent to which the UKB cohort differs from the UK population has been described previously by Fry et al. [266]. Of the 9.2 million individuals invited to participate in UKB, 5.5% attended the baseline assessment. Participation was higher in older ages, females, white people, and those from less socioeconomically deprived areas. There was also geographic variation, with lowest uptake in West Scotland. Participants are taller and leaner than the population average, with lower smoking prevalence, lower prevalence of common medical conditions (based on self-report), and lower all-cause mortality.

There are implications of non-representative samples for prediction model development. For example, as a consequence of the disparity in cancer risk, baseline hazards calculated in in this work are likely to be lower than would be expected in the general population, particularly for men, and so recalibration is likely to be needed on application to a more general population. In addition, model performance varies between populations with different prevalence or risk of a disease (the 'spectrum effect'). With a higher risk of CRC in the general population of screening age, compared to UKB, sensitivity might be anticipated to increase in the general population improving performance as a screening test [440].

This healthy volunteer bias [437] is also seen in my comparison of the integrated modelling and QCancer-10 derivation cohorts. As discussed in Chapter 1, QCancer-10 is a more population-based cohort. The Integrated Modelling Cohort is older with a much narrower age range, is less ethnically diverse and more affluent. There are fewer current smokers, and higher levels of alcohol consumption. Far more participants in UKB report a positive family history of CRC, likely reflecting the differences in ascertainment methods. Medical conditions included in the QCancer-10 model are also more prevalent (except lung cancer), probably due to the different

age distributions of the cohorts. As a result, particularly as age is such a strong predictor of CRC risk, performance of QCancer-10 in UKB has been shown to be lower than on validation in a population cohort [172, 182], with implications for the relative performance of my integrated prediction models.

Several other forms of selection bias should also be considered. Informative censoring is induced by differential loss to follow-up, that is, systematic differences in exposures or outcomes between those lost and retained. The use of linked datasets in UKB follow-up dramatically minimises this (~0.002% of UKB participants with available genetic data were lost to follow-up within 5 years of enrolment), and thus is unlikely to have a significant impact on my analysis. In my case/control PRS analysis, the use of the all remaining unaffected participants as controls avoids matching bias, where over-matching can lead to underestimation of associations.

Prevalence-incidence bias (or Neyman's bias), results from the inclusion of prevalent cases (as in my case-control cohort, and many other PRS studies [228, 441]) and distorts the frequency of the exposure in a dataset due to sampling of cases [437], typically leading to underestimation of effect sizes. However this has been shown to occur only if the exposure under consideration influences mortality [442]. Whilst a PRS for CRC incidence may influence CRC survival (for example polymorphisms may increase the chance of both cell proliferation and cancer development and of metastatic spread) this has not been demonstrated in the literature. Of co-predictors in the PRS models, age may also be subject to Neyman's bias, as those diagnosed at an older age may have a higher associated mortality due to co-morbidities. There is the potential therefore for the effect sizes to be underestimated.

Lower rates of CRC in UKB, and subsequently the number of cases included in UKB follow-up, also impacts sample size requirements. For PRS development, it is the sample size of the base GWAS dataset that is key to determining PRS performance, allowing separate validation cohorts to be created. This enables testing of performance in a cohort of more diverse European ancestry than the white-British derivation cohort, assessment of the portability of the models, and a

more detailed analysis of performance in other ethnicities. I explore the participant exclusions in my PRS modelling datasets in more detail in Chapter 5.

Sample size calculations for the Integrated Modelling Cohort showed that a large number of events per predictor (42 EPP) and events (1569) are needed for the model for women, which is unachievable in my modelling dataset. I continued with model development, however the female integrated models may be at higher risk of over-fitting, and risk estimates may be less precise [443]. External validation, to confirm estimates of performance and risk estimates, would be essential prior to planning clinical use of this model. True external validation is performed in an entirely new dataset (and indeed by a separate research team for 'fully independent validation' [284, pp. 308-309]) and was outside of the scope of this study. At the time of this study, a mean of 7 years of follow-up was available due to right-censoring of registry data. Extended registry follow-up has recently been made available (up to July 2019 for most participants). An extension of this thesis work would be to re-estimate my models with the extended follow-up (and increased case numbers) available, which would improve accuracy of performance estimates, and permit estimation of absolute risks over a longer time horizon.

The thorough phenotyping of the UKB dataset enabled closely matched coding of all Qcancer-10 predictors. While the highly standardised and comprehensive data collection used in UKB reduces the risk of information bias, some risk of bias could persist. Misclassification biases occurs when exposures are misclassified due to random measurement errors or non-random biases, and several forms of misclassification bias are particularly relevant here. Reporting bias occurs where participants give answers they feel are more likely to be of interest to researchers, or under-report socially unacceptable exposures such as heavy smoking ('underreporting bias'). Quantitative/frequency exposures (for example for alcohol here) tend to be underestimated, as participants tend to report modal rather than mean exposures ('mode for mean bias'). Recall bias, particularly in case control studies, arises where knowledge of disease status influences recollection and perception of potential causes. This is less problematic in prospective studies,

particularly where prospective outcomes are obtained through linked registries. Detection bias occurs when the measurement or recording of outcome differs systematically between different exposure groups. The risk of this is low in UKB due to the use of linked data for follow-up.

As data collection in UKB is blind to the outcome of interest any misclassification bias should be 'non-differential', i.e. the risk of bias is the same in cases and controls. In binary variables misclassification bias tends towards the null, however in predictors with multiple categories the bias can tend in either direction [437].

A further type of information bias pertinent to cohort studies is regression-dilution bias. The random error in exposure measurements which causes this bias is due to natural biological variability (for example measurement of weight at different times of the day) or measurement imprecision. Random error in exposures tend to bias regression estimates downwards [426]. Few exposures considered in my QCancer-10 or PRS modelling are likely to be subject to significant error (aside from BMI), nor is my outcome measure.

Whilst I have not re-estimated the effect sizes of QCancer-10 predictors individually in my analyses (instead validating the performance of the original model with limited recalibration, see Chapter 6) biases in exposure measures could lead to underestimation of the effect size QCancer-10 risk scores in the combined model (as opposed to the PRS, which is at low risk of most forms of information bias).

High levels of missingness, and handling of this, can also introduce misinformation and selection biases, and lead to inefficient analysis. Missingness was <1% for almost all predictors in my study, aside from genetic data and family history of CRC, which were still <5% missing. Different mechanisms of missingness require different approaches in statistical analysis. Several options are available in handling missing data, including complete case analysis and imputation methods. A complete case analysis includes only individuals with data available for every predictor for each analysis. The drawback of this is loss of data and resulting inefficiency, with potential loss of statistical power. Available case analysis includes cases with available data for the analysis in question. Each analysis is then based on different

cohort sizes, which can make comparisons more difficult - for example changes in hazard ratios may be due to differences in the model, but also potentially due to the differing subject mix [284, p.117].

Alternatives to these complete case approaches include single imputation, in which a single value (for example the mean) replaces missing values. This is straightforward but does not take into account correlation between, and variability of, predictor values. Multiple imputation, in which missing values are replaced with plausible values based on relationships between the predictors amongst other participants, is more comprehensive. In multiple imputation, the imputation process is repeated n times to give n full datasets, with the analysis (i.e. model fitting and estimation of model performance) completed on each imputed dataset. Results are combined to obtain point estimates and variance for model coefficients and performance across the dataset, incorporating uncertainty in the underlying imputation process [284, pp.122-23]. Imputation methods are dependent on missingness being at least MAR (i.e. are not MNAR).

Missingness of most predictors considered here appeared to be MAR rather than MCAR. Notably, missing data of two key CRC predictors (genetic data and family history of CRC), tended to be higher in participants from minority ethnicities, and those with higher Townsend scores (which may be related to one another). It is not possible to directly test that data are MNAR, as this requires collecting and examining some of the missing information. For some predictors, the mechanism of missingness may be MNAR. Overall, as the proportion of missingness for these variables is low, the bias introduced by these values is likely to be small.

It is reasonable, where missingness is <5%, to perform complete-case analysis [285], particularly in very large datasets, which was the approach I took. In the development of QCancer-10, BMI, alcohol, and smoking status were imputed, but these predictors were thoroughly recorded in UKB. Ideally, I would have undertaken a sensitivity analysis with multiply imputed data, which would also potentially have increased the available sample size for the female integrated model. However, Steyerberg notes that one would not impute a predictor of primary interest [284,

pp.131], and so I would not have imputed PRS into the dataset, and missingness for most other predictors was lower in women than men. In addition, in QCancer-10 development for many predictors, absence of information was assumed to mean absence of disease. In UKB all of these predictors are explicitly queried and so I did not use this as my primary approach, however conducting a sensitivity analysis with missing values set as no, would be useful further work.

A potential issue with complete case analysis is bias introduced when there are systematic differences between participants with missing and complete data [444]. I observed systematic differences between individuals with and without missing family history and genetic data, particularly with regards to ethnicity. The exclusion of a disproportionate number of minority ethnic participants is significant, exacerbating the under-representation that already exists in UKB. This bias in prediction modelling is particularly an issue when missingness is associated with the outcome [284, p.117], which is highly unlikely with a prospective study such as this. However, as a result of these exclusions, the power to evaluate model performance in minority ethnic individuals is further reduced, which may mean that accurate inferences about performance cannot be made.

In conclusion, the size of dataset, extensive and complete phenotyping and genetic data, and reliable follow-up through data linkage afforded by UKB make it a valuable dataset for risk model development and evaluation, as evidenced by the number of predicting modelling studies already conducted using the resource (discussed in Chapter 1). The risk of bias overall is low, but should be considered in interpreting the results presented in the following chapters. UKB is arguably an approporiate dataset for risk modelling in the context of bowel cancer screening. The age range of UKB is similar to that of the bowel cancer screening programme. In addition, although UKB is not completely representative of the UK population, the same applies to bowel screening participants, in whom we see higher screening uptake in less deprived areas (measured by Index of Multiple Deprivation), less ethnically diverse areas, and among women [56, 445] - mimicking the healthy response bias seen in UKB. Thus model performance estimated in this cohort may

be reasonably representative of that of the current (though not an ideal) bowel cancer screening cohort. The next chapter presents the risk modelling work in UKB, with a specific focus on the development of polygenic risk scores.

5

Polygenic risk scores for colorectal cancer

This chapter describes the derivation and evaluation of novel polygenic risk scores in UK Biobank, based on the most recent GWAS meta-analyses.

5.1 Background

A number of PRS for colorectal cancer (CRC) have been published previously, most containing a small number of GWAS-significant SNPs [215]. As discussed in Chapter 1, the predictive power of PRS are limited by the size of the GWAS datasets available, and the genetic architecture of the phenotype under consideration. As GWAS sample sizes increase, the accuracy of effect size estimates, and the power to detect rarer variants also increases [232].

Early studies of PRS used modelling to estimate the probable discrimination achieved by known SNPs. Pharoah et al. [446] used SNP effect sizes and allele frequency to estimate the variance of a PRS, assuming a log-normal distribution of genetic risk, and derived relative risk distributions for breast cancer in the general population. Subsequently, early polygenic risk scores comprised a simple summation of risk alleles, typically weighted by their effect sizes from GWAS. With the addition of more risk SNPs, predictive performance has improved. In a recent external validation study of GWAS-significant PRS for CRC in UKB, models

including a greater number of SNPs performed better [215], but discrimination and explained variation remained poor.

A strong argument has recently been made for including a greater number of genome-wide SNPs in PRS, including those below the GWAS-significance threshold of $p<5x10^{-8}$. "Ignorance cannot be bliss", state Wray et al. [232] - it must be better to use all available information in this context. Many studies have since demonstrated, across a variety of phenotypes, that the inclusion of thousands or millions of SNPs, selected using a range of methodologies, achieve greater predictive power [441, 447, 448].

Genomic prediction studies first need to identify the optimal set of predictive SNPs to include, and then the weight for each SNP representing its contribution to the predictive model [208]. The initial (and simplest) approach taken to expanding PRS was to evaluate including SNPs across a range of significance thresholds, retaining the weights from GWAS. However, this does not take into account correlation between SNPs (linkage disequilibrium, LD), which exaggerates the effect of a given locus if many SNPs in high LD are included. An extension of this approach used 'pruning and thresholding', removing SNPs in LD with one another at successive loci across the genome [447]. However, the most predictive SNPs can be pruned away using this approach, and so pruning was replaced with clumping (C+T). Here, the most strongly associated variant is selected, and correlated variants located within a given genetic distance are excluded, in an iterative process [207]. Clumped variants reaching a given association significance threshold, are then retained, weighted by their GWAS effect size. A range of clumping R^2 and thresholding significance values are compared, and the optimal parameters selected based on the best correlation, or most significant association with, the phenotype on testing in a regression model. The optimal threshold for inclusion is dependent on both GWAS sample size and genetic architecture of the trait in question [447].

A limitation of this approach is the failure to allow joint estimation of SNP effect sizes (that is, these analyses assume SNPs act independently). A range of statistical methods for effect size estimation alongside SNP selection have been evaluated

to account for this. Abraham et al. [208] demonstrated that sparse penalised machine learning methods such as lasso and elastic-net had significantly better predictive performance than traditional approaches, leveraging LD information to improve their accuracy. Several early such methodologies required access to primary GWAS data rather than summary statistics, limiting their implementation by most researchers with access to only the summary data [209]. A number of publicly available tools have since been developed utilising these approaches for wider use. Mak et al. [209] developed Lassosum, using penalized regression for effect estimation. Wray et al. [232] highlighted Bayesian approaches, incorporating information about the prior assumptions of distribution of SNP effects, as an optimal approach, and Vilhjálmsson et al. [210] developed LDpred, using Bayesian methods to estimate SNP effect sizes based on summary statistics and an LD matrix, facilitating wider use of this approach. This has subsequently been updated to LDpred2 [211]. Ge et al.'s PRS-CS method also uses a Bayesian approach with a continuous prior [212].

Two studies have examined genome-wide PRS for CRC, both using the Huyghe et al. [222] meta-analysis as a 'base' dataset. Thomas et al. [228] compared a GWAS-significant PRS with models derived using LDpred and machine learning approaches. Their LDpred model, including 1.2 million SNPs, had the highest discrimination. The area under the ROC curve (AUROC), adjusted for age and sex, was 0.654 in external validation. Fritsche et al. [219] evaluated fixed thresholds, C+T, and lassosum, examining performance in two different cohorts - UK Biobank, and the Michigan Genomics Initiative (MGI). They compared a variety of different base GWAS analyses, mostly derived in UKB. The top-performing CRC PRS in the MGI cohort was obtained using lassosum but contained just 150 SNPs, and achieved an AUROC of just 0.567 (95% CI 0.540-0.594).

A key issue in many PRS studies to date has been overlap between the base and PRS modelling datasets. This results in over-estimation of the association between the PRS and outcome in the training dataset [282], resulting in over-fitting to the training dataset, and inflated performance estimates [232].

In this chapter I aimed to develop and evaluate PRS for CRC using a number of genome-wide approaches, and compare these with a GWAS-significant PRS. I hypothesised that the use of a large base GWAS dataset, and recently developed PRS methodologies, would produce PRS which outperformed those previously published.

5.1.1 Chapter Outline

As an initial evaluation of PRS, I use the risk distribution methodology of Pharoah et al. [446] to model and evaluate a PRS and the impact of risk-stratified screening on screening numbers.

I then develop 6 different PRS for CRC in the UKB dataset using 3 main methodologies. Firstly, I derive a weighted GWAS-significant PRS model. Secondly, I use two different C+T approaches - a standard approach, and an extension to this, stacked clumping and thresholding (SCT), which has not previously been evaluated in CRC, which learns an optimal combination of C+T across chromosomes. Thirdly, I evaluate three different approaches using the LDpred2 programme. I use the GWAS meta-analysis presented in Chapter 3, recalculated without the UKB dataset (to remove overlap with UKB), as my base dataset, to produce less biased performance estimates than those previously published.

I compare the apparent and internally validated performance of my PRS in a white-British cohort of UKB participants from England and Wales, and externally validate performance in both a white European Geographic Validation Cohort and a Minority Ethnic Validation Cohort (see Section 4.6). I present subgroup analysis of PRS performance by age, and in individuals with a family history of CRC. The genome-wide PRS with the best predictive performance, and the GWAS-significant PRS, are then used in combined modelling with QCancer-10 in Chapter 6.

5.2 Methods

Results of my polygenic risk modelling are presented according to PRS-RS guidelines [171], and reporting of all modelling follows the TRIPOD guidelines [168].

5.2.1 Modelling colorectal cancer risk from a polygenic log-normal distribution

The potential utility of a PRS based on GWAS-significant SNPs can be evaluated by comparing cancer cases detected using age-based screening with those detected using a model based on calculated absolute risks in the population, as shown by Pashayan et al. [449]. Here the polygenic relative risk distribution is estimated in cases and controls (based on earlier work by Pharoah et al. [446]), and overlaid on the calculated cancer-specific incidence of CRC, to derive polygenic risk based absolute risks.

In a polygenic model of complex disease, genetic risk distribution in the general population is assumed to be log-normal, with mean μ and standard deviation σ, [446]. The μ is an arbitrary constant, which may be set to $\mu = -\sigma^2/2$, so that mean population relative risk at birth is unity. The variance, σ^2 is calculated from the per-allele relative risk (taken from the GWAS-derived odds ratios) and the risk allele frequency in the population, assuming a log-additive model of risk allele interactions:

$$\sigma^2 = 2\Sigma_\kappa p_\kappa (1 - p_\kappa)\beta_\kappa^2$$

where β and p correspond to the log odds ratio and risk allele frequency for SNP κ [446, 449]. The genetic risk distribution in cases is also log-normal with the same variance, but with mean relative risk shifted to the right on a log scale by σ^2 [446]. From this calculated variance, percentile relative risks, and proportion of cases occurring at these relative risks, can be calculated [449].

I manually curated a list of SNPs from meta-analyses in Law et al. [115] and Huyghe et al. [222], and references within. Where SNPs had been identified in both datasets the effect size (β) and MAF were taken from Law et al. [115]. I excluded several conditional SNPs from Huyghe et al. as the conditioning SNPs were in LD (r^2 >0.1) with existing SNPs in our meta-analysis, and it was unclear how best to handle these. Fifteen SNPs identified in Huyghe et al. [222] but not in Law et al. [115] were excluded. The rare variant on 5q21.1 from Huyghe et al. [222] was also excluded. The final PRS included 97 SNPs.

In GWAS, SNP effect size estimates are subject to the 'winner's curse', a form of ascertainment bias in which the effect estimates in discovery studies tend to be biased upwards [450]. Signals with the greatest effect size are most affected by this bias. I corrected SNP betas for the winner's curse using the False Discovery Rate Inverse Quantile Transformation (FIQT) method [451]. Whilst other methodologies have been proposed for correction, the code was not available for these [452, 453].

I derived absolute risks for CRC from population data. I obtained data for diagnosis from CRC registrations, deaths from CRC and all causes, and mid-year population estimates for England from the ONS for the years 2012-2017 [454, 455]. I calculated mean cancer incidence rates, and cancer-related and all-cause mortality rates across this time frame. Age-conditional 10-year absolute risks of CRC diagnosis were then derived in 1-year age bands for men and women up to the age of 75 using DevCan software vsn.6.7.6 [456].

I used percentile relative risks from the 97-SNP PRS to calculate the 10-year absolute risk of CRC at 40-74 years of age across risk centiles. I then evaluated the number of individuals eligible for screening and cases detected in a hypothetical screening cohort of 100,000 individuals, in which the number of cases at each year of age was based on CRC rates in the ONS data. I compared age-based screening of 50-74 year olds with risk-based screening assuming participation in at a 10-year absolute risk level equivalent to that of an average 50 year old.

5.2.2 Base genome-wide association study meta-analysis

In order to avoid overlap between base and target datasets in PRS development, for the PRS base dataset I re-performed the meta-analysis described in Chapter 3, excluding the UKB dataset. This included 14 GWAS cohorts, with 26,397 cases and 41,481 controls of European ancestry based on PCA. I used the same software and inclusion criteria as previously described in Section 3.3.1.

5.2.3 GWAS-significant polygenic risk score

From the list of SNPs curated in Section 5.2.1 I excluded SNPs which did not reach genome-wide significance ($P_{assoc}<5\times10^{-8}$) in my base meta-analysis. I used the effect sizes from my base dataset, again adjusting for the winner's curse using FIQT correction [451]. Where SNPs were reported at the same loci in different studies and were correlated at $r^2>0.1$ I retained the most significantly associated SNP. One SNP, rs9537521, was not present in UKB data, with no LD proxy available, and so was excluded.

I used QCtool v2.0.1 [295] to extract SNP dosages from the UKB bgen format imputed dataset, flipping dosages where required to match the effect allele. I weighted dosages by the effect sizes per SNP, and summed these to derive the scores. The minimum INFO score for SNPs included in the GWAS-significant model was 0.932.

5.2.4 Clumping and Thresholding

I initially planned to use PRSice2 PRS software for C+T approaches [457]. In this analysis, I removed ambiguous SNPs, those failing imputation in >50% of dataset, structural variants, and SNPs absent in the UKB dataset. I included SNPs passing QC criteria of MAF (calculated on the target dataset only using QCTOOL) of ≥ 0.01 and INFO score of ≥ 0.6. I then attempted to run PRSice2 using default parameters of an additive model with mean imputation of missing values, clumping r^2 of 0.2. Despite using the maximum available computational power available, discussion with the author, and support from central computing services, I could not successfully run PRSice2 on my dataset. I therefore subsequently used the R package bigsnpr for C+T and SCT approaches [300].

Prior to C+T, SCT and LDpred2 modelling, I undertook further QC steps beyond the standard QC described in Chapter 4. I followed the methodology of Privé, Arbel, and Vilhjálmsson [211] and restricted the dataset to individuals classed by UKB as of white-British ancestry (identified through a combination of self-report and principal components analysis). I included individuals included in

UKB PCA analysis (removing related individuals at 3 degrees of relatedness or more) [267], and used computation of robust Mahalanobis distances to create a more genetically homogeneous dataset, removing individuals with log-distance >5 [211]. This resulted in a dataset of 310664 individuals with 4676 cancer cases.

Following per-SNP QC steps suggested by Privé, Arbel, and Vilhjálmsson [211], I restricted variants considered for genome-wide PRS calculated to those included in HapMap3 (n = 1,117,493), which matched the 12,972,739 SNPs in the base GWAS summary statistics. I removed ambiguous SNPs (i.e. A/T or G/C SNPs), and further matched variants between the base dataset using chromosome, genetic position and reference and alternate allele, leaving 1,117,002 SNPs. I then compared standard deviations of genotypes in the summary statistics (SD_{ss}) and LDpred2 LD Training Cohort (SD_{ldtr}), and removed variants where $SD_{ss} < 0.5$ (SD_{ldtr}), $SD_{ss} > 0.1 + SD_{ldtr}$, $SD_{ss} < 0.1$ or $SD_{ldtr} < 0.05$ (see Figure 5.1). This left 1,104,409 SNPs for inclusion in the PRS. The minimum INFO score of these SNPs was 0.411, and I performed no further filtering on this.

Both C+T and LDpred2 approaches require an LD matrix. I constructed the LD matrix from 10,000 randomly selected individuals from the Derivation Dataset (the LD training dataset), which calculates Pearson correlations between SNPs within 500kb windows, calculating p-values using two-sided t-tests [211]. A minimum of 1000 individuals is recommended for the LD reference panel [210]. To the 10,000 used in LD matrix calculation I added 20,000 randomly selected individuals to generate the Training Cohort, used for parameter selection, which included 30,000 participants with 446 cases.

I generated C+T scores across a grid of r^2 (0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 0.95), 50 thresholds of -log10(P-values) (spaced equally between the most significant p-value and 0.1 on a log scale), and base clumping window size values (50, 100, 200 and 500) [211]. The actual window size in kb is the base size divided by clumping r^2, so that for an r^2 of 0.1, window sizes would be 500, 1000, 2000 and 5000kb; this adjustment is made as LD between variants is inversely proportional to genetic distance [458, 459]. Clumping used data on 10,000 individuals from the LD training dataset for

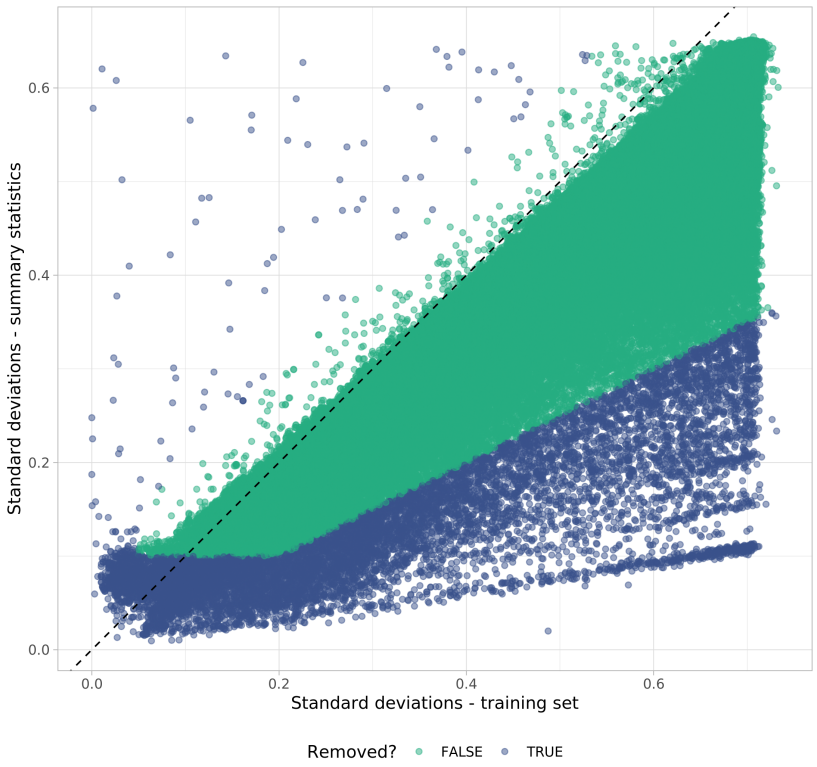

Figure 5.1: Quality control based on genotype matching between summary statistics and Training Cohort

computational efficiency; thresholding was performed using the expanded 30,000 person Training Cohort. From this grid of 1,400 C+T PRS, the top-performing C+T score was selected based on AUROC in linear regression.

I also evaluated stacked clumping and thresholding (SCT), developed by Privé, Aschard, and Blum [460], which develops C+T by learning the optimal linear combination of C+T scores generated through efficient penalised logistic regression, and outperforms C+T in other phenotypes [458]. This facilitates joint estimation of effect sizes which can improve predictive performance [460]. Stacked clumping and thresholding uses the vectors of C+T polygenic scores per chromosome (therefore here 22 x 1400 = 30,800 scores) as explanatory variables in a penalised regression, in which weights are fitted for each score. A single vector of effect sizes is then derived from the linear combination of these scores [458].

5.2.5 LDpred2

LDpred2 [211], a recent update of LDpred [210], uses a Bayesian approach to SNP selection and shrinkage for PRS, based on an LD matrix and GWAS summary statistics, implemented in the R package bigsnpr [300]. The prior for the effect sizes includes two parameters - the heritability explained by the included genotypes, h^2, and the proportion of causal SNPs, $prop_{causal}$.

The updated version of LDpred [210] provides higher predictive performance, particularly with large GWAS sample size [211], corrects previous instability issues [461], and evaluates more hyper-parameters (126 different iterations instead of 7 in LDpred). A larger window of 3cM (using genetic distance rather than the number of bases) improves performance particularly for causal variants in regions of long-range LD (for example HLA regions). Colorectal cancer-associated variants in HLA regions have recently been reported [115, 222] and this improvement may therefore be of benefit in CRC-prediction. The authors provide a tutorial and code accompanying their paper [462], on which my analysis was based.

There are multiple options for PRS construction within LDpred2:

- An infinitesimal model (LDpred2-inf) assumes all markers are causal;
- Grid models (LDpred2-grid) require hyper-parameters SNP heritability, h^2, (calculated from constrained LD score regression), proportion of causal variants, $prop_{causal}$, and sparsity (allowing variant effects to be zero), are tuned in a training set (called the validation set in the original paper);
- An auto model (LDPred2-auto) which automatically estimates sparsity and SNP heritability, removing the need for a training set.

I evaluated LDpred2-inf and LDpred2-grid models (sparse and non-sparse), running these genome-wide (rather than by chromosome) as recommended. My grid of tuning parameters for LDpred2-grid models included 3 heritability estimates (0.1121, 0.1602, 0.2243), 21 values for $prop_{causal}$ evenly spaced between 0.00005 and 1 on the log scale, and sparse/non-sparse. I then selected the optimally

performing sparse and non-sparse models based on the best Z-score for the logistic regression slope, adjusted for array platform and first 4 principal components (PCs). The use of Z-score is recommended in this context by the authors as being more robust than AUROC [211].

5.2.6 Apparent polygenic risk score performance and internal validation

Logistic regressions models, using prevalent and incident cases, have commonly been used in other PRS studies to assess PRS performance [228, 441]. In addition, I was interested in performance for prospective prediction of absolute risk, as this is preferred in decision making, and is also the methodology used for QCancer-10 models [182]. For each PRS I therefore evaluated the association with CRC risk in both logistic regression models and Cox proportional hazards (Cox) models (see Section 2.8).

Prediction models included the PRS, age (taken at enrolment in to UKB, equal to the study entry time point), sex, genotyping array and 4 principal components (PCs), with apparent performance evaluated in the Test Cohort (280,664 individuals, with 4,230 cases; case prevalence was 1.5% in both Training and Test Cohorts). Age and sex were included as key co-variates known to be independent predictors of CRC risk. The inclusion of PCs accounts for population stratification due to underlying population genetic structure (see Section 2.8.3), and genotyping array accounts for possible differential genotyping between the two platforms. Age, PRS, and PCs were modelled as continuous variables, assuming a linear relationship (on the log scale).

For Cox models I confirmed proportional hazards assumptions held through visual inspection of plots of $log(-log)$survival against log(survival) [284, pp. 78, 285], which form parallel lines if the model has a constant hazard ratio, and through statistical testing of proportional hazards, per variable and overall, with inspection and plotting of Schoenfeld residuals.

I evaluated interactions between age and PRS by fitting an interaction term along with other predictors in both logistic regression and Cox models, and examining the

strength and significance of Wald χ^2 statistics (using a stringent p-value cut-off for interaction terms of < 0.01 to account for multiple testing) for these interactions [284, pp.217]. I also examined plots of marginal effects of PRS in the interaction.

PRS model performance was compared to a reference model which included age, sex, array, and 4 PCs, in order to evaluate the contribution of PRS to performance. In addition, I evaluated models which did not include age and sex, to evaluate the contribution which these factors (known to be independent predictors for CRC risk) made to the performance of the full model [282].

I plotted distributions of the PRS in each cohort, standardised to a mean of 0 and standard deviation of 1 in the Derivation Test Cohort. Assuming PRS are based on a summation of independent SNPs with the same distributions, the central limit theorem indicates the PRS in a sample should follow an approximately normal distribution. Deviation from this can indicate that either included SNPs are correlated, or the inclusion of SNPs with different distributions (for example in divergent ancestries) [282]. I calculated odd ratios and hazard ratios per-standard deviation of PRS for each model. (These standardised scores were used for the plots of marginal effects if PRS in interaction with age; all other analyses used non-standardised scores).

Overall model performance was assessed using the scaled Brier Score (see Section 2.8.8). I assessed discrimination using the C-statistic and Somers' D_{xy} for logistic regression models, and Harrell's c-index and Royston and Sauerbrei's D statistic for Cox models. For Cox models I additionally plotted Kaplan Meier cumulative incidence plots in four risk groups (Section 2.8.8). To assess explained variation, for logistic regression models I used Nagelkerke's R^2, and calculated Nagelkerke's R^2 attributable to the PRS (R^2 of the full model minus R^2 of the null model). For Cox models, I used Royston and Sauerbrei's R^2_D.

I used 500 bootstrap samples to generate confidence intervals and for internal validation [463]. This randomly samples from the original data with replacement to generate new datasets, estimating performance in each, permitting estimation of the distribution of performance measures, and optimism within these estimates.

I evaluated the significance of the differences in performance of the models using paired t-tests. I used the optimism-adjusted calibration slope as a global shrinkage factor to adjust the linear predictor for the models, and re-estimated the intercept for logistic regression models, and $S_0(t)$ for survival models [285], by refitting the models with the adjusted LP as an offset (i.e. as the model covariate with the coefficient constrained to 1).

5.2.7 External validation and subgroup analysis

Performance of the adjusted models was validated in the Geographic Validation Cohort and Minority Ethnic Validation Cohort. In addition to measures described above, I evaluated the calibration slope and calibration in the large (CITL), and visually assessed calibration plots (see Section **ref**(meth-modelcalibration)). Notably, whilst the Hosmer-Lemeshow goodness of fit test is often used to report calibration, this is over-sensitive in large sample sizes [169], and was not used here. I used recalibration-in-the-large (Section 2.8.11) to recalibrate the models in the external datasets.

I examined model performance in males and females separately in the Geographic Validation Cohort to facilitate comparison with the sex-specific models developed in Chapter 6, and in individuals with a family history of CRC. I also plotted calibration by age by plotting expected and observed risk in 5-year age bands.

5.3 Modelling polygenic risk scores from the log-normal polygenic distribution

Following the publication of the most recent CRC GWAS meta-analyses [115, 222], I evaluated the potential for a PRS derived from discovered GWAS_{meta} significant SNPs to discriminate CRC risk, and of a PRS containing all common variation, following the methodology of Pharoah and colleagues [446].

The 97 SNPs of this PRS (listed in Appendix A) result in a log-normal distribution of relative risk with σ^2 of 0.34 (Figure 5.2 A). For a PRS including all known possible variants, σ^2 is 1.164 (Figure 5.2 D). The 97-SNP PRS confers

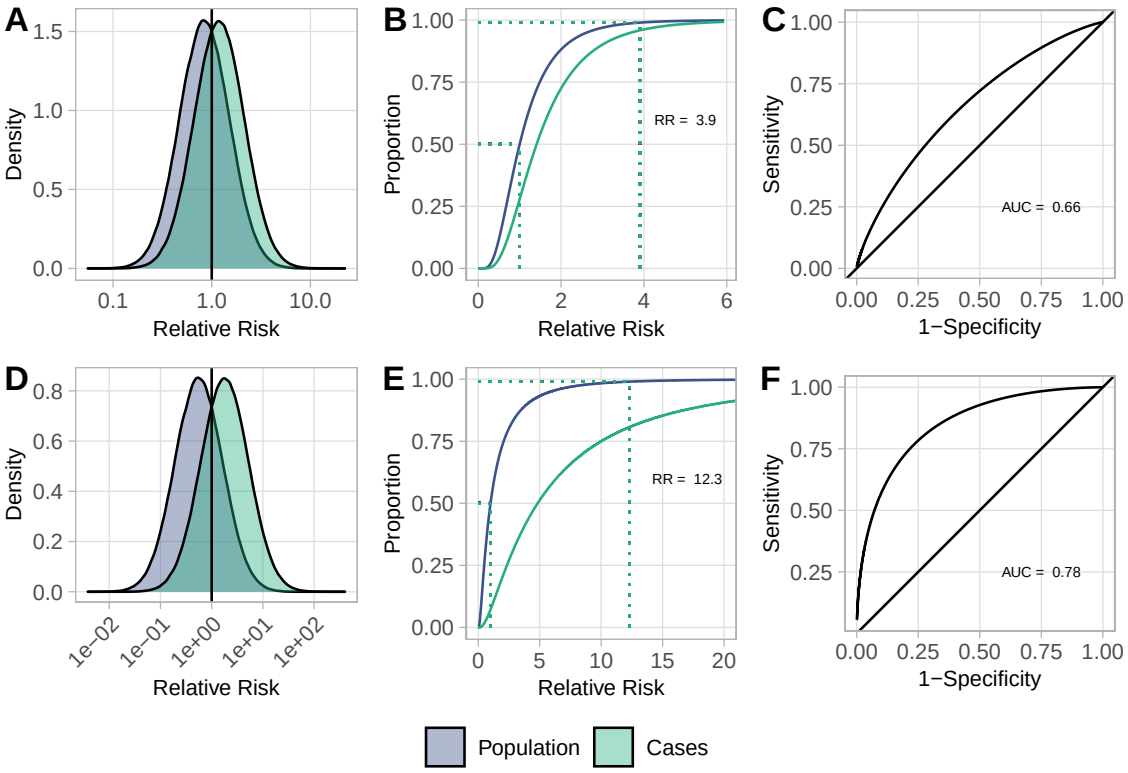

Figure 5.2: Performance of PRS models derived from modelling of log-normal distributions. Plots show density distributions of relative risk (A), cumulative population risk (B) and AUC (C) for a PRS containing 97 GWAS-significant SNPs, and for a PRS assuming all common variants are known (D-F).

a 3.9 fold difference in RR between those in the 99th and 50th centiles; and a 2.1 fold difference in risk between top decile and median risk for 97 SNPs. Including all common variation results in a 12.3-fold increase in risk at the 99th centile. The AUROCs for the 97-SNP PRS, and all-known PRS are shown in (Figure 5.2 C,F).

The age conditional 10-year absolute risk of CRC from birth to 75 is shown in Figure 5.3. Ten-year absolute risk of CRC on entry to the screening programme at 50 is 0.48% for women and 0.65% for men.

Plots of 10-year absolute risks across PRS centile thresholds (Figure 5.4) illustrate the wide variation in risk for different groups of polygenic risk, and the marked difference between males and females: women in the 95th centile have broadly the same risk as men in the 80th centile of risk.

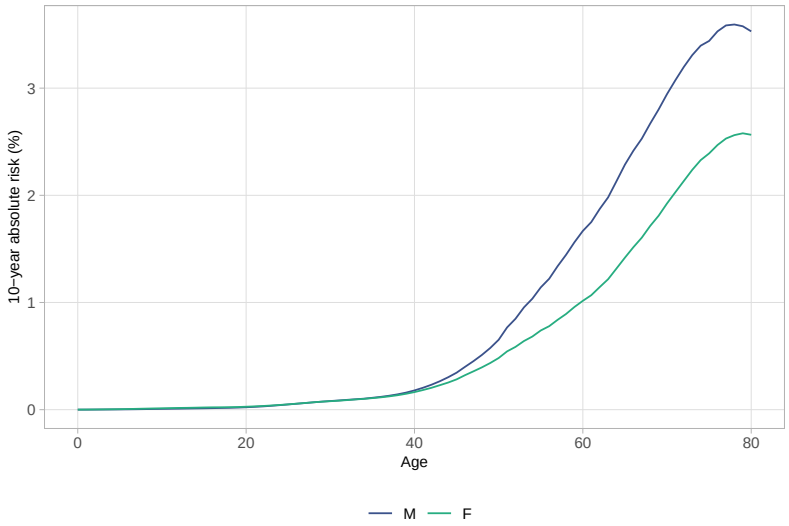

Figure 5.3: 10-year absolute risk of CRC in males and females aged 0-90 years, derived from 2012-2017 ONS data

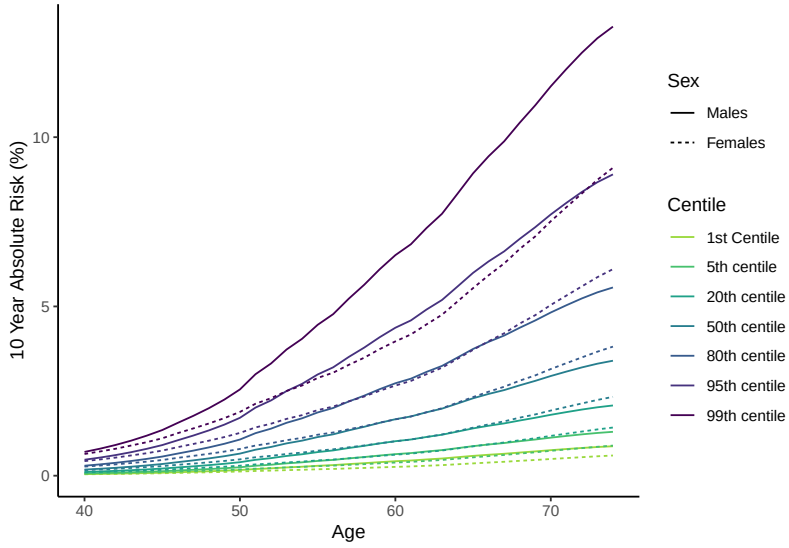

Figure 5.4: 10-year absolute risk of CRC in males and females by PRS centile

5.3.1 Modelling the impact of risk-stratification on a hypothetical screening cohort

Based on ONS data, the screening-age (50-74) population in England in between 2012-2017 included an average of 7,734,486 women and 7,407,102 men, with 6,938 cases of CRC in women (overall crude incidence rate 90 per 100,000 person years at risk), and 10,388 cases of CRC in men (overall crude incidence rate 140 per 100,000 person years at risk).

If a broader age range of 40-74 was considered for risk-based screening, 11,549,934 women would be eligible for screening with 7,572 cases of CRC (overall incidence rate 66 per 100,000 person years at risk), and 11,148,784 men with 11,073 cases of CRC (overall incidence rate 99 per 100,000 person years at risk) in men in this age group.

Applying risk-based screening, including individuals with a threshold of over 0.57% ten-year risk to the current screening programme would result in 11,990,321 individuals (6,139,016 women and 5,851,907 men) being eligible for screening, with 16,729 cancer cases (6,784 in women and 9,949 in men). This is a 3.4% reduction in cancers detected for a 20.8% reduction in numbers of screening participants compared to age-based screening.

Extending risk-based screening to 40 year olds would result in 13,008,636 individuals (6,654,379 women and 6,355,476 men) participating in screening, with 17,432 detectable cancer cases (7,024 in women and 10,409 in men). This results in 0.6% fewer cancers detected for a 14.1% reduction in numbers of screening participants compared to age-based screening.

Thus both of these risk-based approaches would reduce numbers participating in screening compared to the current age-based screening programme, with extension of the age range to 40 years having essentially no impact on the total number of cancers detected.

5.4 Polygenic risk scores in UK Biobank

5.4.1 Polygenic risk score construction

5.4.1.1 GWAS-significant polygenic risk scores

Of SNPs previously reported as associated with CRC at $p<5\times10^{-8}$ in European populations, 50 SNPs replicated at GWAS-significance in my meta-analysis and included in the GWAS-sig PRS. Table 5.1 lists these SNPs with their risk alleles and adjusted betas and odds ratios.

Table 5.1: Risk alleles and adjusted effect sizes of SNPs included in the GWAS-significant PRS

Locus	rsID	Chromosome	Position	Risk Allele	Base P value	Adjusted Beta	Adjusted OR
1p34.3	rs61776719	1	38461319	C	4.13e-08	0.0711379	1.073729
1p32.3	rs12143541	1	55247852	G	3.39e-08	0.0950466	1.099710
1q25.3	rs4546885	1	183025555	G	1.72e-11	0.0834519	1.087033
1q41	rs6658977	1	222049820	T	3.97e-08	0.0693903	1.071854
2q35	rs13020391	2	219184436	C	2.78e-10	0.0805173	1.083848
3p22.1	rs35470271	3	40915239	G	3.15e-09	0.0994733	1.104589
3q13.2	rs12635946	3	112916918	C	4.16e-12	0.0873575	1.091287
4q24	rs17035289	4	106048291	T	7.03e-10	0.0998087	1.104960
4q31.21	rs75686861	4	145621328	A	1.89e-08	0.1177675	1.124982
5p13.1	rs1445011	5	40280202	C	7.05e-13	0.0947643	1.099400
5q31.1	rs639933	5	134467751	C	3.44e-08	0.0720464	1.074705
6p21.31	rs16878812	6	35569562	A	4.34e-08	0.1062480	1.112098
6p21.2	rs1321310	6	36623124	C	7.65e-10	0.0864985	1.090350
6p12.1	rs62404966	6	55712124	C	2.88e-08	0.0790045	1.082209
6q21	rs6928864	6	105966894	C	1.37e-08	0.1253851	1.133585
7p12.3	rs3801081	7	47511161	G	6.28e-09	0.0753624	1.078275
8q23.3	rs16892766	8	117630683	C	7.35e-28	0.2259679	1.253535
8q24.21	rs6983267	8	128413305	G	7.59e-39	0.1566475	1.169583
9p21.3	rs1412834	9	22110131	T	4.56e-15	0.0931829	1.097662
10p14	rs7894531	10	8734761	G	2.91e-21	0.1225597	1.130387
10q22.3	rs704017	10	80819132	G	1.13e-14	0.1020743	1.107466
10q24.2	rs2193352	10	101346609	G	2.4e-13	0.1090038	1.115167
11q13.4	rs57796856	11	74338355	T	6.25e-13	0.0860008	1.089807
11q13.4	rs4944940	11	74415252	G	2.49e-16	0.2617104	1.299150
11q23.1	rs3087967	11	111156836	T	9.41e-28	0.1418661	1.152422
12p13.31	rs10849438	12	6412036	G	2.17e-08	0.1117075	1.118186
12q13.12	rs11169572	12	51216890	C	1.49e-12	0.0866486	1.090513
12q24.12	rs597808	12	111973358	G	1.09e-12	0.0864705	1.090319
12q24.21	rs7315438	12	115891403	T	4.02e-11	0.0808168	1.084172
13q13.3	rs12427600	13	37460648	C	1.71e-09	0.0844332	1.088100
13q22.1	rs45597035	13	73649152	A	1.26e-08	0.0733021	1.076056
13q22.3	rs1330889	13	78609615	C	2.05e-08	0.1029962	1.108487
14q22.2	rs35107139	14	54419106	C	1.11e-10	0.0848918	1.088599
15q13.3	rs16969681	15	32993111	T	1.07e-20	0.1909614	1.210413
15q13.3	rs73376930	15	33012502	G	3.24e-25	0.1512445	1.163281
15q13.3	rs17816465	15	33156386	A	1.38e-10	0.0969737	1.101831
15q26.1	rs7495132	15	91172901	T	7.74e-09	0.1078986	1.113935
16q23.2	rs61336918	16	80007266	A	1.57e-11	0.0918747	1.096227
16q24.1	rs899244	16	86700030	T	8.13e-09	0.0849503	1.088663
17p12	rs1078643	17	10707241	A	8.25e-09	0.0894635	1.093587
18q21.1	rs7226855	18	46454048	A	2.44e-57	0.1937059	1.213739
19q13.11	rs73039434	19	33524919	T	5.9e-15	0.2636508	1.301674
19q13.33	rs12979278	19	49218602	T	8.18e-09	0.0716786	1.074310
20p12.3	rs961253	20	6404281	A	4.23e-17	0.1036345	1.109195
20p12.3	rs994308	20	6603622	C	2.55e-10	0.0778976	1.081012
20p12.3	rs6085661	20	6693128	T	1.45e-12	0.0863733	1.090213
20q13.13	rs6066825	20	47340117	A	7.02e-12	0.0870635	1.090966
20q13.13	rs4811050	20	48980670	A	1.63e-10	0.0998156	1.104967
20q13.13	rs6091213	20	49384745	C	3.75e-08	0.0775145	1.080598
20q13.33	rs1741640	20	60932414	C	5.38e-25	0.1615189	1.175295

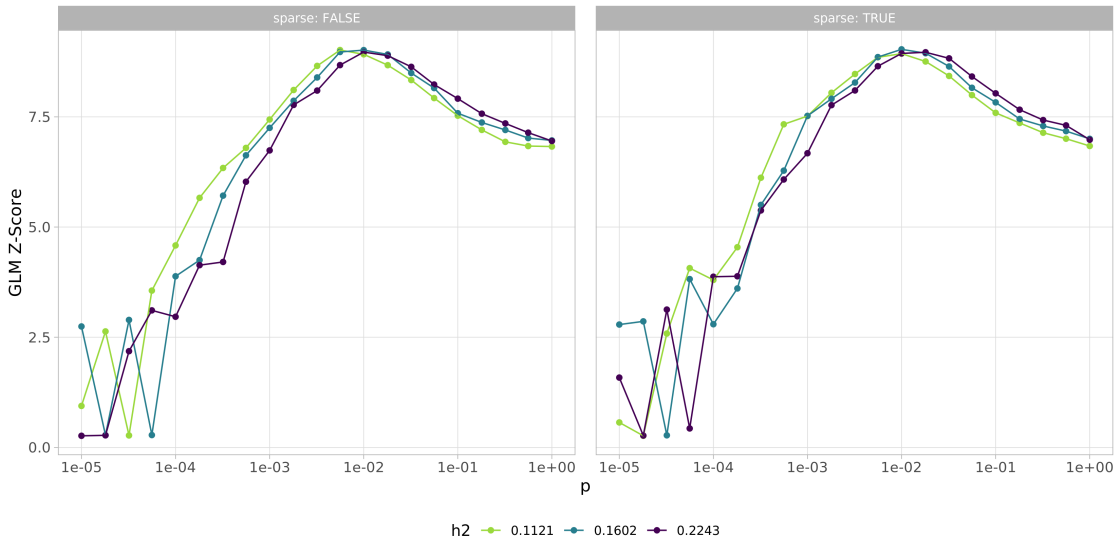

Figure 5.5: Z-scores across tuning parameters for LDpred2 grid models

5.4.1.2 Clumping and Thresholding polygenic risk scores

The top performing C+T PRS was obtained with clumping r^2 of 0.5, clumping window of 1000kb, and p value threshold of 0.01897556. The C+T PRS includes 13,446 SNPs with an AUROC in the Training Cohort of 0.608. The SCT model includes 194,756 SNPs.

5.4.1.3 LDpred2 polygenic risk scores

The heritability estimate (on the observed scale) from dataset is 0.1602065. Figure 5.5 shows the Z-scores across the range of tuning parameters for LDpred2-grid models. The optimal parameters for these models were $prop_{causal}$ 0.0056, and h^2 of 0.1121 for the non-sparse model, which contained all 1,104,409 SNPs and $prop_{causal}$ 0.0056, h^2 0.1602 and sparsity 0.44137 for the sparse model, which contained 616,956 SNPs.

5.5 Evaluation of polygenic risk scores in logistic regression models

Figure 5.6 shows the standardised distribution plots of the 6 PRS scores in the Test, Geographic Validation and Minority Ethnic Validation Cohorts; Table 5.2 shows

Table 5.2: Standardised Mean (SD) PRS in Test, Geographic Validation and Minority Ethnic Validation Cohorts

	Test			Geographic Validation			Minority Ethnic Validation		
models	Controls	Cases	Δ Mean	Controls	Cases	Δ Mean	Controls	Cases	Δ Mean
LDpred2-inf	-0.01 (1.00)	0.35 (1.03)	0.36	0.10 (1.05)	0.72 (1.18)	0.62	-1.33 (1.23)	-0.95 (1.33)	0.38
LDpred2-grid	-0.01 (1.00)	0.44 (1.00)	0.45	0.09 (1.02)	0.70 (1.07)	0.61	-0.80 (1.00)	-0.36 (1.00)	0.44
LDpred2-grid-sp	-0.01 (1.00)	0.44 (1.00)	0.45	0.09 (1.03)	0.71 (1.09)	0.62	-0.83 (1.01)	-0.38 (1.01)	0.45
SCT	-0.01 (1.00)	0.34 (1.02)	0.35	0.07 (1.02)	0.60 (1.09)	0.53	-1.19 (1.36)	-0.83 (1.38)	0.36
C+T	-0.01 (1.00)	0.34 (1.00)	0.35	0.10 (1.03)	0.65 (1.13)	0.55	-0.56 (1.02)	-0.19 (1.03)	0.37
GWAS-sig	0.00 (1.00)	0.32 (0.99)	0.32	0.01 (1.00)	0.40 (1.00)	0.39	-0.55 (1.03)	-0.31 (1.12)	0.24

Table 5.3: Wald Chisq values for interactions between PRS and age in logistic regression model

	Chi-Square (P)		
model	PRS	age	PRS * age
LDpred2-inf	529 (0.000)	1254 (0.000)	8 (0.004)
LDpred2-grid	860 (0.000)	1254 (0.000)	3 (0.065)
LDpred2-grid-sp	829 (0.000)	1254 (0.000)	3 (0.068)
SCT	500 (0.000)	1254 (0.000)	2 (0.136)
C+T	509 (0.000)	1252 (0.000)	3 (0.064)
GWAS-sig	447 (0.000)	1248 (0.000)	1 (0.457)

the corresponding values. As anticipated the PRS follow the normal distribution in the Test Cohort; deviation from the normal distribution in the Minority Ethnic Validation Cohort reflects the differences in allele frequencies and effects. The greater separation of mean PRS scores between cases and controls in the Geographic Validation cohort indicates greater discrimination in this cohort compared to the Test Cohort. The globally lower PRS in both cases and controls in the Minority Ethnic Validation cohort are indicative of the differences in allele frequency in these populations, while the increase in range likely represents the increased genetic diversity of this population.

Plots of marginal effects of PRS in interaction with age (Figure 5.7) indicate a negative interaction, suggesting a smaller effect of PRS with increasing age. This is less marked for the GWAS-significant model. Interaction terms between age and sex, fitted in a full model with all other predictors (Table 5.3), had a small χ^2, and was significant (at $p<0.01$) for LDpred2-inf model alone. Given the weakness of the interactions, I did not include the interaction term in the models.

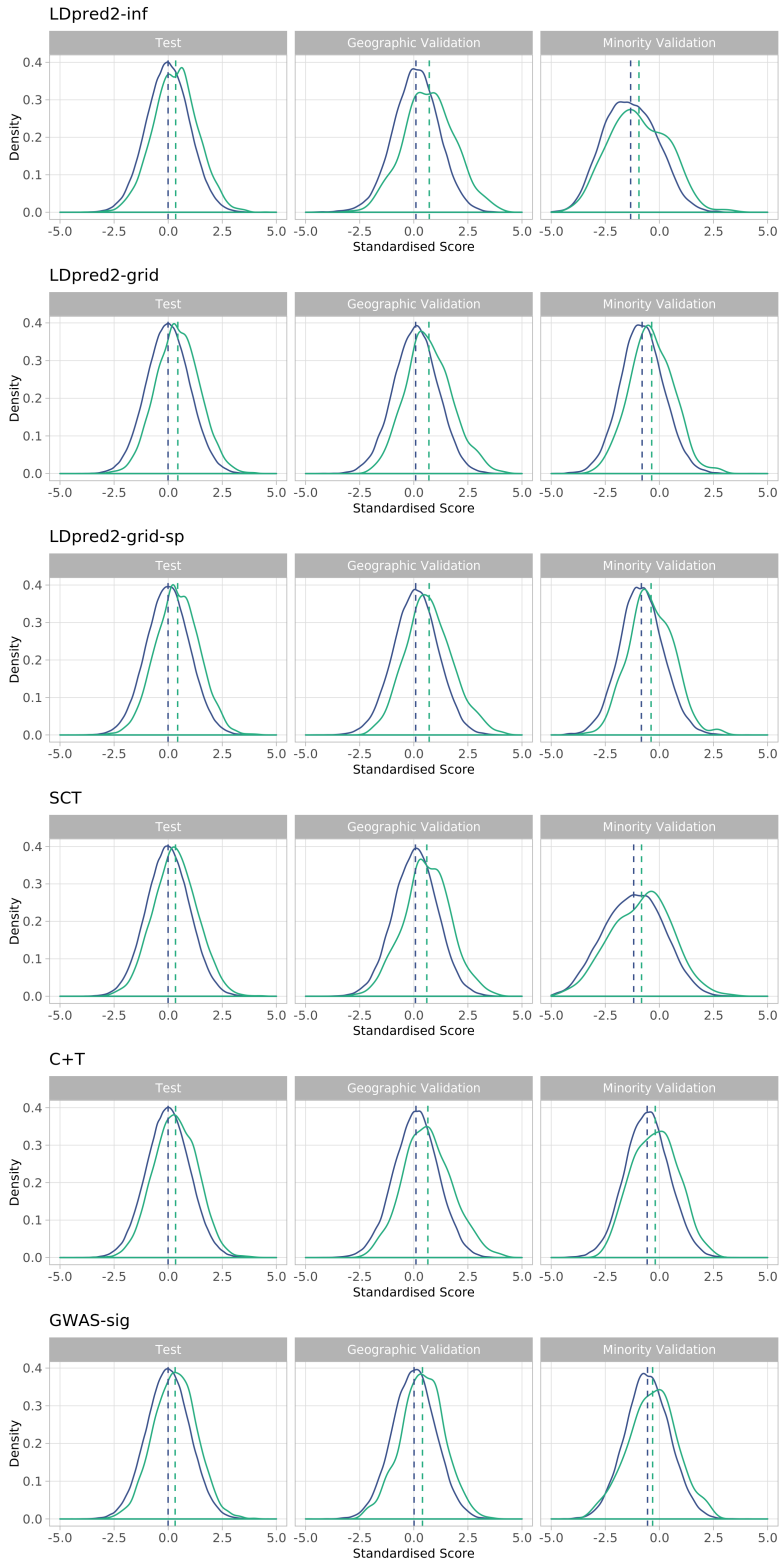

Figure 5.6: Distribution of standardised PRS scores in modelling cohorts. Blue lines represents controls, green lines cases

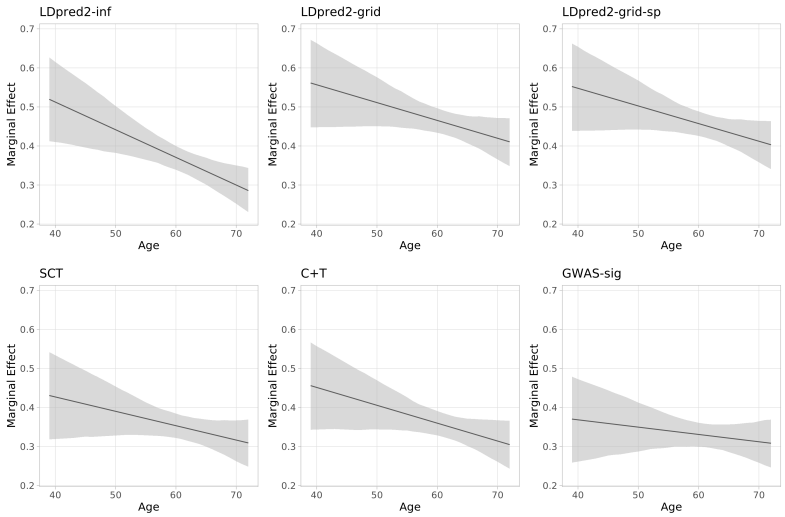

Figure 5.7: Plots of marginal effects of standardised PRS in logistic regression models interaction with age

Table 5.4 shows the adjusted effect sizes for predictors in the PRS logistic regression models. LDpred2 grid-based PRS have greatest effect size, with an OR of ~1.5. Being female reduces odds by around 1/3, whilst age increases risk by about 9% per year. Model specifications for these models are in Appendix B.

5.5.1 Apparent polygenic risk score performance in logistic regression and internal validation

Apparent and internally validated performance of the logistic regression PRS models and Reference model are shown in Table 5.5. The top-performing models across all metrics are those derived using LDpred2 grid approaches. The odds ratio per SD of the LDpred2-grid PRS was 1.584 (95% CI 1.536-1.633), with top C-statistics of 0.717 (0.711-0.725) and explained variation of 6.3% (5.9-6.8%). The sparse and non-sparse LDpred2 grid models performed similarly despite the sparse model having almost half the number of SNPs. All models performed significantly better than the Reference model containing age, sex, array and four PCs, demonstrating that PRS improves CRC prediction above age and sex alone. Internal validation showed minimal bias in all measures.

Table 5.4: Adjusted effects of PRS logistic regression model predictors

Term	Beta	OR	P
LDpred2-inf			
PRS (per SD)	0.361	1.435 (1.391 - 1.480)	0.000
Sex = Female	-0.401	0.670 (0.629 - 0.712)	0.000
Age (per year)	0.086	1.090 (1.085 - 1.095)	0.000
Array = UKBL	0.053	1.055 (0.959 - 1.159)	0.271
PC1	-0.011	0.989 (0.969 - 1.009)	0.262
PC2	0.000	1.000 (0.979 - 1.020)	0.965
PC3	-0.007	0.993 (0.974 - 1.013)	0.504
PC4	-0.012	0.988 (0.977 - 0.999)	0.039
LDpred2-grid			
PRS (per SD)	0.460	1.584 (1.536 - 1.633)	0.000
Sex = Female	-0.404	0.668 (0.628 - 0.710)	0.000
Age (per year)	0.086	1.090 (1.085 - 1.095)	0.000
Array = UKBL	0.053	1.054 (0.959 - 1.159)	0.274
PC1	-0.012	0.988 (0.969 - 1.008)	0.243
PC2	0.000	1.000 (0.979 - 1.020)	0.965
PC3	-0.007	0.993 (0.973 - 1.013)	0.484
PC4	-0.009	0.991 (0.980 - 1.002)	0.112
LDpred2-grid-sp			
PRS (per SD)	0.452	1.571 (1.524 - 1.620)	0.000
Sex = Female	-0.404	0.668 (0.628 - 0.710)	0.000
Age (per year)	0.086	1.090 (1.085 - 1.095)	0.000
Array = UKBL	0.052	1.054 (0.959 - 1.158)	0.278
PC1	-0.012	0.988 (0.969 - 1.008)	0.251
PC2	0.000	1.000 (0.979 - 1.020)	0.964
PC3	-0.007	0.993 (0.974 - 1.013)	0.488
PC4	-0.010	0.990 (0.979 - 1.001)	0.082
SCT			
PRS (per SD)	0.349	1.417 (1.375 - 1.461)	0.000
Sex = Female	-0.405	0.667 (0.627 - 0.710)	0.000
Age (per year)	0.086	1.090 (1.085 - 1.095)	0.000
Array = UKBL	0.058	1.060 (0.964 - 1.165)	0.229
PC1	-0.012	0.988 (0.968 - 1.008)	0.225
PC2	0.000	1.000 (0.980 - 1.021)	0.990
PC3	-0.007	0.993 (0.974 - 1.013)	0.489
PC4	-0.003	0.997 (0.986 - 1.009)	0.647
C+T			
PRS (per SD)	0.354	1.425 (1.382 - 1.470)	0.000
Sex = Female	-0.403	0.668 (0.628 - 0.711)	0.000
Age (per year)	0.086	1.090 (1.085 - 1.095)	0.000
Array = UKBL	0.054	1.055 (0.960 - 1.160)	0.266
PC1	-0.012	0.988 (0.968 - 1.008)	0.223
PC2	0.000	1.000 (0.980 - 1.021)	0.999
PC3	-0.009	0.991 (0.972 - 1.011)	0.388
PC4	-0.007	0.993 (0.982 - 1.005)	0.258
GWAS-sig			
PRS (per SD)	0.329	1.390 (1.348 - 1.433)	0.000
Sex = Female	-0.405	0.667 (0.627 - 0.709)	0.000
Age (per year)	0.086	1.090 (1.085 - 1.095)	0.000
Array = UKBL	0.056	1.058 (0.962 - 1.162)	0.244
PC1	-0.014	0.986 (0.966 - 1.006)	0.161
PC2	-0.001	0.999 (0.978 - 1.020)	0.908
PC3	-0.009	0.991 (0.971 - 1.010)	0.353
PC4	0.008	1.008 (0.997 - 1.020)	0.145

Table 5.5: Apparent and internally validated performance of PRS in logistic regression models in the Test Cohort.

	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig	Reference
Number of SNPs	1,104,409	1,104,409	616,956	194,756	13,446	50	0
Apparent performance (95% CI)							
PRS OR per SD	1.435 (1.391 - 1.480)	1.584 (1.536 - 1.633)	1.571 (1.524 - 1.620)	1.417 (1.375 - 1.461)	1.425 (1.382 - 1.470)	1.390 (1.348 - 1.433)	-
C statistic	0.704 (0.697 - 0.712)	0.717 (0.711 - 0.725)	0.716 (0.710 - 0.723)	0.702 (0.695 - 0.711)	0.704 (0.697 - 0.711)	0.700 (0.693 - 0.707)	0.680 (0.672 - 0.687)
Somers' D_{xy}	0.407 (0.394 - 0.423)	0.435 (0.422 - 0.451)	0.432 (0.419 - 0.446)	0.404 (0.389 - 0.422)	0.407 (0.394 - 0.423)	0.400 (0.386 - 0.414)	0.359 (0.344 - 0.374)
Nagelkerke's R^2 (%)	5.5 (5.1 - 5.9)	6.3 (5.9 - 6.8)	6.2 (5.8 - 6.7)	5.4 (5.0 - 5.9)	5.4 (5.1 - 5.9)	5.3 (4.9 - 5.7)	4.2 (3.8 - 4.6)
R^2_{PRS} (%)	1.3 (1.1 - 1.5)	2.1 (1.9 - 2.4)	2.0 (1.8 - 2.3)	1.2 (1.0 - 1.4)	1.2 (1.0 - 1.5)	1.1 (0.9 - 1.3)	-
Scaled Brier Score (%)	0.87	1.05	1.03	0.86	0.85	0.83	0.6
Internally validated performance							
C statistic	0.703	0.717	0.716	0.701	0.703	0.700	0.679
Somers' D_{xy}	0.406	0.434	0.432	0.403	0.406	0.400	0.358
Nagelkerke's R^2 (%)	5.4	6.3	6.2	5.4	5.4	5.3	4.2
Calibration slope	0.996	0.997	0.998	0.996	0.995	0.999	0.996
Scaled Brier Score (%)	0.85	0.94	1.06	0.84	0.76	0.85	0.58

LDpred2-inf – LDpred2 infinitesimal model; LDpred2-grid – LDpred2 grid model; LDpred2-grid-sp – LDpred2 sparse grid model;
SCT – stacked clumping and thresholding; C+T – clumping and thresholding; GWAS-sig – GWAS significant.

Without adjusting for sex and age in the models, performance is considerably poorer, as would be anticipated (Table 5.6). The top C-statistic is 0.626 (95% CI 0.618-0.634), explained variation is 2.1% (1.8-2.4%), and scaled Brier score is 0.34%.

Table 5.6: Apparent performance of PRS in LR models in the Test Cohort with and without adjustment for sex and age. Internal validation used 500 bootstrap samples.

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig
With sex and age						
C statistic	0.704 (0.697 - 0.712)	0.717 (0.711 - 0.725)	0.716 (0.710 - 0.723)	0.702 (0.695 - 0.711)	0.704 (0.697 - 0.711)	0.700 (0.693 - 0.707)
Somers' D_{xy}	0.407 (0.394 - 0.423)	0.435 (0.422 - 0.451)	0.432 (0.419 - 0.446)	0.404 (0.389 - 0.422)	0.407 (0.394 - 0.423)	0.400 (0.386 - 0.414)
Nagelkerke's R^2 (%)	5.5 (5.1 - 5.9)	6.3 (5.9 - 6.8)	6.2 (5.8 - 6.7)	5.4 (5.0 - 5.9)	5.4 (5.1 - 5.9)	5.3 (4.9 - 5.7)
Scaled Brier Score (%)	0.87	1.05	1.03	0.86	0.85	0.83
Without sex and age						
C statistic	0.597 (0.589 - 0.606)	0.626 (0.618 - 0.634)	0.623 (0.614 - 0.632)	0.594 (0.587 - 0.603)	0.597 (0.589 - 0.606)	0.592 (0.584 - 0.601)
Somers' D_{xy}	0.194 (0.178 - 0.212)	0.251 (0.235 - 0.268)	0.247 (0.229 - 0.264)	0.189 (0.175 - 0.206)	0.193 (0.178 - 0.211)	0.185 (0.169 - 0.202)
Nagelkerke's R^2 (%)	1.3 (1.1 - 1.5)	2.1 (1.8 - 2.4)	2.0 (1.8 - 2.3)	1.2 (1.0 - 1.5)	1.3 (1.1 - 1.5)	1.1 (0.9 - 1.3)
Scaled Brier Score (%)	0.21	0.34	0.33	0.19	0.19	0.17

LDpred2-inf – LDpred2 infinitesimal model; LDpred2-grid – LDpred2 grid model; LDpred2-grid-sp – LDpred2 sparse grid model; SCT – stacked clumping and thresholding; C+T – clumping and thresholding; GWAS-sig – GWAS significant.

5.5.2 External validation of polygenic risk score logistic regression models

In the Geographic Validation Cohort, logistic regression PRS models show improved discrimination, explained variation and overall fit compared to the Test Cohort (Table 5.7). The highest performance statistics are seen with the LDpred-grid-sp model, with a C-statistic of 0.733 (95% CI 0.710-0.753), explained variation of 7.6% (6.1-8.9%) and scaled Brier score of 1.66%. Whilst confidence intervals for many of the PRS models overlap, paired t-tests suggest statistically significant differences between models across almost all metrics.

In terms of calibration, all models under-predict CRC risk, with CITL >0. This is particularly evident in the highest risk groups (Figure 5.8A), and is improved by recalibration (Figure 5.8B). Genome-wide PRS models also tend to be overfitted (i.e. show insufficient variation at the extremes, indicated by calibration slopes >1), though confidence intervals span 1 for all but the LDpred2-inf model.

Table 5.7: Performance of PRS in logistic regression models in Validation Cohorts. Values show performance indices plus 95% confidence intervals.

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig	Reference
Geographic Validation Cohort							
C statistic	0.726 (0.704 - 0.748)	0.732 (0.710 - 0.752)	0.733 (0.710 - 0.753)	0.718 (0.696 - 0.739)	0.719 (0.696 - 0.740)	0.703 (0.679 - 0.724)	0.677 (0.654 - 0.699)
Somers' D_{xy}	0.452 (0.408 - 0.496)	0.464 (0.420 - 0.504)	0.466 (0.421 - 0.507)	0.436 (0.392 - 0.477)	0.438 (0.392 - 0.480)	0.405 (0.358 - 0.447)	0.353 (0.308 - 0.397)
Nagelkerke's R^2 (%)	7.0 (5.7 - 8.4)	7.6 (6.1 - 8.9)	7.6 (6.1 - 8.9)	6.4 (5.0 - 7.7)	6.6 (5.2 - 7.9)	5.4 (4.0 - 6.7)	3.8 (2.6 - 5.0)
Calibration slope	1.137 (1.010 - 1.268)	1.091 (0.967 - 1.199)	1.104 (0.980 - 1.213)	1.076 (0.946 - 1.200)	1.098 (0.958 - 1.222)	0.994 (0.861 - 1.113)	0.936 (0.795 - 1.075)
CITL	0.206 (0.120 - 0.272)	0.198 (0.113 - 0.262)	0.199 (0.115 - 0.264)	0.194 (0.107 - 0.260)	0.195 (0.110 - 0.261)	0.191 (0.104 - 0.258)*	0.191 (0.105 - 0.257)*
Scaled Brier Score (%)	1.48	1.64	1.66	1.27	1.38	1.08	0.67
Minority Ethnic Validation Cohort							
C statistic	0.588 (0.545 - 0.627)	0.602 (0.558 - 0.640)	0.601 (0.559 - 0.640)	0.589 (0.546 - 0.626)	0.597 (0.554 - 0.636)	0.587 (0.543 - 0.624)	0.585 (0.542 - 0.623)
Somers' D_{xy}	0.176 (0.090 - 0.254)	0.203 (0.116 - 0.279)	0.203 (0.118 - 0.281)	0.179 (0.093 - 0.253)	0.195 (0.108 - 0.271)	0.174 (0.086 - 0.247)	0.171 (0.084 - 0.245)
Calibration slope	0.175 (0.096 - 0.258)	0.204 (0.122 - 0.288)	0.208 (0.126 - 0.294)	0.161 (0.088 - 0.240)	0.195 (0.110 - 0.281)	0.143 (0.071 - 0.213)	0.144 (0.069 - 0.217)
CITL	1.299 (1.155 - 1.417)	1.336 (1.194 - 1.456)	1.325 (1.183 - 1.446)	1.360 (1.217 - 1.479)	1.310 (1.167 - 1.429)	1.392 (1.251 - 1.511)	1.343 (1.200 - 1.459)
Scaled Brier Score (%)	-0.02	0.04	0.03	-0.06	-0.06	-0.07	-0.15

LDpred2-inf – LDpred2 infinitesimal model; LDpred2-grid – LDpred2 grid model; LDpred2-grid-sp – LDpred2 sparse grid model;
SCT – stacked clumping and thresholding; C+T – clumping and thresholding; GWAS-sig – GWAS significant.
Pairwise comparisons of performance metrics in validation cohorts were all significantly different P<0.001 except comparisons marked *P=0.6.
R^2 for all models in the Minority Ethnic Validation Cohort <0 (indicating poorer performance than a model with no explanatory variables)

In the Minority Ethnic Validation Cohort, all models perform poorly. As a baseline, the Reference model also performs poorly, with a C-statistic of 0.585 (95% CI 0.542-9.623). Only LDpred2 grid models show any marked improvement above this (C-statistic 0.602 (0.558-0.640) for the LDpred2-grid model). The models are highly under-fitted (calibration slope far below 1) and systematically under-predict CRC risk, as demonstrated by CITL>0 and calibration plots showing predictions lying above the reference line (Figure 5.8C). Following recalibration of the intercept, models are markedly over-fitted (under-predicting risk in lower risk groups, and over-predicting in higher risk groups) (Figure 5.8D).

5.5.3 Subgroup analysis of polygenic risk score logistic regression models

Examination of model performance by sex in the Geographic Validation Cohort shows that models have higher discrimination and explain a greater proportion of variation in males compared to females (Table 5.8). The C-statistics for the LDpred2-grid-sp model in men and women are 0.741 (95% CI 0.715-0.768) and 0.714 (0.684-0.743) respectively, with an additional 2% of variation explained in men (R^2 8.3% (0.066-0.101) compared to 6.1% (4.3-7.8%)). Measures of calibration by sex are variable: calibration slopes are closer to 1 in women in genome-wide models, however calibration plots show that risk is under-predicted to a greater extent in women in the top tenth of risk compared to men (Figure 5.9), and CITL is further from 0 in women.

In individuals with a first degree family history in the Geographic Validation Cohort, discrimination is poorer than the cohort overall (LDpred2-grid-sp C-statistic of 0.706 (95% CI 0.647-0.758) compared to 0.733 (0.710-0.753)), with a lower proportion of variation explained (3.6% (-0.9-7.5%) compared to 7.6% (6.1-8.9%), Table 5.9), and scaled Brier score of 1.45 compared to 1.66 (see Table 5.8). Calibration is also poorer, with predictions systematically too low (CITL >0, Figure 5.10).

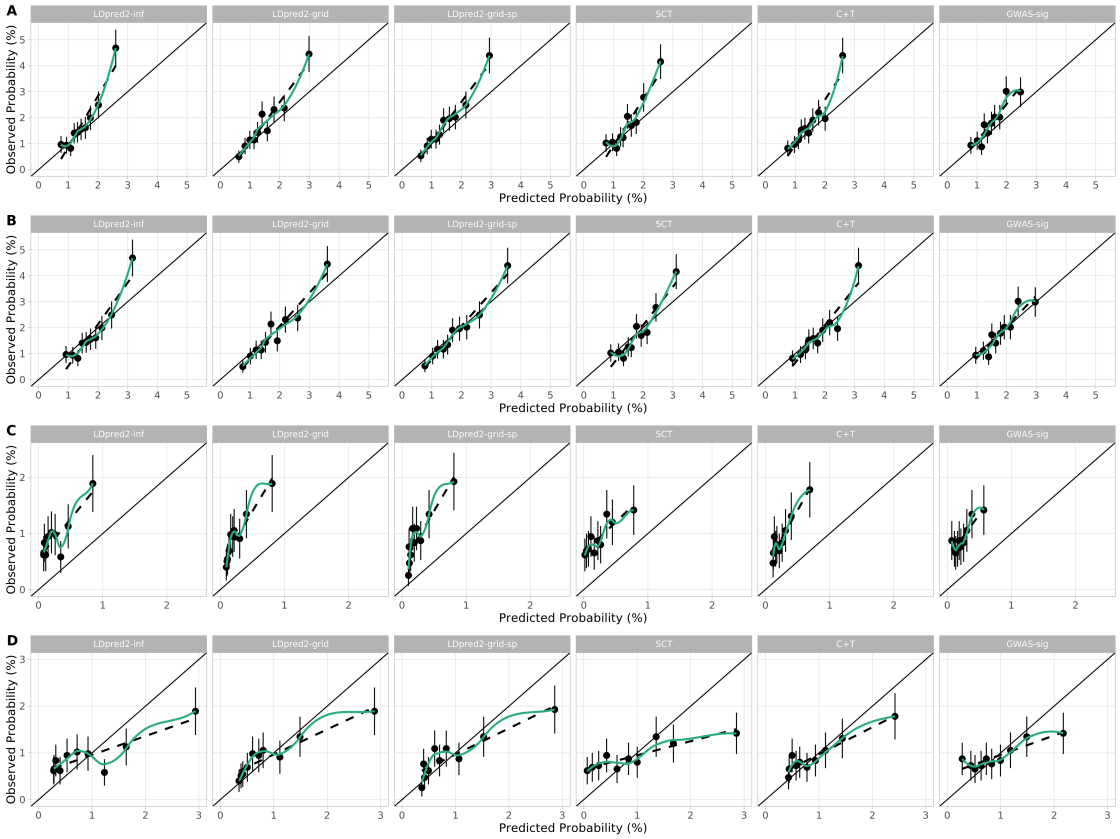

Figure 5.8: Calibration plots for PRS in logistic regression models in the Geographic Validaiton Cohort before (A) and after (B) recalibration, and in the Minority Ethnic Validation Cohort before (C) and after (D) recalibration.

Figure 5.11 demonstrates that calibration by age is near-identical for all models, and shows that whilst models are well calibrated in younger age bands, they under-predict risk to a greater extent in the oldest age group. I observed a jump in observed risk for 55-59 year olds which results in apparent miscalibration in this age group.

Table 5.8: Performance of PRS logistic regression models by sex in the Geographic Validation Cohort

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig	Reference
Males							
C statistic	0.731 (0.705 - 0.760)	0.740 (0.716 - 0.767)	0.741 (0.715 - 0.768)	0.728 (0.702 - 0.753)	0.726 (0.702 - 0.755)	0.716 (0.689 - 0.743)	0.686 (0.662 - 0.714)
Somers' D_{xy}	0.463 (0.410 - 0.519)	0.481 (0.433 - 0.534)	0.481 (0.431 - 0.536)	0.455 (0.404 - 0.507)	0.453 (0.403 - 0.510)	0.433 (0.378 - 0.486)	0.372 (0.324 - 0.428)
Nagelkerke's R^2	7.6 (6.0 - 9.3)	8.3 (6.6 - 10.1)	8.3 (6.6 - 10.1)	7.2 (5.5 - 8.7)	7.2 (5.6 - 8.9)	6.6 (5.0 - 8.2)	4.6 (3.2 - 6.1)
Scaled Brier Score (%)	1.731	1.895	1.904	1.473	1.606	1.374	0.839
CITL	0.186 (0.075 - 0.287)	0.178 (0.068 - 0.279)	0.180 (0.070 - 0.281)	0.174 (0.067 - 0.275)	0.176 (0.066 - 0.278)	0.170 (0.061 - 0.273)	0.171 (0.064 - 0.275)
Calibration Slope	1.216 (1.047 - 1.409)	1.171 (1.025 - 1.343)	1.182 (1.034 - 1.357)	1.178 (1.006 - 1.354)	1.187 (1.026 - 1.371)	1.137 (0.968 - 1.320)	1.067 (0.900 - 1.277)
Females							
C statistic	0.709 (0.680 - 0.739)	0.712 (0.682 - 0.741)	0.714 (0.684 - 0.743)	0.694 (0.666 - 0.723)	0.699 (0.669 - 0.731)	0.673 (0.643 - 0.703)	0.648 (0.621 - 0.674)
Somers' D_{xy}	0.419 (0.360 - 0.477)	0.423 (0.365 - 0.481)	0.427 (0.368 - 0.486)	0.387 (0.331 - 0.447)	0.397 (0.338 - 0.462)	0.346 (0.287 - 0.406)	0.295 (0.242 - 0.348)
Nagelkerke's R^2	5.7 (4.0 - 7.4)	6.0 (4.1 - 7.7)	6.1 (4.3 - 7.8)	4.8 (3.2 - 6.6)	5.2 (3.3 - 6.9)	3.4 (1.8 - 5.1)	2.2 (0.8 - 3.5)
Scaled Brier Score (%)	1.039	1.201	1.221	0.896	0.978	0.588	0.336
CITL	0.230 (0.111 - 0.352)	0.221 (0.100 - 0.343)	0.223 (0.101 - 0.344)	0.217 (0.098 - 0.339)	0.218 (0.098 - 0.340)	0.215 (0.097 - 0.342)	0.213 (0.093 - 0.336)
Calibration Slope	1.102 (0.919 - 1.292)	1.035 (0.881 - 1.196)	1.055 (0.896 - 1.214)	1.002 (0.839 - 1.185)	1.041 (0.863 - 1.236)	0.862 (0.700 - 1.036)	0.804 (0.642 - 0.976)

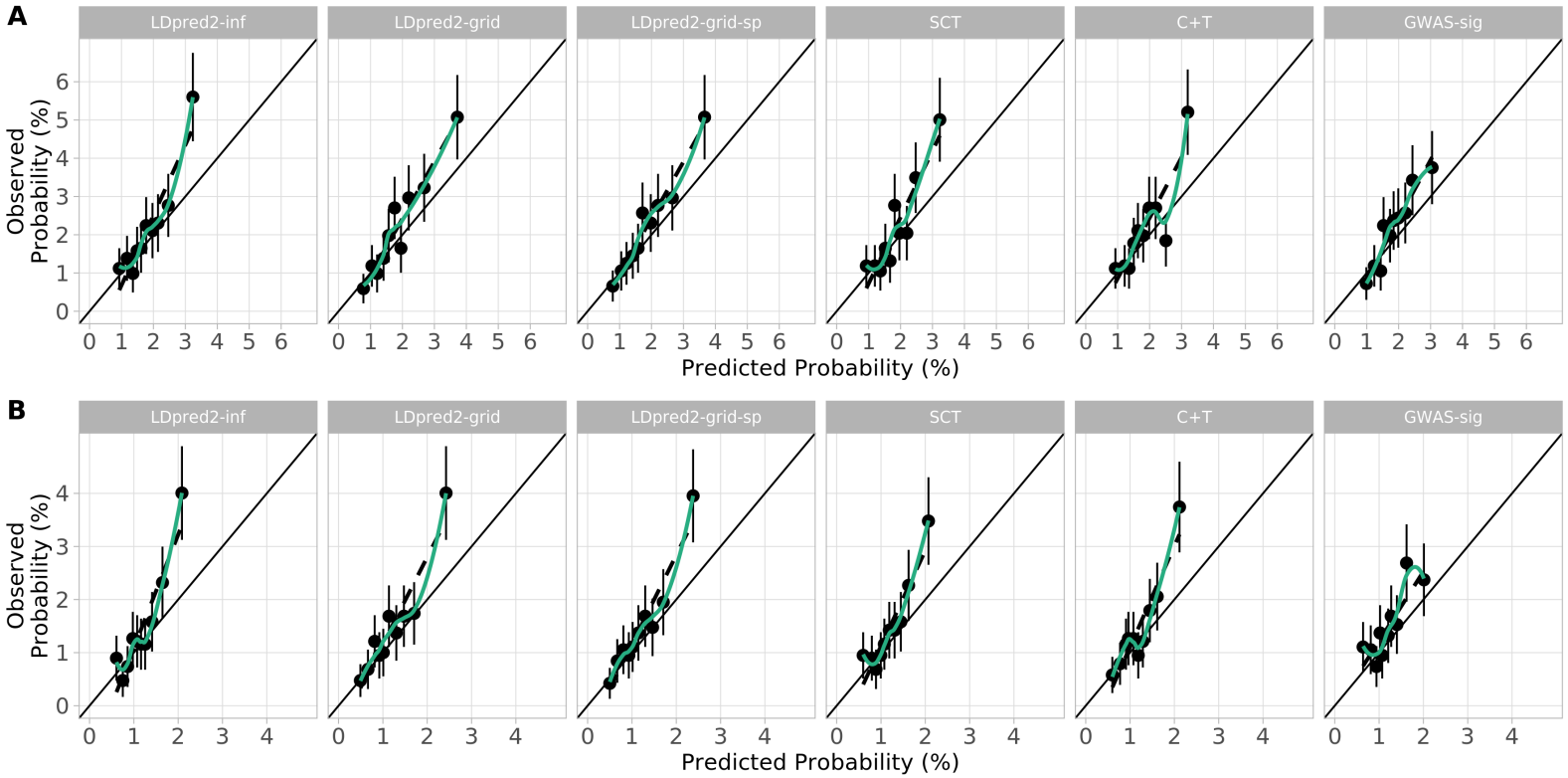

Figure 5.9: Calibration plots for PRS in logistic regression models in males (A) and females (B) in the Geographic Validation Cohort

Table 5.9: Performance of PRS logistic regression models in individuals with a family history of CRC in the Geographic Validation Cohort

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig	Reference
C statistic	0.697 (0.637 - 0.748)	0.701 (0.642 - 0.754)	0.706 (0.647 - 0.758)	0.685 (0.625 - 0.738)	0.703 (0.646 - 0.752)	0.668 (0.608 - 0.721)	0.653 (0.593 - 0.703)
Somers' D_{xy}	0.394 (0.275 - 0.496)	0.402 (0.283 - 0.509)	0.412 (0.293 - 0.515)	0.369 (0.251 - 0.475)	0.406 (0.292 - 0.504)	0.335 (0.217 - 0.443)	0.306 (0.186 - 0.406)
Nagelkerke's R^2	2.4 (-2.2 - 6.0)	3.4 (-1.1 - 7.4)	3.6 (-0.9 - 7.5)	1.8 (-2.8 - 5.8)	2.9 (-1.7 - 6.6)	0.5 (-4.1 - 4.6)	‘-1.1 (-5.4 - 2.5)‘
Scaled Brier Score (%)	1.074	1.449	1.447	1.031	1.138	0.767	0.239
CITL	0.658 (0.462 - 0.827)	0.607 (0.409 - 0.771)	0.613 (0.414 - 0.776)	0.643 (0.451 - 0.812)	0.644 (0.453 - 0.810)	0.633 (0.443 - 0.809)	0.677 (0.483 - 0.851)
Calibration Slope	1.021 (0.714 - 1.322)	0.971 (0.683 - 1.259)	0.997 (0.714 - 1.286)	0.948 (0.633 - 1.246)	1.052 (0.745 - 1.363)	0.838 (0.518 - 1.145)	0.822 (0.479 - 1.164)

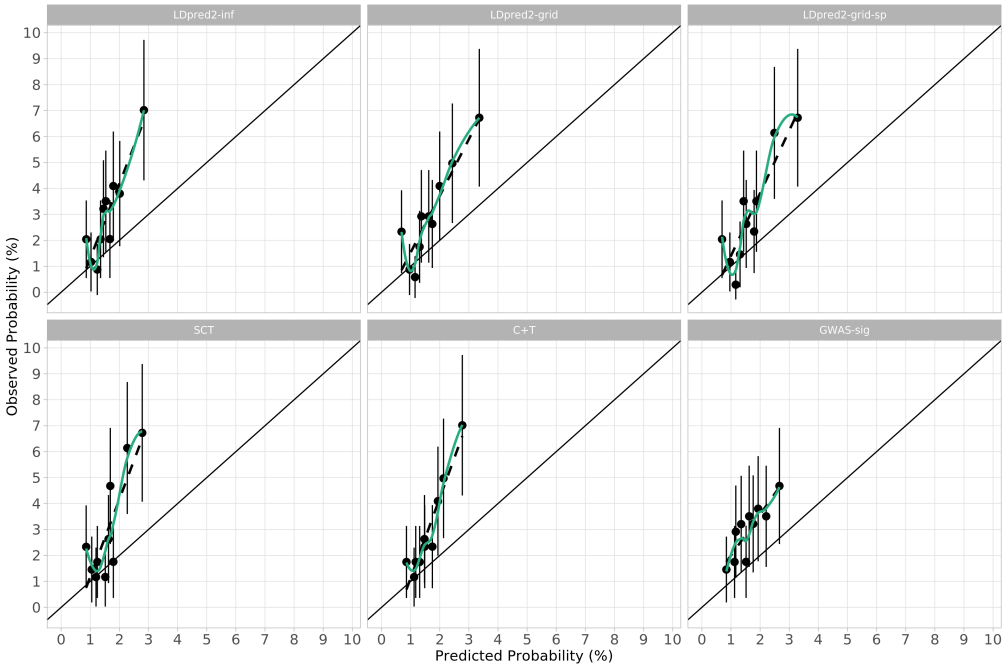

Figure 5.10: Calibration plots for PRS in logistic regression models in individuals with a family history of CRC in the Geographic Validation Cohort

5.6 Evaluation of polygenic risk scores in Cox regression models

Tests for adherence to Cox proportional hazards assumptions indicate that proportional hazards assumptions hold. In $log(-log(survival))$ plots (Figure 5.12) parallel lines are evident for the majority of predictors (though for the C+T model and age these were less convincing). These plots can be unreliable for continuous variables, and χ^2 tests for deviation from proportional hazards were not significant, with plots indicating fulfilment of proportional hazards assumptions (see Appendix C).

Wald χ^2 tests of significance of interactions between age and PRS in Cox models (Table 5.10) show significant interactions (P<0.01) for LDpred2-grid, LDpred2-grid-sp and C+T models. However, given the weakness of the interaction terms relative to the other predictors, and consistency of effect, I elected not to include these in the models.

Plots of marginal effects of PRS in interaction with age (Figure 5.13) show a negative quantitative interaction between PRS and age (this is, PRS increase across

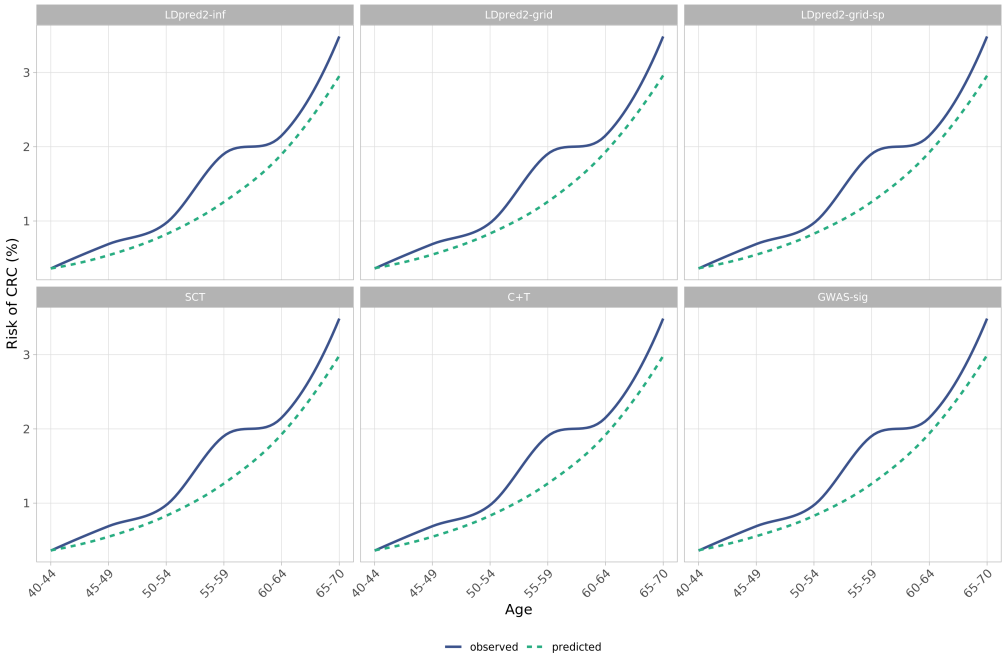

Figure 5.11: Calibration plots for PRS in logistic regression models by age in the Geographic Validation Cohort

Table 5.10: Wald Chi-Square values for interactions between PRS and age in Cox models

	Chi-Square (P)		
model	PRS	age	PRS * age
LDpred2-inf	207 (0.000)	575 (0.000)	4 (0.038)
LDpred2-grid	428 (0.000)	578 (0.000)	9 (0.003)
LDpred2-grid-sp	405 (0.000)	577 (0.000)	8 (0.005)
SCT	222 (0.000)	576 (0.000)	4 (0.035)
C+T	242 (0.000)	576 (0.000)	9 (0.003)
GWAS-sig	225 (0.000)	574 (0.000)	7 (0.011)

all age ranges, but to a lesser extent as age increases).

Table 5.11 shows the effect sizes for Cox PRS model predictors. As with logistic regression models, LDpred2-grid PRS have greatest effect size, with HR of 1.563 (95% CI 1.498-1.631), whilst age increases risk by 8.3% per year, and being female reduces risk by just over 1/3. Model specifications for Cox models are in Appendix B.

Table 5.11: Adjusted effects of PRS Cox model predictors

Term	Beta	HR	P
LDpred2-inf			
PRS (per SD)	0.313	1.368 (1.310 - 1.428)	0.000
Sex = Female	-0.457	0.633 (0.581 - 0.690)	0.000
Age (per year)	0.079	1.083 (1.076 - 1.090)	0.000
Array = UKBL	0.074	1.077 (0.946 - 1.228)	0.263
PC1	-0.012	0.988 (0.961 - 1.016)	0.405
PC2	0.000	1.000 (0.971 - 1.029)	0.988
PC3	-0.007	0.993 (0.966 - 1.021)	0.635
PC4	-0.008	0.992 (0.977 - 1.008)	0.345
LDpred2-grid			
PRS (per SD)	0.447	1.563 (1.498 - 1.631)	0.000
Sex = Female	-0.460	0.631 (0.579 - 0.688)	0.000
Age (per year)	0.080	1.083 (1.076 - 1.090)	0.000
Array = UKBL	0.074	1.076 (0.945 - 1.226)	0.269
PC1	-0.012	0.988 (0.961 - 1.016)	0.405
PC2	0.000	1.000 (0.971 - 1.029)	0.984
PC3	-0.007	0.993 (0.966 - 1.021)	0.638
PC4	-0.007	0.993 (0.977 - 1.009)	0.361
LDpred2-grid-sp			
PRS (per SD)	0.435	1.545 (1.480 - 1.612)	0.000
Sex = Female	-0.460	0.632 (0.579 - 0.689)	0.000
Age (per year)	0.080	1.083 (1.076 - 1.090)	0.000
Array = UKBL	0.073	1.076 (0.944 - 1.226)	0.272
PC1	-0.012	0.988 (0.961 - 1.016)	0.413
PC2	0.000	1.000 (0.971 - 1.029)	0.984
PC3	-0.007	0.993 (0.966 - 1.021)	0.638
PC4	-0.008	0.992 (0.976 - 1.008)	0.321
SCT			
PRS (per SD)	0.321	1.378 (1.321 - 1.438)	0.000
Sex = Female	-0.461	0.631 (0.579 - 0.688)	0.000
Age (per year)	0.080	1.083 (1.076 - 1.090)	0.000
Array = UKBL	0.078	1.081 (0.949 - 1.232)	0.242
PC1	-0.012	0.988 (0.960 - 1.015)	0.379
PC2	0.000	1.000 (0.972 - 1.029)	0.984
PC3	-0.007	0.993 (0.966 - 1.021)	0.626
PC4	0.000	1.000 (0.984 - 1.016)	0.978
C+T			
PRS (per SD)	0.335	1.397 (1.338 - 1.459)	0.000
Sex = Female	-0.459	0.632 (0.580 - 0.689)	0.000
Age (per year)	0.079	1.083 (1.076 - 1.090)	0.000
Array = UKBL	0.074	1.077 (0.945 - 1.227)	0.265
PC1	-0.012	0.988 (0.961 - 1.016)	0.381
PC2	0.000	1.000 (0.972 - 1.029)	0.987
PC3	-0.008	0.992 (0.965 - 1.019)	0.553
PC4	-0.004	0.996 (0.980 - 1.012)	0.601
GWAS-sig			
PRS (per SD)	0.320	1.377 (1.320 - 1.436)	0.000
Sex = Female	-0.461	0.631 (0.578 - 0.688)	0.000
Age (per year)	0.079	1.083 (1.076 - 1.090)	0.000
Array = UKBL	0.076	1.079 (0.947 - 1.230)	0.252
PC1	-0.014	0.986 (0.959 - 1.014)	0.318
PC2	-0.001	0.999 (0.971 - 1.028)	0.948
PC3	-0.009	0.991 (0.964 - 1.019)	0.529
PC4	0.010	1.010 (0.994 - 1.026)	0.218

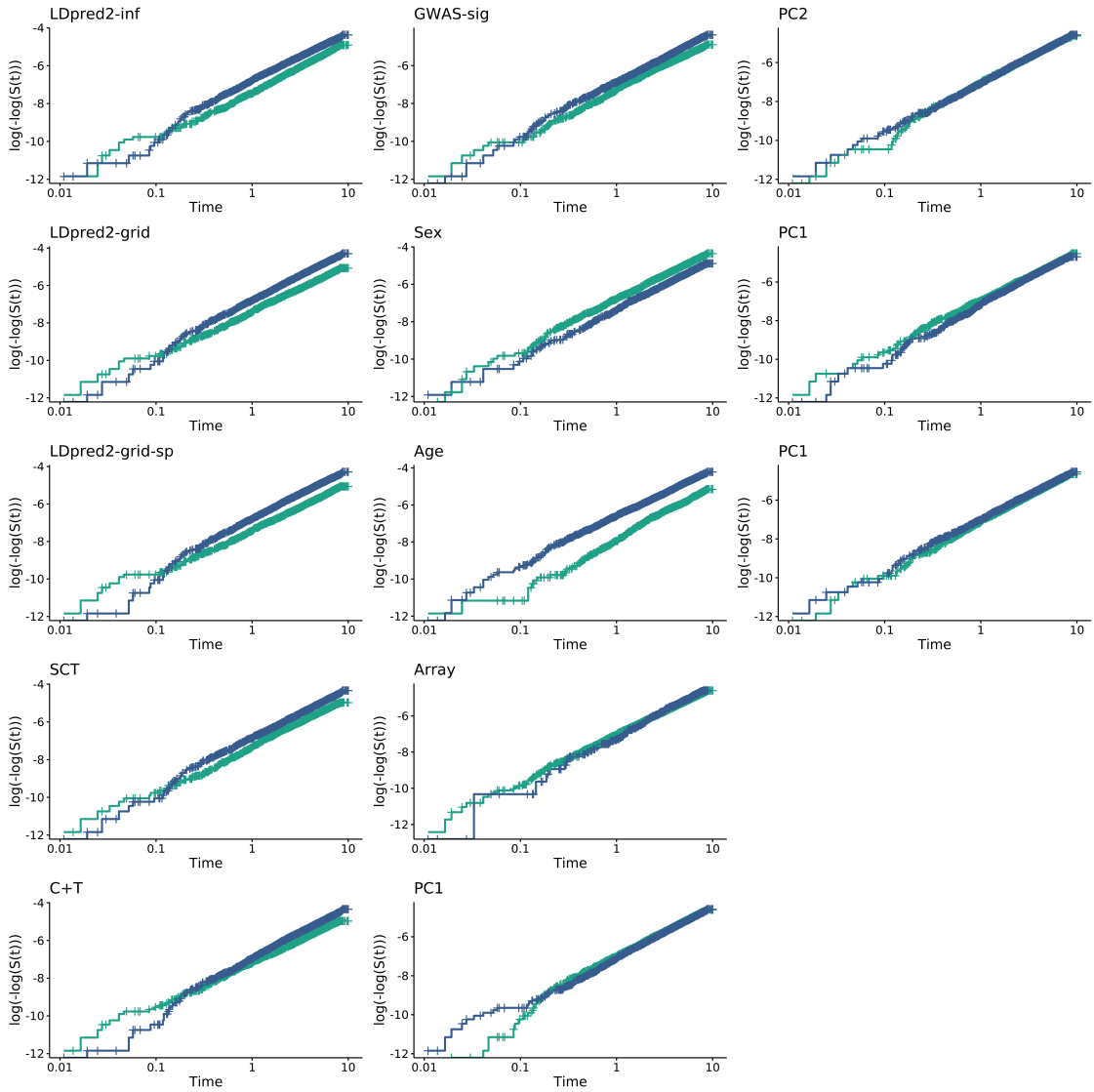

Figure 5.12: Plots of $log(-log(Survival))$ against $log(Survival)$ for PRS model predictors

5.6.1 Apparent polygenic risk score performance in Cox regression model and internal validation

Apparent and internally validated performance of Cox PRS models and the Reference model are detailed in Table 5.12. These closely reflect those of the logistic regression analysis with top performance across all metrics observed with the LDpred2-grid model, and minimal bias in all measures. The C-index for the LDpred2-grid model is 0.714 (0.704-0.726) and the model explains 25.6% (95% CI 23.8-27.7%) of observed variation. Performance of the sparse and non-sparse LDpred2 grid models is again

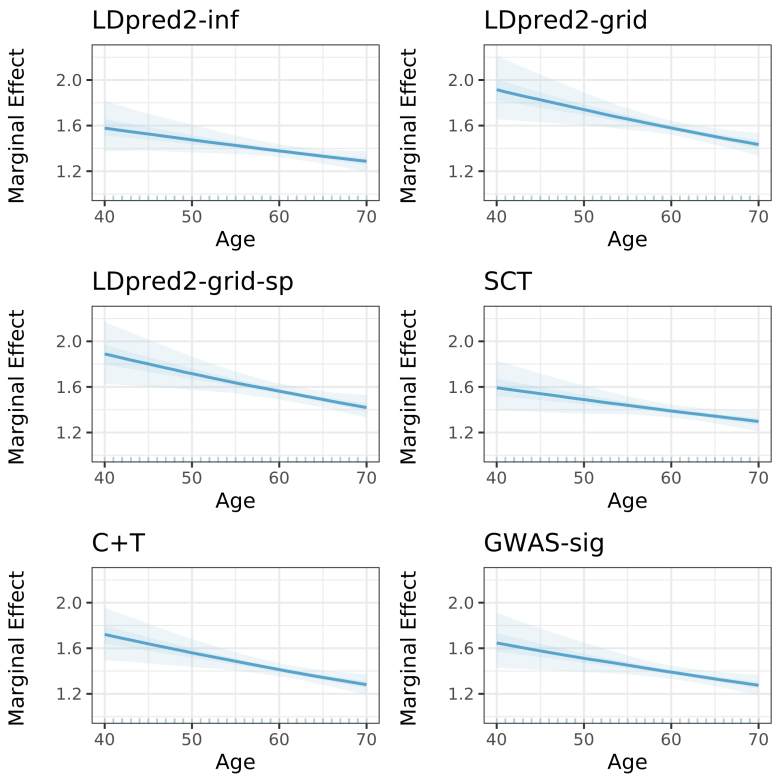

Figure 5.13: Plots of marginal effects of PRS in interaction with age in Cox models

similar. As for the logistic regression models, performance of Cox PRS models unadjusted for age and sex is poor (Table 5.13).

Table 5.12: Apparent and internally validated performance of PRS in Cox models in the Test Cohort

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig	Reference
Apparent performance (95% CI)							
C index	0.696 (0.685 - 0.707)	0.714 (0.704 - 0.726)	0.712 (0.702 - 0.723)	0.695 (0.685 - 0.706)	0.698 (0.689 - 0.709)	0.695 (0.685 - 0.706)	0.675 (0.665 - 0.687)
Somers' D_{xy}	0.391 (0.370 - 0.414)	0.427 (0.409 - 0.451)	0.424 (0.403 - 0.447)	0.391 (0.370 - 0.412)	0.396 (0.378 - 0.417)	0.390 (0.370 - 0.412)	0.350 (0.331 - 0.373)
D statistic	1.085 (1.027 - 1.150)	1.201 (1.143 - 1.268)	1.190 (1.132 - 1.255)	1.096 (1.034 - 1.163)	1.099 (1.043 - 1.162)	1.094 (1.031 - 1.162)	0.961 (0.902 - 1.021)
R^2_D (%)	22.0 (20.1 - 24)	25.6 (23.8 - 27.7)	25.3 (23.4 - 27.3)	22.3 (20.3 - 24.4)	22.4 (20.6 - 24.4)	22.2 (20.2 - 24.4)	18.1 (16.3 - 19.9)
Scaled Brier Score (%)	0.448	0.564	0.550	0.485	0.474	0.505	0.390
Internally validated performance							
C index	0.694	0.713	0.711	0.694	0.697	0.694	0.674
Somers' D_{xy}	0.389	0.425	0.422	0.389	0.393	0.387	0.347
D statistic	1.078	1.194	1.183	1.089	1.091	1.088	0.954
R^2_D (%)	21.7	25.4	25.1	22.1	22.1	22.0	17.8
Scaled Brier Score (%)	0.437	0.551	0.537	0.474	0.463	0.495	0.381
Calibration Slope	0.992	0.994	0.995	0.994	0.992	0.992	0.992

Table 5.13: Apparent performance of PRS in Cox models in the Test Cohort with and without adjustment for sex and age

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig
With sex and age						
C index	0.696 (0.685 - 0.707)	0.714 (0.704 - 0.726)	0.712 (0.702 - 0.723)	0.695 (0.685 - 0.706)	0.698 (0.689 - 0.709)	0.695 (0.685 - 0.706)
Somers' D_{xy}	0.391 (0.370 - 0.414)	0.427 (0.409 - 0.451)	0.424 (0.403 - 0.447)	0.391 (0.370 - 0.412)	0.396 (0.378 - 0.417)	0.390 (0.370 - 0.412)
D statistic	1.085 (1.027 - 1.150)	1.201 (1.143 - 1.268)	1.190 (1.132 - 1.255)	1.096 (1.034 - 1.163)	1.099 (1.043 - 1.162)	1.094 (1.031 - 1.162)
R^2_D (%)	22.0 (20.1 - 24)	25.6 (23.8 - 27.7)	25.3 (23.4 - 27.3)	22.3 (20.3 - 24.4)	22.4 (20.6 - 24.4)	22.2 (20.2 - 24.4)
Scaled Brier Score (%)	0.448	0.564	0.550	0.485	0.474	0.505
Without sex and age						
C index	0.588 (0.577 - 0.603)	0.622 (0.611 - 0.634)	0.620 (0.608 - 0.633)	0.587 (0.577 - 0.601)	0.592 (0.581 - 0.606)	0.589 (0.578 - 0.603)
Somers' D_{xy}	0.176 (0.154 - 0.206)	0.245 (0.223 - 0.268)	0.240 (0.216 - 0.267)	0.174 (0.153 - 0.203)	0.183 (0.162 - 0.211)	0.179 (0.157 - 0.206)
D statistic	0.496 (0.437 - 0.570)	0.709 (0.647 - 0.773)	0.690 (0.629 - 0.756)	0.509 (0.451 - 0.581)	0.530 (0.472 - 0.601)	0.513 (0.445 - 0.581)
R^2_D (%)	5.5 (4.4 - 7.2)	10.7 (9.1 - 12.5)	10.2 (8.6 - 12)	5.8 (4.6 - 7.5)	6.3 (5.1 - 7.9)	5.9 (4.5 - 7.5)
Scaled Brier Score (%)	0.05	0.16	0.15	0.08	0.08	0.10

5.6.2 External validation of polygenic risk score Cox regression models

In external validation of Cox PRS models in the Geographic Validation Cohort (Table 5.14), discrimination and explained variation are greater than in the Derivation Cohort, as was seen in logistic regression models. The top C-index is seen with the LDpred-grid-sp model, 0.725 (95% CI 0.696-0.752), explaining 28.5% (23.4-33.7%) of observed variation.

Model fit is generally good as measured by the calibration slope, though the LDpred-inf model in particular is slightly overfitted (calibration slope 1.123 (0.950-1.291)), though confidence intervals spanned 1 for all models.

Table 5.14: Performance of PRS in Cox models in Validation Cohorts

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig	Reference
Geographic Validation Cohort							
C index	0.715 (0.686 - 0.743)	0.724 (0.696 - 0.751)	0.725 (0.696 - 0.752)	0.713 (0.686 - 0.740)	0.707 (0.681 - 0.734)	0.701 (0.675 - 0.729)	0.673 (0.644 - 0.702)
Somers' D_{xy}	0.430 (0.372 - 0.485)	0.448 (0.391 - 0.501)	0.450 (0.393 - 0.504)	0.426 (0.372 - 0.480)	0.415 (0.361 - 0.468)	0.402 (0.350 - 0.458)	0.345 (0.288 - 0.404)
D statistic	1.243 (1.075 - 1.406)	1.285 (1.124 - 1.448)	1.293 (1.130 - 1.460)	1.184 (1.029 - 1.346)	1.182 (1.023 - 1.348)	1.145 (0.992 - 1.319)	0.945 (0.790 - 1.113)
R^2_D (%)	26.9 (21.6 - 32.1)	28.3 (23.2 - 33.3)	28.5 (23.4 - 33.7)	25.1 (20.2 - 30.2)	25.0 (20.0 - 30.3)	23.8 (19.0 - 29.4)	17.6 (13.0 - 22.8)
Scaled Brier Score (%)	0.75	0.76	0.78	0.63	0.61	0.59	0.37
Calibration Slope	1.123 (0.950 - 1.291)	1.058 (0.911 - 1.204)	1.073 (0.925 - 1.220)	1.070 (0.919 - 1.234)	1.054 (0.897 - 1.223)	1.023 (0.869 - 1.204)	0.947 (0.774 - 1.142)
Minority Ethnic Validation Cohort							
C index	0.647 (0.593 - 0.700)	0.666 (0.610 - 0.720)	0.664 (0.609 - 0.718)	0.650 (0.596 - 0.705)	0.658 (0.606 - 0.710)	0.659 (0.605 - 0.715)	0.647 (0.595 - 0.702)
Somers' D_{xy}	0.293 (0.185 - 0.399)	0.331 (0.221 - 0.440)	0.329 (0.219 - 0.437)	0.300 (0.192 - 0.410)	0.316 (0.212 - 0.420)	0.319 (0.210 - 0.430)	0.293 (0.189 - 0.403)
D statistic	0.931 (0.650 - 1.273)	1.033 (0.736 - 1.374)	1.030 (0.734 - 1.363)	0.940 (0.640 - 1.281)	0.981 (0.682 - 1.320)	0.995 (0.693 - 1.335)	0.889 (0.610 - 1.229)
R^2_D (%)	17.2 (9.2 - 27.9)	20.3 (11.5 - 31.1)	20.2 (11.4 - 30.7)	17.4 (8.9 - 28.1)	18.7 (10.0 - 29.4)	19.1 (10.3 - 29.9)	15.9 (8.2 - 26.5)
Scaled Brier Score (%)	0.16	0.26	0.26	0.16	0.21	0.19	0.14
Calibration Slope	0.262 (0.161 - 0.397)	0.314 (0.205 - 0.452)	0.318 (0.207 - 0.455)	0.252 (0.154 - 0.384)	0.297 (0.188 - 0.442)	0.251 (0.151 - 0.389)	0.232 (0.136 - 0.366)

Kaplan Meier cumulative incidence curves of the Geographic Validation Cohort plotted alongside the Test Cohort indicate a degree of miscalibration, particularly in the highest risk groups, where the curves deviate significantly between the two datasets (Figure 5.14).

In calibration plots of PRS Cox models (Figure 5.15A) over 5-8 years of follow-up, genome-wide models under-predict risk in the highest risk groups, echoing the findings of the Kaplan Meier curves. This miscalibration is well corrected with recalibration-in-the-large for all but the LDpred2-inf model (Figure 5.15B).

In the Minority Ethnic Validation Cohort performance overall is poorer across all metrics (Table 5.14), with a top C-index of 0.666 (95% CI 0.610-0.720) and explained variation of 20.3% (11.5-31.1%) seen with the LDpred2-grid model. Models were highly underfitted with calibration slopes of between 0.251 and 0.318. This miscalibration is seen in the Kaplan Meier cumulative incidence curves (Figure 5.16) which demonstrate lack of overlap between the Test and Validation cohorts across both high and low risk groups.

Calibration plots show marked miscalibration across all risk groups (Figure 5.17A), which was significantly improved with recalibration-in-the-large. This is particularly improved for the GWAS-sig model, and also much better for the C+T, LDpred2-grid and LDpred2-grid-sp models, particularly over a longer follow-up period (Figure 5.17B).

5.6.3 Subgroup analysis of polygenic risk score Cox regression models

Subgroup analysis of Cox PRS model performance by sex in the Geographic Validation Cohort shows that models have higher discrimination and explain a greater proportion of variation in males compared to females (Table 5.8). For example, for the LDpred2-grid model the C-index in men and women are 0.724 (0.691-0.761) and 0.711 (0.671-0.746) respectively, with an additional 1.5% of variation explained in men (R^2 27.9% (21.5-34.5%) compared to 26.4% (18.8-33.7%)). Calibration slopes are closer to 1 in women indicating better fit of the

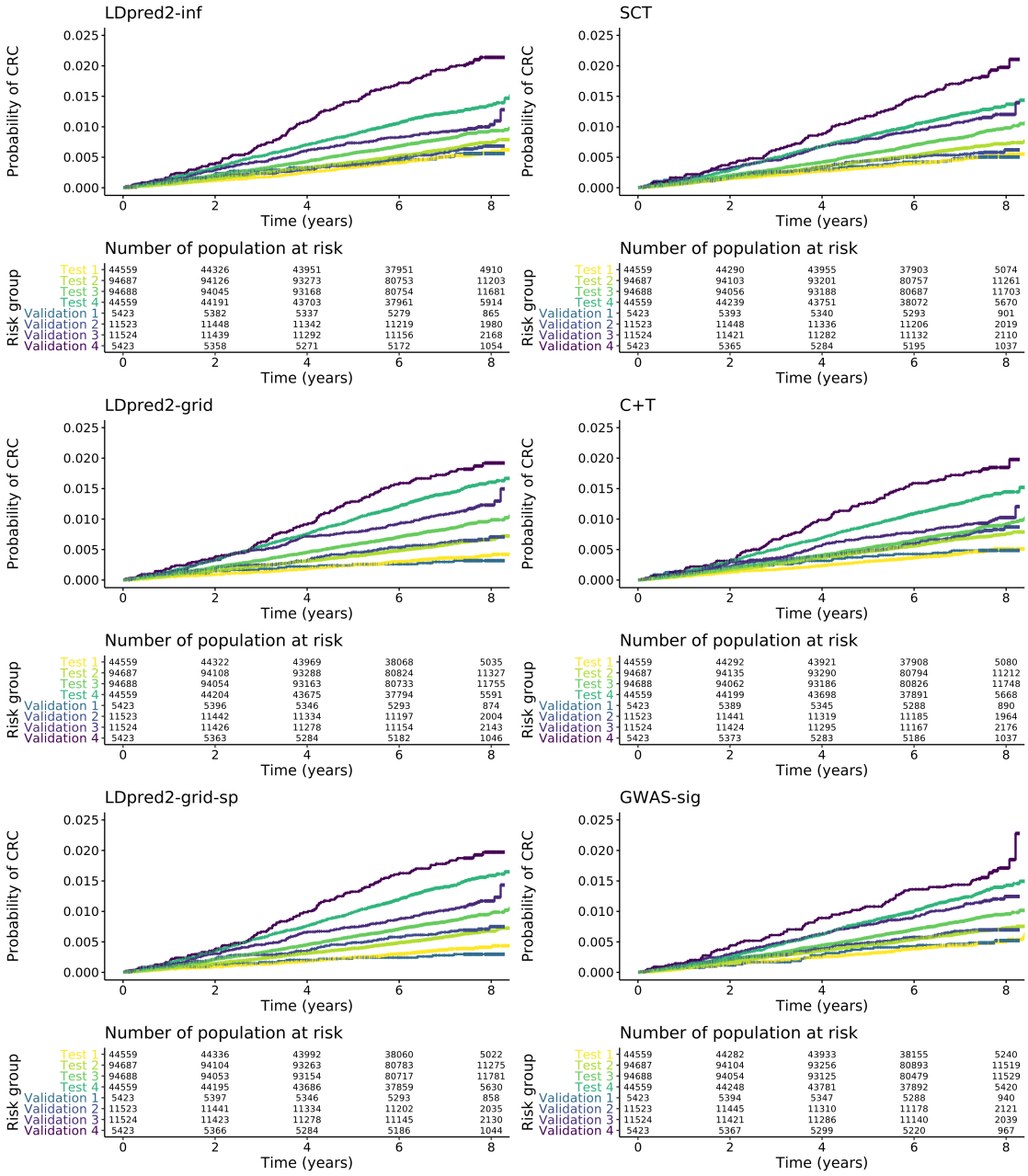

Figure 5.14: Kaplan-Meier curves of Cox PRS models in Test and Geographic Validation Cohorts across four risk groups (where Group 1 is lowest risk)

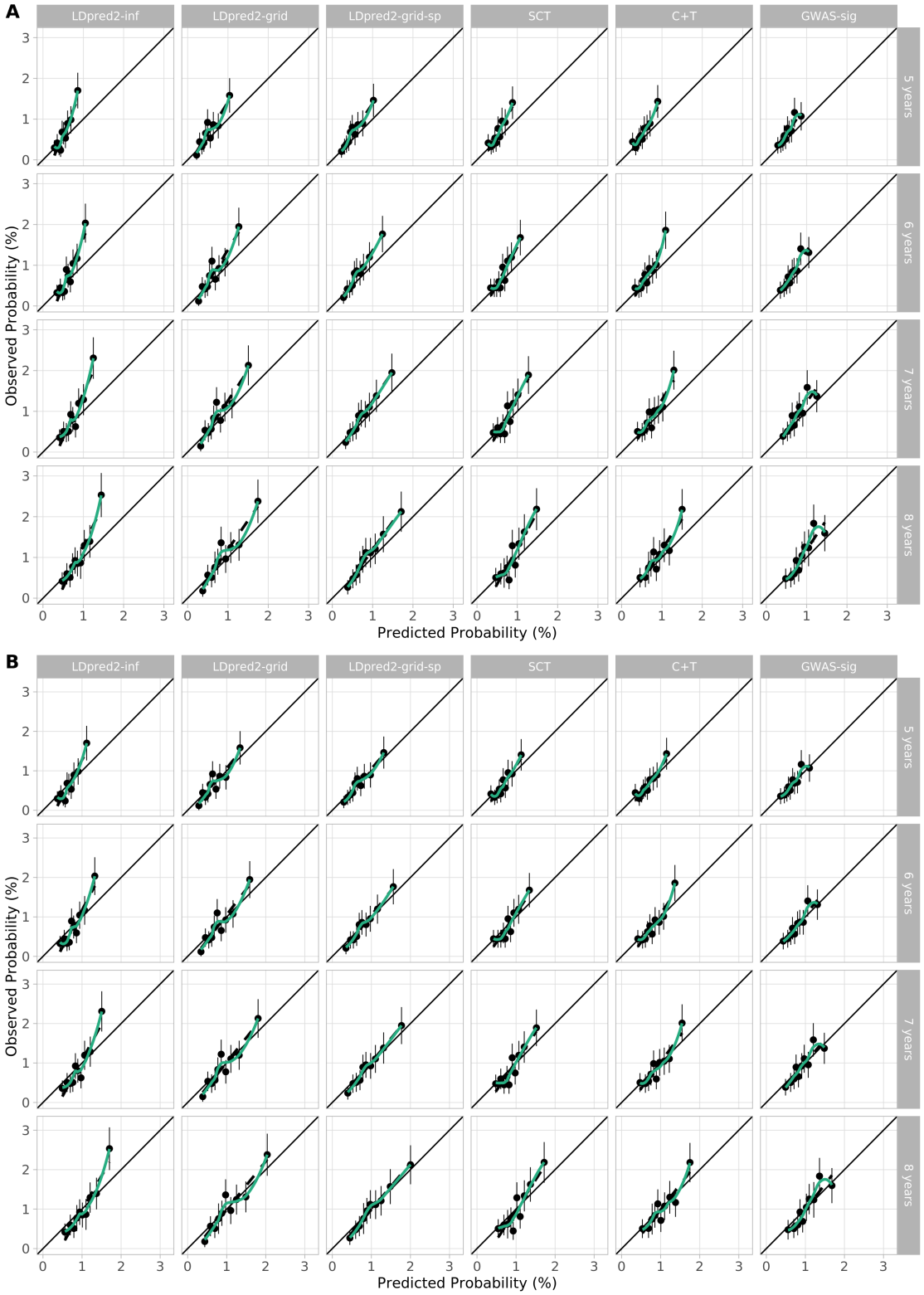

Figure 5.15: Calibration plots of Cox PRS models in the Geographic Validation Cohort before (A) and after (B) recalibration

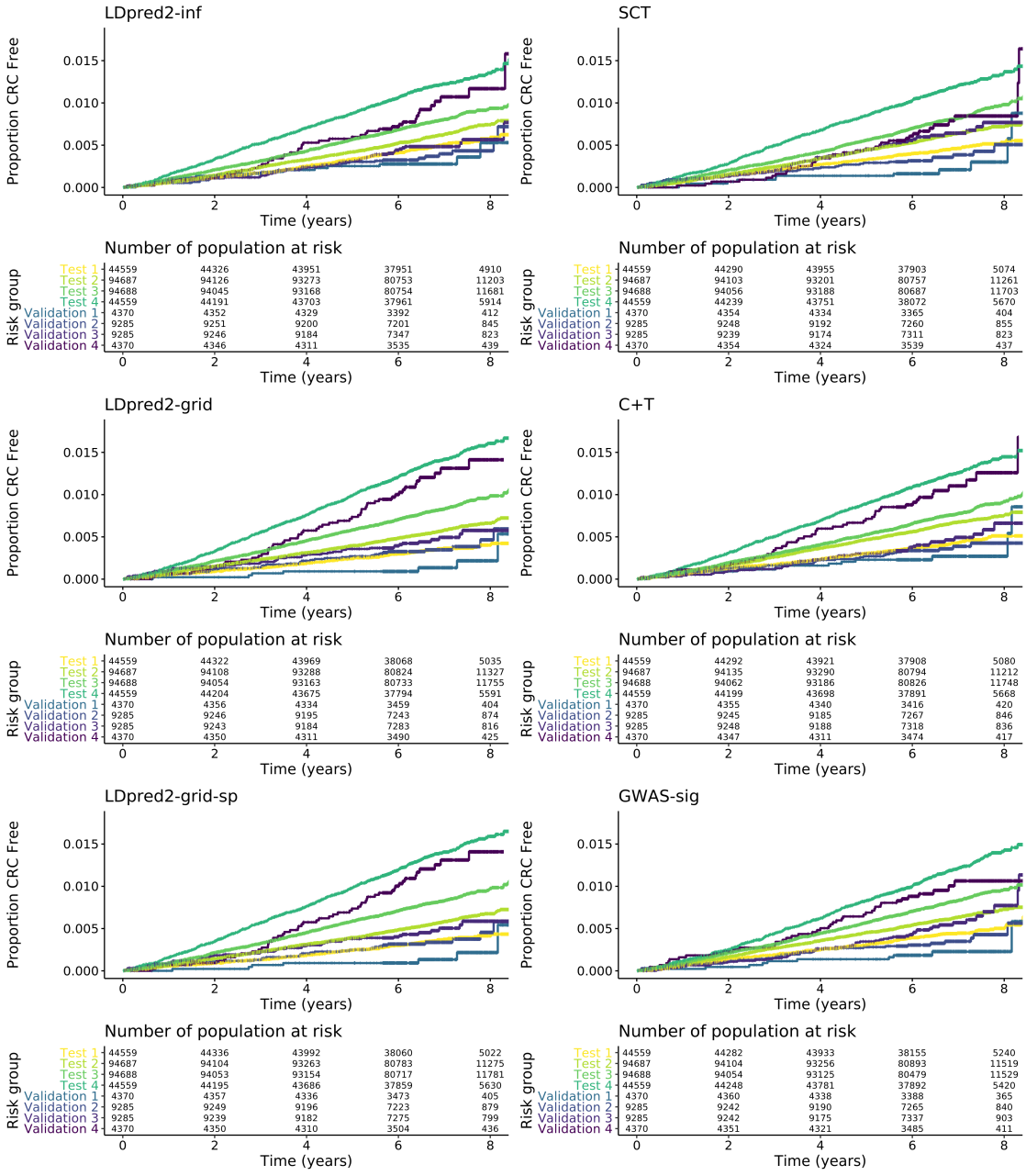

Figure 5.16: Kaplan-Meier curves of Cox PRS models in Test and Geographic Validation Cohorts across four risk groups (where Group 1 is lowest risk)

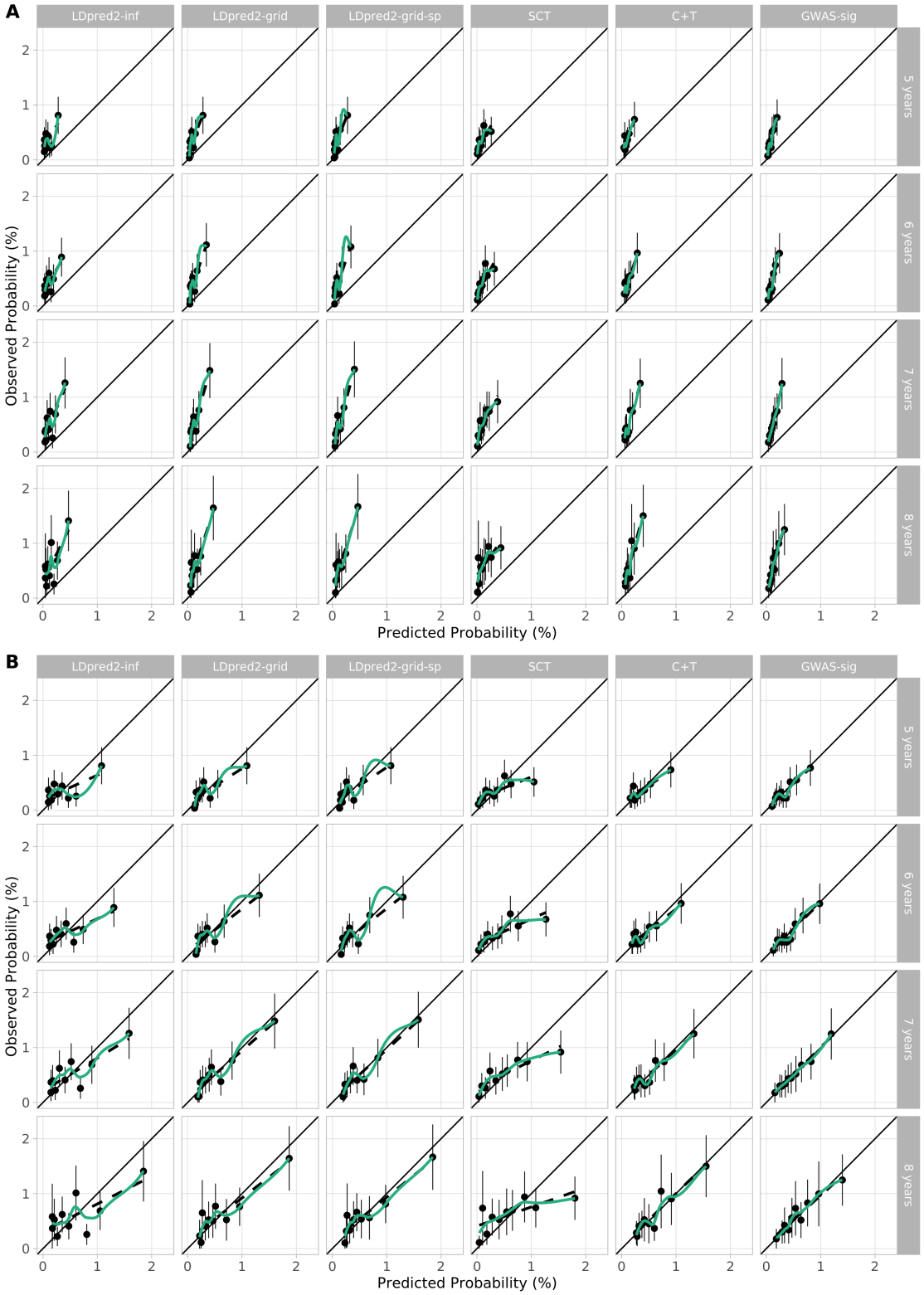

Figure 5.17: Calibration plots of Cox PRS models in the Minority Ethnic Validation Cohort before (A) and after (B) recalibration

models, however calibration plots show that risk is under-predicted to a greater extent in women in the top risk groups compared to men (Figure 5.18).

Calibration by age is near-identical for all models (Figure 5.19). Though well calibrated in younger age bands, models under-predict risk to a greater extent in the oldest age group. The jump in observed risk seen in analysis of logistic regression models is again evident, here in 55-64 year olds, resulting in apparent miscalibration in this age group. There are too few incident cases (n=55) in individuals with positive family history of CRC in this cohort to examine performance separately.

Table 5.15: Performance of PRS in Cox models by sex in the Geographic Validation Cohort

Index	LDpred2-inf	LDpred2-grid	LDpred2-grid-sp	SCT	C+T	GWAS-sig
Males						
C index	0.709 (0.675 - 0.747)	0.724 (0.691 - 0.761)	0.723 (0.691 - 0.760)	0.711 (0.677 - 0.745)	0.704 (0.668 - 0.740)	0.707 (0.675 - 0.740)
Somers' D_{xy}	0.419 (0.349 - 0.493)	0.448 (0.382 - 0.522)	0.446 (0.382 - 0.520)	0.422 (0.354 - 0.489)	0.408 (0.337 - 0.481)	0.414 (0.350 - 0.481)
D statistic	1.197 (0.989 - 1.430)	1.272 (1.072 - 1.486)	1.271 (1.063 - 1.499)	1.149 (0.963 - 1.359)	1.156 (0.938 - 1.383)	1.185 (0.991 - 1.394)
R^2_D (%)	25.5 (18.9 - 32.8)	27.9 (21.5 - 34.5)	27.8 (21.2 - 34.9)	24.0 (18.1 - 30.6)	24.1 (17.3 - 31.3)	25.1 (19.0 - 31.7)
Scaled Brier Score (%)	0.82	0.84	0.85	0.67	0.67	0.72
Calibration Slope	1.172 (0.954 - 1.431)	1.120 (0.942 - 1.327)	1.128 (0.944 - 1.350)	1.139 (0.947 - 1.370)	1.117 (0.896 - 1.365)	1.157 (0.950 - 1.390)
Females						
C index	0.707 (0.670 - 0.745)	0.711 (0.671 - 0.746)	0.713 (0.673 - 0.749)	0.700 (0.657 - 0.738)	0.696 (0.655 - 0.731)	0.680 (0.638 - 0.720)
Somers' D_{xy}	0.414 (0.340 - 0.490)	0.421 (0.342 - 0.492)	0.427 (0.345 - 0.498)	0.399 (0.313 - 0.476)	0.393 (0.309 - 0.461)	0.360 (0.276 - 0.439)
D statistic	1.244 (0.994 - 1.492)	1.227 (0.985 - 1.459)	1.250 (1.005 - 1.485)	1.142 (0.882 - 1.398)	1.133 (0.887 - 1.377)	1.004 (0.769 - 1.245)
R^2_D (%)	27.0 (19.1 - 34.7)	26.4 (18.8 - 33.7)	27.2 (19.4 - 34.5)	23.8 (15.7 - 31.8)	23.5 (15.8 - 31.1)	19.4 (12.4 - 27.0)
Scaled Brier Score (%)	0.55	0.52	0.56	0.44	0.41	0.30
Calibration Slope	1.171 (0.912 - 1.460)	1.053 (0.827 - 1.278)	1.080 (0.847 - 1.304)	1.076 (0.822 - 1.354)	1.058 (0.802 - 1.312)	0.944 (0.708 - 1.203)

5.7 Discussion

In this chapter I evaluate a range of genome-wide approaches to developing PRS and compared these with a "standard" GWAS-significant model. I demonstrate that models derived using an LDpred2 grid-based approach perform best in both logistic and Cox regression models, providing greater discrimination, and explaining a greater proportion of variation in risk. These models also tend to be the better-calibrated of the genome-wide approaches. There was little difference in performance metrics for sparse and non-sparse LDpred2-grid models, despite the sparse model containing just over half the number of SNPs. I have demonstrated portability of these models through validation in a geographically external cohort and shown that miscalibration in a new population can be improved by recalculating the baseline risk (recalibration-in-the-large).

My first evaluation of PRS, modelling relative and absolute risk from the log-normal distribution of the PRS (following modelling approaches used by Pashayan et al. [449]) predicted an AUROC of 0.66 achieved with 97 SNPs. This is a marked improvement on the AUROC of 0.63 predicted by Frampton et al.'s 37 SNP PRS [157], but as demonstrated by my subsequent modelling in UKB and the work of others [215, 219], represents a significant over-estimate of PRS performance achieved in "real life" datasets. Including all potential common variants in the model increased the AUROC to 0.78, and significantly increased relative risk discrimination. As noted by Wray et al. [232], there is an upper limit to the possible predictive performance of a polygenic predictor. This is limited by the genetic architecture and heritability (h^2) of the condition, which is the upper limit of the phenotypic variation in complex traits arising from genetics; the remaining variation is due to non-genetic factors [232, 233]. The C-statistic for a genetic risk prediction test reports the accuracy of prediction of disease status, which is influenced by many additional non-genetic risk factors, rather than the accuracy of the genetic risk test for predicting true genetic risk [233]. Thus a PRS will never be able to achieve "perfect" prediction.

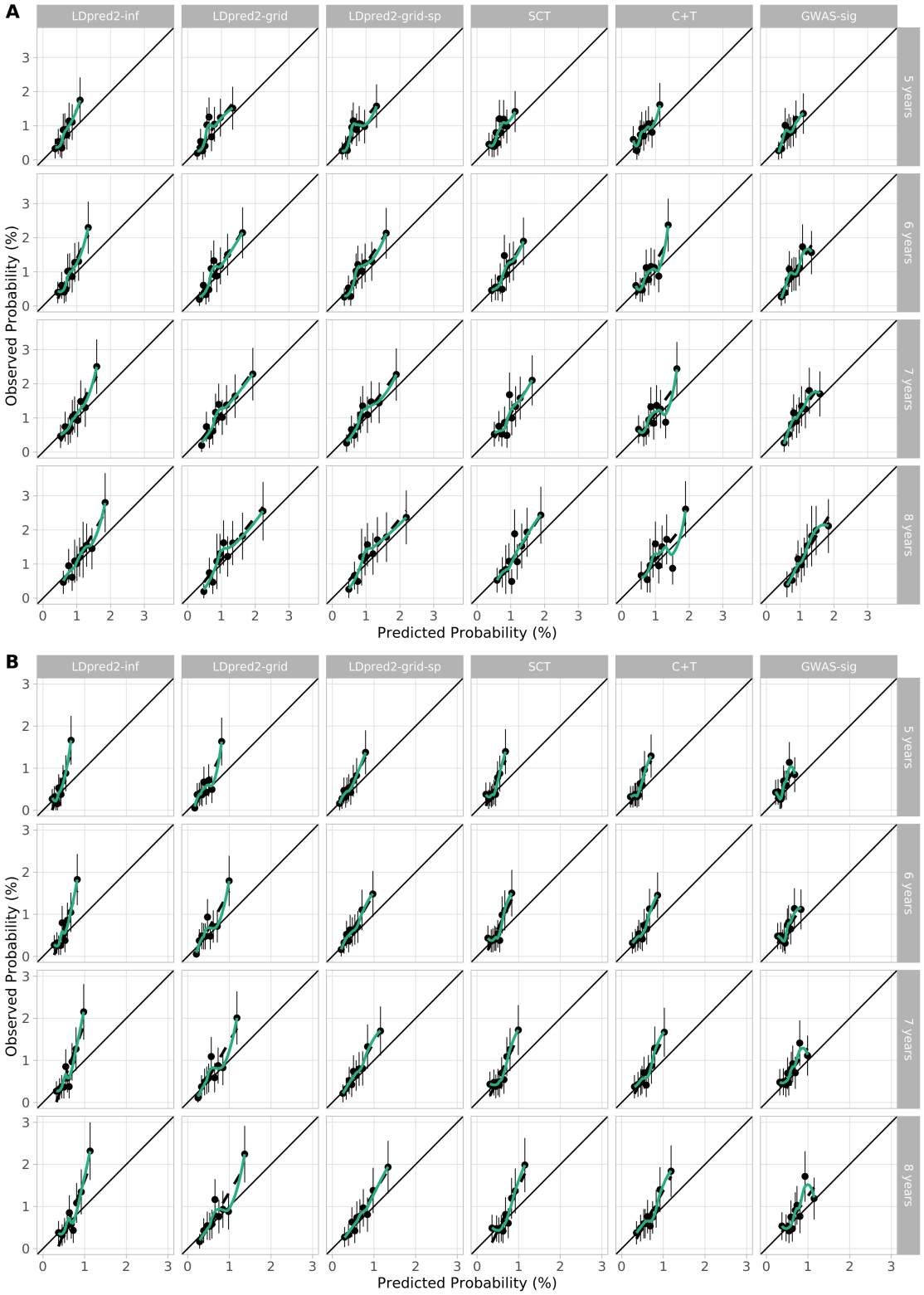

Figure 5.18: Calibration plots of Cox PRS models in the Geographic Validation Cohort in males (A) and females (B)

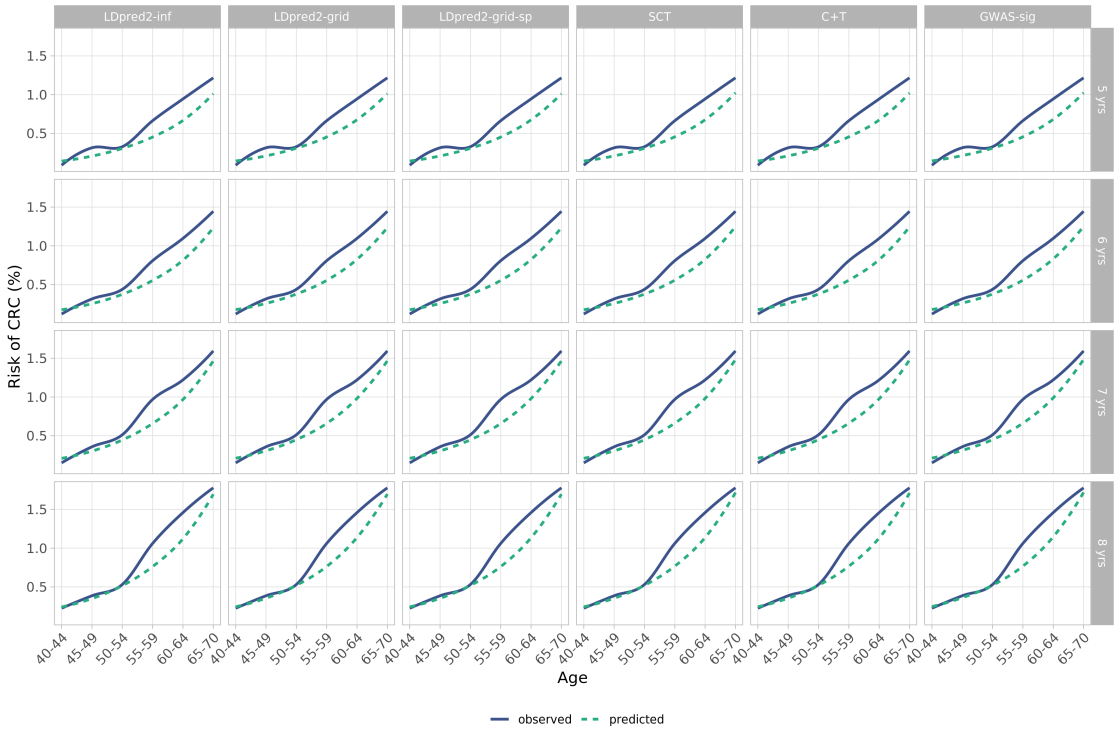

Figure 5.19: Predicted and observed risk of CRC in 5 year age bands over 5-8 years of follow-up in the Geographic Validation Cohort

My modelling of absolute risk variation using the 97-SNP PRS demonstrated the significant variation in risk by PRS, sex, and age. I show that compared to the UK Bowel Cancer Screening Programme screening approach currently adopted of screening men and women aged 50-74, a risk-based approach would result in a significant reduction in screening burden, with only a small reduction in the number of cancers detected. Though quite a crude evaluation, this demonstrates the potential for risk-modified screening to improve the efficiency of a bowel screening programme.

As noted in Chapter 1, most previously published PRS for CRC have included GWAS-significant SNPs alone, with two previous evaluations of genome-wide approaches using LDPred, machine learning, C+T, and Lassosum [219, 228]. Time constrains meant that whilst there now are many statistical approaches available for PRS, I selected a more limited number of methodologies for my own evaluation of genome-wide PRS. At the time I began the work presented here, no studies had been published evaluating genome-wide methodologies. I resolved to evaluate C+T-based approaches, which was the most commonly used genome-wide methodology at the

time. I initially attempted to use PRSice software, which extends manual C+T approaches to evaluate many thousands of thresholds. This programme has been used in UKB for other phenotypes using smaller case-control studies [464], but failed to run with the size of dataset and computational power available to me. On discussion with other researchers and the authors it became clear that this limitation was not unique to me; I subsequently used C+T and SCT implemented in *bigsnpr*. LDpred2 was published during this time, and had been shown to outperform multiple other methodologies (including another commonly used package, lassosum), across a range of phenotypes [211]. I therefore evaluated this in addition. An extension of this project would be to evaluate other available methods.

With regard to the relative performance of my models, the optimal performance of LDpred2 was in line with existing studies [228]. Initial work on C+T methodologies had suggested that SCT out-performed C+T methods except where highly underpowered. Here SCT and C+T methods performed equally well, as was the case for many phenotypes in comparisons made by Privé, Arbel, and Vilhjálmsson [211]. The absolute differences in performance between models presented here, though statistically significant, are small. On external validation the improvement in C-statistic in the LDpred2-grid-sp model over GWAS-sig model in logistic regression analysis was just 0.029, with an improvement in R^2 of 2.2%. Calibration was worst for the LDpred2-inf model, which might be anticipated as this approach is likely to be most prone to over-fitting in the Training Cohort. The GWAS-sig model was less prone than genome-wide models to over-predict risk in the upper risk groups, a reflection of the fact that these SNPs were selected from meta-analysis of external datasets, rather than using UKB data.

The base GWAS dataset chosen is of prime importance in PRS studies. The GWAS meta-analysis by Huyghe et al. [222] has supplied the base data for a number of studies of PRS in CRC, many of which have been undertaken in UKB. As noted previously, including the same individuals in the base GWAS (used to identify SNPs and derive weights) and the derivation dataset used to train PRS models, results in over-estimation of the association between the PRS and outcome [282], and so

the model is over-fitted to the training dataset, and performance estimates are inflated. The extent of inflation is proportional to the fraction of the training dataset overlapping the base (GWAS) dataset [232]. Examining the methods of previously published CRC GWAS, and UKB-based studies PRS studies, almost all of the UKB training cases will have been present in the base meta-analysis. This overfitting will also contribute to poor relative performance of PRS when applied elsewhere.

The optimism conferred by this approach is evident in several published studies. Using the Huyghe et al. [222] base dataset and a number of other UKB-derived GWAS studies, Fritsche et al. [219] generated multiple PRS using clumping and thresholding and lassosum approaches, for 35 common phenotypes including CRC. They evaluated their PRS, adjusting for age, sex, array, and first 4 PCs, in both UKB and in the MGI cohort. There are clear differences in performance between the cohorts. Their top CRC model in UKB containing 87 SNPs (using a C+T approach) has an AUROC of 0.617 (95% CI 0.605-0.630). In the MGI cohort, which does not overlap with the base dataset, the maximal PRS performance was lower (AUROC 0.567 (0.540-0.594)). Similarly, in external validation of existing PRS in UKB, Saunders et al. [215] reported that the PRS from Huyghe et al. [222] was the top-performing, with an unadjusted AUROC of 0.64 in men and 0.62 in women. However, this was the only PRS to be derived from data which included UKB; other PRS, without overlap, resulted in AUROCs <0.6.

Ideally, samples present in the training dataset should be removed from the base meta-analysis to avoid this overlap, though this comes at a cost of reduced power to detect new variants [282]. As I had access to individual GWAS summary statistics I was able to take this approach, reducing this bias in this study. Notably, the UKB dataset contains a small number of individuals who are also present in, or have relatives in, the Scottish GWAS cohort (n = 185 in the UKB case-control dataset used for GWAS in Law et al. [115]). I was unable to identify these individuals, but there may still be a small remaining overlap between base and derivation dataset which could lead to slight over-fitting. The cost of this approach is that the reduction in the size of the base GWAS resulted in a smaller

GWAS-significant PRS (50 SNPs) compared to other recent studies which have used ~140 SNPs. For my GWAS-significant PRS, correction for 'winners curse' reduces potential bias in the PRS further.

A further strength of this work is the use of genotype dosages rather than allele counts, which incorporates uncertainty around imputation into the PRS. Whilst I followed a similar approach to previously published studies in fitting my adjusted PRS models [e.g. 441, 238], several amendments to the methodology might have improved model fit. Steyerberg [284] recommends setting outlying values of a continuous predictor to the outer bounds, as these outlying values can have an overweighted effect on the effect size, which may affect model fit. This approach might have improved model calibration, and I would have undertaken a sensitivity analysis evaluating this approach had I had more time.

Overall my study provides a relatively unbiased estimation of the performance of the PRS compared with previous studies. The use of a geographic validation cohort further improves the robustness of performance estimates, and demonstrates generalisability of the models [465]. Estimates of model performance tend to be optimistic in their original datasets. On external validation therefore, performance metrics are often lower. A model can perform less well in external validation but still be clinically useful, depending on context and clinical judgement [169, 466]. I observed an improvement in performance in the Geographic Validation Cohort in comparison to the Derivation Test Cohort. This is likely to be a result of the relative homogeneity of the Derivation Dataset (as a result of the QC measures used in my analysis), in comparison to the Validation Cohort. The Geographic Validation Cohort included proportionally more women than the Derivation Cohort, though they were well-matched in terms of age distribution. Colorectal cancer prevalence was also greater in the Geographic Validation Cohort (1.79% compared to 1.51%).

Calibration was reasonable in the Geographic Validation Cohort however all models under-predicted risk in the highest risk group, most probably due to demographic differences between the two cohorts. This is more evident in women, suggesting that the differences between the two cohorts may be greater for women.

Models could be recalibrated for male and female populations separately prior to implementation. Given this miscalibration, one would need to be cautious in giving risk estimates to individuals in the highest risk group on the basis of PRS. However, risk counselling is challenging even in speciality cancer genetics clinics, and giving ranges of risk is a standard approach.

Polygenic risk scores may not perform uniformly across all age groups. Previous studies in CRC found a 95-SNP PRS to be more strongly predictive of CRC risk in early onset than late onset CRC, with an almost linear decrease in log-OR of 0.06 with each decade of age [467, 468]. Similar reductions in predictive strength across age groups were seen in Thomas et al.'s study [228, 469], but not in PRS derived by Li et al. [216]. Other phenotypes also show a reduction in predictive strength of PRS with age [470]. Evaluating this interaction in my dataset, I found that the effect size of PRS did decrease with age, but the interaction was only significant for one LDpred2 PRS in logistic regression models. I ultimately chose not to include this in my evaluation of PRS models given the relative weakness of the interaction, particularly as the unnecessary inclusion of interaction terms can lead to overfitting and may not improve model performance [284]. Had time permitted, I would have undertaken a sensitivity analysis including this interaction.

In the Geographic Validation Cohort, I noted miscalibration in 55-59 year olds in logistic regression analysis, and in 55-59 and 60-64 year old age groups in Cox regression, due to a jump in observed risk. This step in observed risk could perhaps be due to increased diagnosis of asymptomatic CRC as participants enter to the bowel screening programme, which begins at 50 years of age in Scotland (and did so at the time of UKB recruitment). Notably I didn't see this step in observed risk in the integrated modelling cohort in my evaluation of combined models in Chapter 6.

As was seen in measures of calibration, PRS generally performed less well for women across all performance metrics. However, the confidence intervals around these risk estimates are wide and these results should be interpreted with caution. Hypothetically it is possible that, were there to be an interaction between sex and effect size of risk alleles, the demographics of the Base GWAS datasets (for

example, disproportionally large numbers of men) could result in miscalibration of these models in women.

Models were poorly calibrated in individuals with a first degree family history of CRC, systematically underpredicting risk (CITL ~ 0.6 for all models), seen across most risk groups in calibration plots (Figure 5.10). This is likely to be in part due to lack of characterisation of individuals with Mendelian syndromes, and inclusion of Mendelian genes in a genetic prediction model would almost certainly improve risk prediction accuracy overall. Whole exome data was not available in UKB at the time of my analysis, but this would be a natural extension to this work. In recent evaluations of exome sequencing data within UKB, 0.15-0.2% of the initial release of 50,000 exomes from UKB carried pathogenic or likely pathogenic variants for Lynch Syndrome [471, 472]. Modelling of Lynch mutation status (pathogenic or likely pathogenic variants in any Lynch gene) alongside a 95-SNP PRS in 48,812 individuals with exome sequencing data in UKB found odds ratios for CRC ranging between 8 to 117 in Lynch carriers (n=76) across percentiles of PRS, compared to 0.3-3.8 for non-carriers (compared to a non-carrier with median PRS). Similar findings were noted for BRCA1/2 carriers [472]. However, in an evaluation of PRS risk modulation in Lynch syndrome in much larger cohort of Lynch mutation carriers (n=826, with 504 CRC cases), a 107-SNP PRS was not associated with CRC risk [473]. Further work is needed to clarify the impact of PRS on risk in both Lynch syndrome, and other familial CRC syndromes. A potential use of PRS could be to inform risk estimates for those with hereditary cancers.

In this study, the base dataset included individuals of European ancestry, and I restricted the UKB Derivation Cohort to white-British individuals. This is an approach taken by other studies [211], but potentially limits portability to other settings. Were I to repeat this work I would use a broader genetic mix of northern-European individuals in PRS training, which might improve PRS portability further. My PRS perform well in individuals of European ancestry, and one might expect performance in a screening cohort of Northern European individuals to be reasonably similar to that seen in the Geographic Validation Cohort.

Poor performance of PRS in individuals from minority ethnicities is a well-recognised issue in PRS, which stems largely from biases introduced by Eurocentric GWAS studies when transferred to other populations. The genetic architecture of populations of differing ancestries differs significantly, with fundamental differences in LD and allele frequencies. European GWAS will miss-estimate effect sizes and miss causal loci from other populations. As a result, PRS based on European data do not reliably and accurately transfer to other populations [474]. I discuss this issue and implications in more detail in my Discussion chapter.

Demographic differences may also have contributed to poorer performance in the Minority Ethnic Validation Cohort. Whilst 43% of CRC cases in the Derivation Dataset were female, in the Minority Ethnic Validation Cohort this figure was 50%, despite a similar proportion of the overall cohort being female (~53%). This may suggest a higher CRC risk for minority ethnic women, reasons for which might include differences in screening uptake by sex in the two groups, or differences in environmental exposures leading to increased risk in women. The mean age of both cases and controls was also lower in the Minority Ethnic Validation Cohort, potentially indicating that older minority ethnic invitees were less likely to participate than older white participants.

The Minority Ethnic Validation Cohort is heterogeneous, including all participants not identifying as White. It would have been useful to evaluate model performance in a more granular way, for example evaluating cohorts from South Asia separately from those with Caribbean or Black African heritage. However given the small size of these populations and low case numbers within them such analyses would have been highly underpowered. Significant efforts are currently underway to improve representation in GWAS and PRS research, which I explore further in my Discussion Chapter.

The work of this chapter demonstrates that genome-wide PRS out-perform GWAS-significant PRS in CRC, though the extent of this improvement is relatively small. I selected LDpred2-grid-sp as the top-performing PRS, based on the C-statistic and R^2 in external validation, given its near equivalent performance with the

LDpred2-grid score, favouring sparsity for its advantages in terms of computational burden and biological plausibility. The differences in PRS performance by sex, and in individuals of divergent ancestry, highlight some of the ethical implications of clinical use of PRS which will need to be addressed prior to implementation. In the next chapter I evaluate the performance of combined models integrating the LDpred2-grid-sp and GWAS-sig PRS with the QCancer-10 risk score.

6

Integrated risk models for colorectal cancer risk

This chapter describes the development of integrated risk scores for colorectal cancer within UK Biobank, based on the polygenic risk scores derived in Chapter 5 and the QCancer-10 (Colorectal) non-genetic risk model, and the evaluation of the improvement in model performance and potential clinic impact of an integrated model compared to QCancer-10 (Colorectal) alone.

6.1 Background

As discussed in Chapter 1, multiple integrated models have previously been developed to predict colorectal cancer (CRC) risk. Nine were externally validated in UKB in a previous study by Saunders et al. [215] (Table 1.4), and following this publication several more have been published [238]. Whilst these models have incorporated varied non-genetic predictors, only GWAS-significant PRS have been included. The number of SNPs included has increased over time with the publication of larger sized GWAS studies, with the largest based on the meta-analysis by Huyghe et al. [222]. As several of these integrated models have been derived in UKB, overlapping with the data included in Huyghe et al.'s meta-analysis, PRS performance may be over-estimated, as discussed in the previous chapter.

Given the number of existing non-genetic models, and good discrimination on external validation of several, there is an argument for using an existing model as the basis for the non-genetic risk modelling in my own study. Of externally validated existing models, QCancer-10 (Colorectal) [182] was the best performing, with an AUROC of 0.70 (95% CI: 0.69-0.72) in men and 0.66 (95% CI: 0.64-0.68) in women in external validation within the UKB cohort [172]. There are additional benefits of using QCancer-10 (Colorectal) (hereafter, QCancer-10) in a risk score for use in bowel screening. It is derived from electronic health-records (EHR), and so could be quite easily integrated with bowel screening data, and potentially embedded at point of care in GP surgeries. In addition, QCancer-10 has already been recommended as a tool to help guide patient-informed screening decisions [475]. I therefore used QCancer-10 as my baseline epidemiological model.

The differences in demographics between the QCancer-10 derivation set and UKB have implications for likely model performance. QCancer-10 is derived from a population based cohort, and so in the healthier volunteer cohort of UKB could over-predict cancer risk due to the lower prevalence of incident cancers in the UKB. Conversely the combined model developed in UKB will be calibrated to this population, and so might be expected to under-predict risk when applied to the general population. Notably, the population of the bowel cancer screening programme (BCSP) is itself biased. In the first 5 years of the screening programme, among 60-64 year olds screening uptake was graded across area-level socio-economic status (SES), with lower participation in lower SES groups, by ethnicity, with lower participation in more ethnically diverse areas, and by sex, with lower uptake in men [445]. These participation biases echo those seen in participation in UKB, and thus UKB may be reasonably representative of the current bowel screening population, though not of the ideal bowel screening cohort. Although efforts are being made to reduce these disparities in uptake, and the introduction of faecal immunochemical testing (FIT) may lessen them further, they are unlikely to be mitigated completely.

The aim of this chapter was to evaluate whether the addition of PRS (developed in Chapter 5) to QCancer-10 would improve performance of the models in a clinically

meaningful way. I also aimed to evaluate whether incorporating the genome-wide PRS would have a greater impact than the GWAS-significant PRS.

6.1.1 Chapter Outline

The chapter starts with the external validation of QCancer-10 [182] in UKB. I then develop integrated models including either LDPred2-grid-sp PRS, identified in Chapter 5 as the top performing genome-wide PRS, or the GWAS-significant PRS, and compare the performance of these with QCancer-10 alone. I assess model performance using standard metrics as used in Chapter 5, and also evaluate the potential clinical impact of any incremental improvement in model performance.

6.2 Methods

All modelling in this chapter was undertaken using the Integrated Modelling Cohort, comprising 238,496 women (1,458 cases) and 196,091 men (1,895 cases) (see Section 2.8.8).

6.2.1 Validation of QCancer-10

Prior to developing the integrated models, I validated the performance of QCancer-10 in UKB for comparison. This provides comparable evidence of performance in a dataset of bowel screening age. As with all other modelling presented here, the model outcome (incident CRC) was as defined in Section 4.2.1, and as previously noted, I did not include anal cancer, although these were included in the original QCancer-10 model. Definition of QCancer-10 predictors, handling of outlying data and missingness, are described in Chapter 4.

Initially I evaluated performance of the original model, details of which can be found at `https://qcancer.org/15yr/colorectal/` [270], in the UKB dataset. The same performance metrics were used for all models in this chapter and are described in Section 4.2.1 and 6.2.3). I then recalibrated the model to the UKB dataset through recalibration-in-the-large (Section 2.8.11).

6.2.2 Integrated model specification

I derived separate integrated Cox regression risk models for men and women including two predictors for each: the QCancer-10 risk score, and a PRS risk score. I compared models including the top-performing genome-wide PRS, LDpred2-grid-sp, and the GWAS-significant PRS. Each of the scores were modelled as continuous predictors. I examined the distributions of each score in the Integrated Modelling Cohort, and visualised outliers using boxplots (Figure 6.1) and subsequently followed Steyerberg's recommendation to set 0.5% of outliers at each end of the distribution to the outer bounds, truncating the distribution [284, 476]. This improves model fitting, as values at the extremes can have an outsized impact on effect estimation.

I evaluated proportional hazards assumption as described in Section 5.2.6. I evaluated the forms of continuous variables using linear forms and multiple fractional polynomials (MFP, [477]) using the R package 'mfp'. I initially evaluated these in univariable analysis, evaluating MFPs with 2 and 4 degrees of freedom, plotting the forms for each predictor.

However, the forms of MFPs can be affected by other predictors and should therefore be evaluated together, and I subsequently evaluated potential MFP forms in combination for each model using the mfp programme, which uses an iterative analysis of model performance with differing MFPs to select the optimal form based on the best Akaike information criterion (AIC).

I also assessed interaction between the predictors by visual inspection of plots of marginal effects of the QCancer-10 risk score by PRS, and used the Wald χ^2 statistic to assess the prognostic strength and significance of interaction terms.

6.2.3 Measures of model performance

I compared performance of the recalibrated QCancer-10 model with integrated prediction models, calculating apparent and internally validated performance for the integrated models. I again used 500 bootstrap samples for confidence intervals and internal validation (see Section 5.2.6).

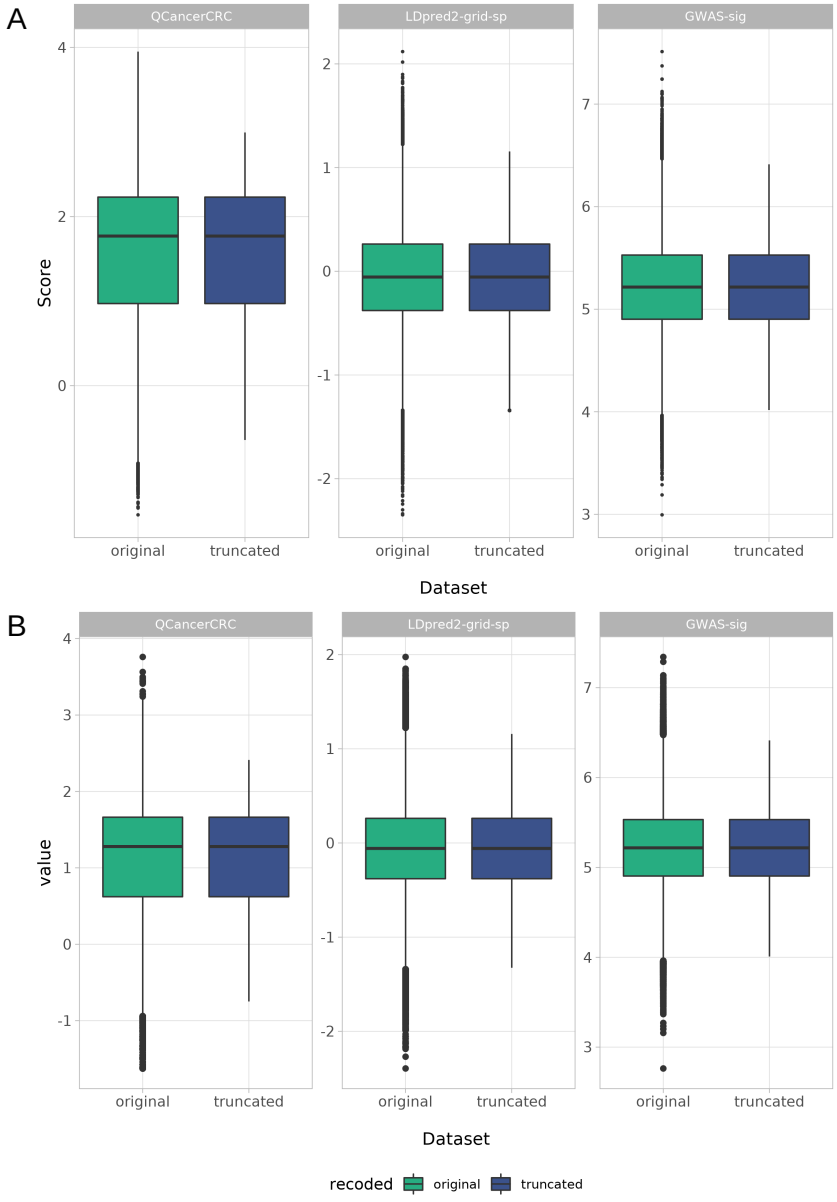

Figure 6.1: Boxplots of QCancer-10 (Colorectal) score and polygenic risk score distributions in males (A) and females (B) before and after removal of outiers in the Integrated Modelling Cohort

As for Cox PRS models, I reported discrimination using Harrell's C-index, and Royston and Sauerbrei's D statistic, and visually assessed discrimination through Kaplan-Meier cumulative incidence curves over 4 risk groups (see Section 2.8.8). I assessed explained variation using Royston and Sauerbrei's R^2_D, I assessed calibration using calibration plots of observed vs. predicted probability of CRC with loess-based smoothing methods, and expected to observed ratios (E/O) over follow-up times of 5-8 years.

I undertook a sensitivity analysis excluding cases diagnosed within the first two years of follow-up to assess possible reverse causality (i.e. the possibility that the outcome - colorectal cancer - has occurred before and caused the measured exposure variable).

6.2.4 Subgroup analysis

I compared calibration by analysis of observed and predicted outcomes in individuals from minority ethnicities, and with a family history of CRC in pre-specified subgroup analysis [169]. As I had observed some miscalibration by age in the PRS evaluation (underpredicting risk in the 55-59 and 65-70 year olds, see Section 5.7), I undertook a post-hoc analysis of model performance across 3 age groups - under 50 year olds, 50-59 year olds, and over 60 year olds.

6.2.5 Assessment of clinical performance

I calculated a number of measures of clinical performance at centile risk thresholds for absolute and relative risk, and present these across the top 25 risk groups:

- sensitivity (proportion of CRC cases correctly identified)
- specificity (true negatives/(true negatives + false positives))
- detection rate (number of cases identified in the tested population, as a percentage)
- false positive rate (false positives/(true negatives + false positives))

I calculated relative risk as risk in comparison to an individual of the same sex and age, with mean PRS by sex, white ethnicity, BMI of 25, mean Townsend Deprivation Score, and no CRC risk factors, representing a 'healthy' average reference individual.

6.2.6 Decision curve analysis

Decision curve analysis (DCA) was developed by Andrew Vickers and Elena Elkin, as a way of applying decision analysis methodologies to prediction models [478]. They noted that traditional measures of model performance, such as discrimination and calibration, were insufficient in evaluating whether risk models improved clinical decision making [479]. In decision analysis, the consequences of testing are incorporated into performance evaluation, to assess whether a test or model is worth using. The "best" model is the one that maximises the outcome. Decision curve analysis is recommended in the TRIPOD guidelines [168] as a measure of assessing prediction model performance. In DCA, the outcome of interest is measured as net benefit, and this is plotted across a range of clinically relevant risk thresholds. The possible theoretical range of net benefit is negative infinity to disease incidence [478], therefore for a less common disease, maximal net benefit will be lower.

In this analysis, I assume that the decision concerned is whether a person should have a screening colonoscopy, based on risk assessment, and compare the performance of the QCancer-10, QCancer-10+LDP and QCancer-10+GWS in guiding these decisions. In order to make a treatment decision, one must identify a risk threshold above which we would choose to undertake a colonoscopy, i.e. the level above which concern for potentially missing a diagnosis outweighs the risks or costs of the procedure (financial or otherwise). This ought to be the threshold level at which one feels neutrally towards the intervention [480]. For example, if we are willing to undertake colonoscopy on 100 individuals to detect one CRC, but no more than this, then the harm-to-benefit ratio is 1:99, equating to a risk threshold of 1% [480].

The net benefit (NB) is a balance of true positives and false positives i.e. colorectal cancers correctly detected minus unnecessary colonoscopies. Benefit (here detecting cancer) and harm (an unnecessary colonoscopy) are not on the same scale, and so

an "exchange rate" is used to address this balance [479]. The false positives are weighted by the odds at a given risk threshold (for example, if the risk threshold were 1% as above the weight would be 1/99, or 0.0101), and subtracted from the true positives to give the "net true positives". The net true positives are then divided by the sample size to obtain the NB, the unit of which is true positives. As per Vickers and Elkin [478],

$$NB = \frac{TruePositives}{N} - \frac{FalsePositives}{N}\left(\frac{p_t}{1-p_t}\right)$$

where N is the number of individuals in the integrated modelling cohort and p_t is the probability (or risk) threshold (ie, at p_t=1%, we are willing to perform colonoscopy for 100 individuals to detect one cancer).

For a survival model, the estimation of true and false positives necessitates conversion of survival data to binary outcomes at a given time-point. As described by Vickers et al. [481], if the individual's predicted probability from the model is $\geq p_t$ x=1, otherwise x=0; s(t) is defined as the Kaplan-Meier probability at the chosen time-point, and n is the size of the dataset. The number of true positives is then [1 - (s(t) | x = 1)] × P(x = 1) × n, and false positives is (s(t) | x = 1) × P(x = 1) × n [481]. This can then be used to calculate NB as above. I used the probability of the event at 8 years derived from each of the three models.

It is recommended that NB is calculated over a range of reasonable thresholds, as the acceptable threshold will vary between individuals and contexts [482], and here plots are extremely useful. The modelling strategies used are usually plotted alongside two default strategies - one in which everyone receives the intervention (i.e. colonoscopy for the whole population) and one in which nobody is treated (i.e. no screening, and therefore no benefit). One can also focus on the true negatives ascertained with the model, rather than the true positives, by evaluating the unnecessary interventions avoided. I therefore plotted NB and unnecessary interventions avoided over the range of risk thresholds for which the models showed benefit.

6. Intergrated-risk-models

To identify relevant thresholds to report NB at, I looked at the risk of cancer in which one might proceed with a screening procedure in other studies. Vickers notes that the threshold probability may also be considered as the acceptable number needed to test, or in the case of this study, number needed to screen (NNS) [483]. I examined the published literature for thresholds already deemed to be clinically acceptable. The NNS in previous randomised trials of colonoscopy-based screening is variable, at 1000 in two studies from the USA [484] and Spain [39], 202 in a multi-nationality European study [38], and 182 in a Dutch trial [485]. In a UK trial of screening sigmoidoscopy [55], the NNS to prevent cancer over 11 years was 191 (and to prevent one death, 489). These results correspond to probability thresholds of between 0.1% and 0.5%.

These numbers reflect immediate risk, rather than longer term risk, and additionally, a significant part of colonoscopy benefit is gained from carcinoma prevention through advanced adenoma removal. This is not reflected here, and could not readily be assessed in this study due to the absence of high resolution data on colorectal polyp diagnosis in UKB. The NNS for advanced neoplasia overall in these studies (i.e. advanced adenoma and CRC combined) is 9-49, equivalent to probability thresholds of 2-10%, which reflects the higher prevalence of advanced adenoma.

In choosing clinically relevant risk thresholds a holistic approach is needed, incorporating the potential risks and benefits for a patient, and potentially including wider considerations such as potential financial and logistical implications of implementing a test. Concern over the 'cost' of colonoscopy, both financially and in terms of opportunity costs in a capacity-limited healthcare system, will be greater in a population-based screening programme compared to a randomised controlled trial. In addition, for individual-level informed decision making, patient and clinician risk preferences will vary based on their level of concern about possible cancer, and around colonoscopy itself. It is recommended to evaluate net benefit metrics over a range of threshold, and I therefore report metrics across pre-specified threshold probabilities of 0.5%, 1%, 1.5% and 2%.

Whilst the strategy which obtains the highest NB across the relevant thresholds is the "best" strategy, when comparing prediction models which have different implications in terms of implementation (for example the requirement to undertake PRS, with blood draw, genotyping etc., compared to risk score calculation using routine health data), it is also useful to consider whether the increase in NB is worthwhile. An assessment of whether the increase in net-benefit would be worth the additional costs of using the model, the test trade-off, can be calculated as $1/\Delta$NB, which gives the minimum number of tests (i.e. PRS) required for one more true positive CRC diagnosis in the extended model [479]. If the financial and practical costs of these additional tests were deemed reasonable by policymakers then the added benefit of the integrated model would be considered worthwhile. I calculated the increase in NB achieved by adding the LDpred2 PRS to the QCancer-10 model, and the test trade-off at the pre-specified thresholds above.

6.3 Validation of QCancer-10

The performance of QCancer-10 is consistent with previous validation studies (Table 6.2) [172]. The model for women show poorer performance than that for men. Both of the models tend to slightly over-predict risk in the highest risk groups. Recalibration corrects this in men, but in women there was a persistent though lesser over-prediction of risk in the top risk group (Figure 6.2).

Subgroup analysis is discussed later in the chapter, in comparison with the integrated models (see Section 6.5).

6.4 Specification of integrated QCancer-10+PRS models

Plots of score distributions show a right-skew in the distribution of the QCancer-10 score in this dataset (Figure 6.3). Plots of $log(-log(Survival))$ against $log(Survival)$ show that the proportional hazard assumption holds (Figure 6.4). There is little difference in prediction when comparing MFP terms across each of the scores in

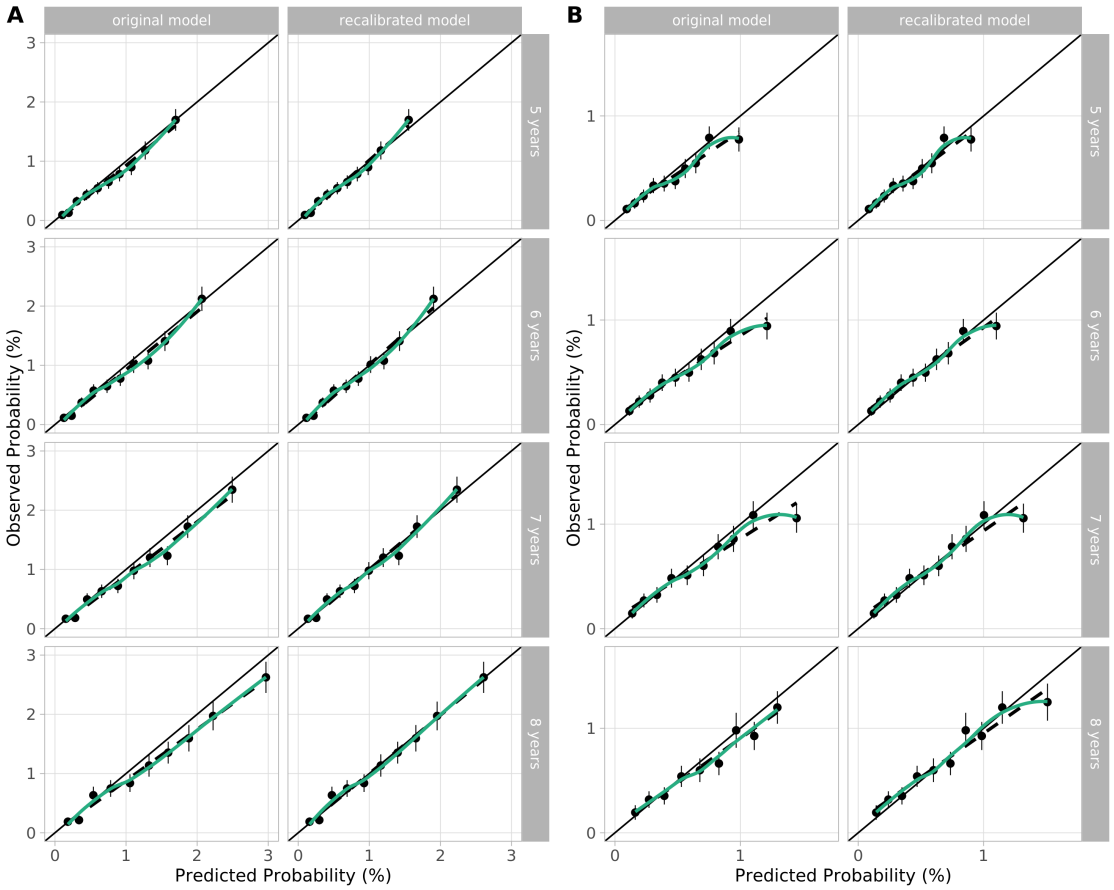

Figure 6.2: Calibration of QCancer-10 in males and females, before (A) and after (B) recalibration

men (Figure 6.5), and when modelled in combination no polynomial terms were selected for either model. However, in women, a polynomial term for LDpred2-grid-sp was selected.

Plots of marginal effects of QCancer-10 across PRS scores (Figure 6.6) indicate that there is a tendency for the QCancer-10 score to have a lesser effect at higher PRS. Wald χ^2 statistics for the terms indicate that there are no significant interactions (Table 6.1). Given the weakness of the interaction terms relative to the two scores, I elected not to include this in the model.

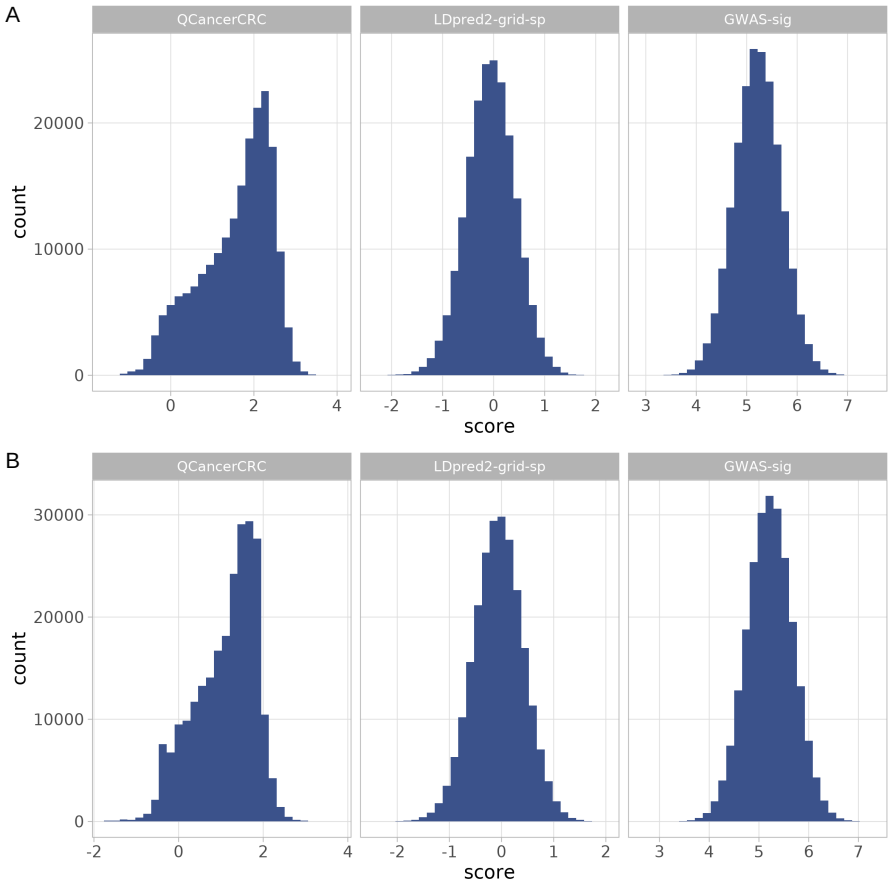

Figure 6.3: Histogram of QCancer-10 score distributions in males (A) and females (B) in the Integrated Modelling Cohort

Table 6.1: Wald Chi2 statistic for interactions between QCancer-10 score and PRS

	QCancer-10+LDP	QCancer-10+GWS
Males		
QCancer-10 LP	635.85 (<0.001)	652.88 (<0.001)
PRS	390.82 (<0.001)	264.36 (<0.001)
Interaction term	6.29 (0.012)	8.50 (0.004)
Females		
QCancer-10 LP	293.43 (<0.001)	302.90 (<0.001)
PRS	297.96 (<0.001)	135.83 (<0.001)
Interaction term	0.15 (0.699)	1.53 (0.216)

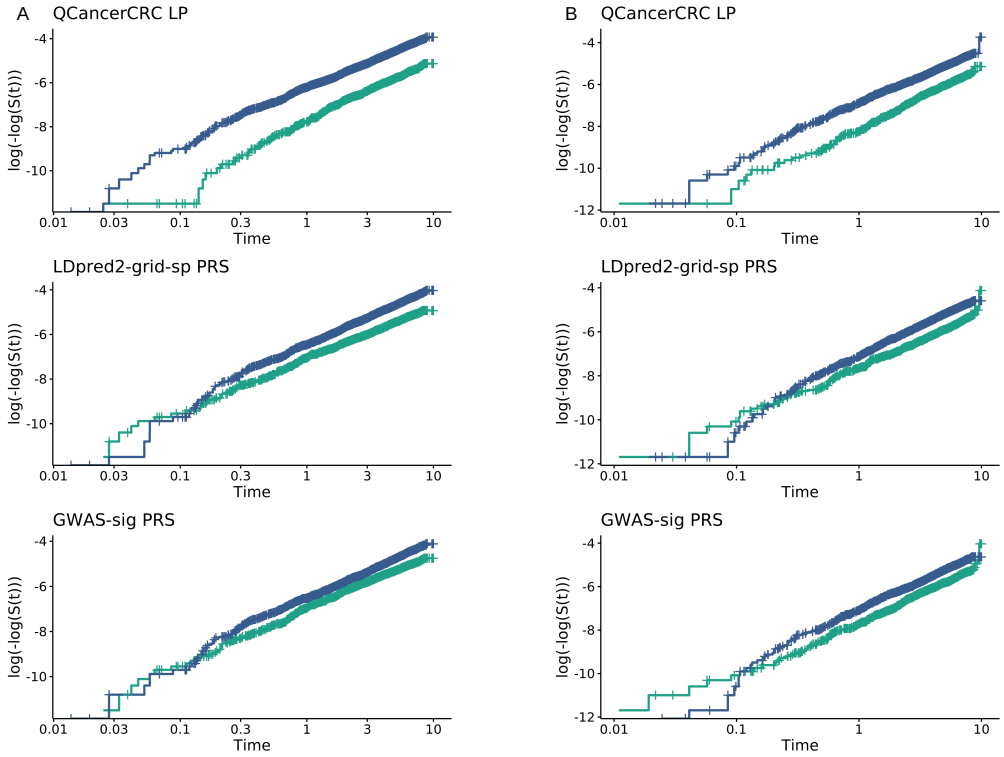

Figure 6.4: Plots of $log(-log(Survival))$ against $log(Survival)$ for PRS and QCancer-10 scores in males (A) and females (B)

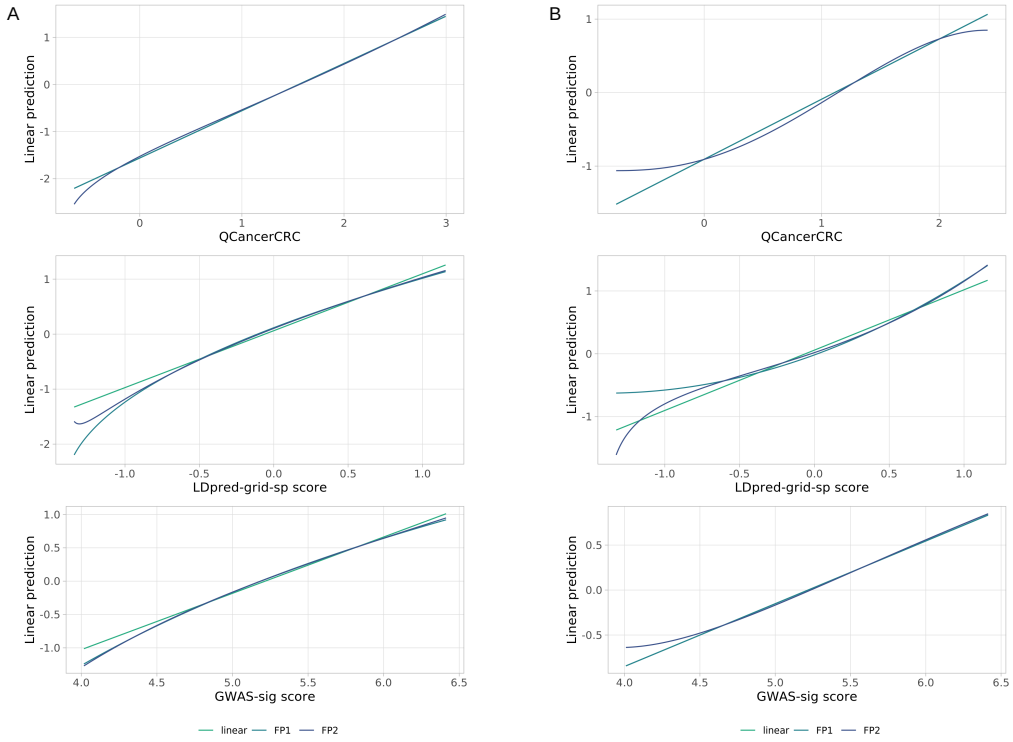

Figure 6.5: Plots of MFP forms in males (A) and females (B) in the Integrated Modelling Cohort

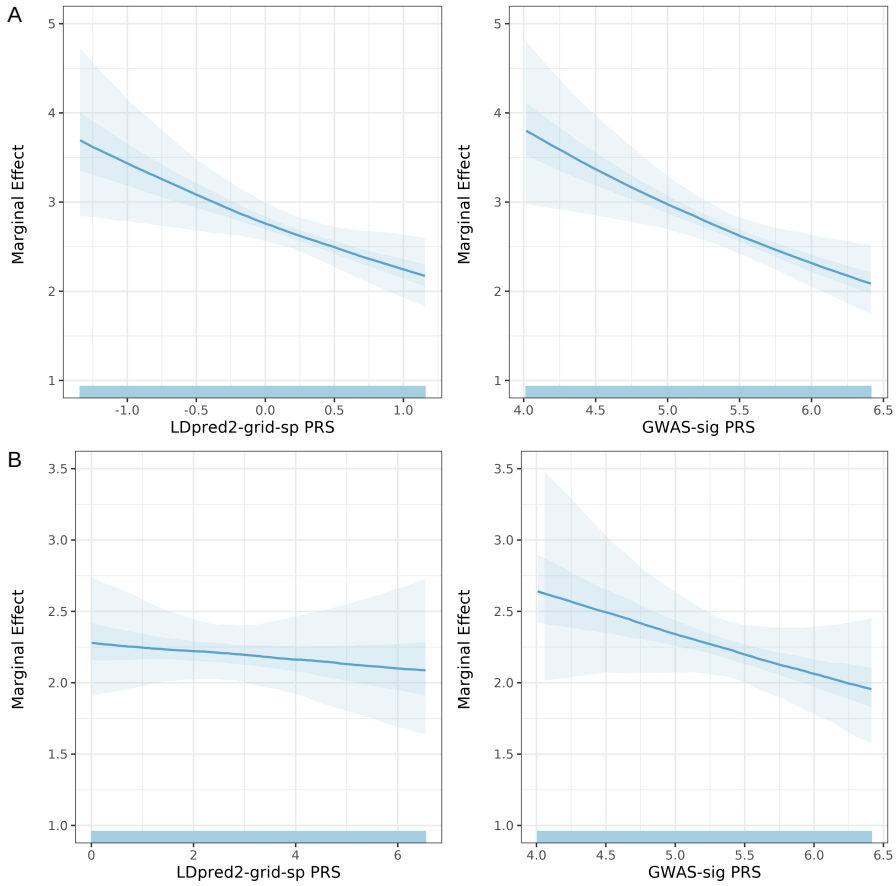

Figure 6.6: Plots of the marginal effects of the QCancer-10 score in interaction with PRS in men (A) and women (B)

6.5 Evaluation and comparison of model performance

Integrated models combining LDpred2-grid-sp PRS (QCancer-10+LDP), or the GWAS-significant PRS (QCancer-10+GWS), both perform better than the QCancer-10 model alone (Table 6.2), with best performance obtained with the addition of the genome-wide PRS. Internal validation showed that there was minimal optimism in performance estimates. As with the QCancer-10 model, integrated models for women perform less well than integrated models for men, though the incremental improvement in performance with the addition of PRS is greater for women.

Table 6.2: Apparent and internally validated performance of QCancer-10+LDP and QCancer-10+GWS, compared to external validation of QCancer-10.

	QCancer-10+LDP		QCancer-10+GWS		QCancer-10
	Apparent	Internally validated	Apparent	Internally validated	
Males					
QCancer-10 HR per SD	2.309 (2.164 - 2.464)	-	2.334 (2.187 - 2.490)	-	-
PRS HR per SD	1.592 (1.520 - 1.668)	-	1.453 (1.388 - 1.521)	-	-
C statistic	0.730 (0.720 - 0.741)	0.730	0.717 (0.707 - 0.727)	0.716	0.693 (0.682 - 0.704)
Somers' D_{xy}	0.460 (0.440 - 0.481)	0.460	0.433 (0.414 - 0.455)	0.433	0.847 (0.841 - 0.852)
D statistic	1.282 (1.224 - 1.341)	1.280	1.209 (1.156 - 1.267)	1.207	1.058 (0.987 - 1.121)
R^2_D (%)	28.2 (26.3 - 30.0)	28.100	25.9 (24.2 - 27.7)	25.800	21.1 (18.9 - 23.1)
Scaled Brier Score (%)	0.81	0.800	0.81	0.800	0.59
Calibration Slope	-	0.998	-	0.998	0.995 (0.914 - 1.063)
Females					
QCancer-10 HR per SD	*	-	1.764 (1.655 - 1.881)	-	-
PRS HR per SD	*	-	1.359 (1.290 - 1.431)	-	-
C statistic	0.686 (0.672 - 0.701)	0.685	0.668 (0.655 - 0.683)	0.668	0.645 (0.631 - 0.659)
Somers' D_{xy}	0.372 (0.345 - 0.402)	0.370	0.337 (0.310 - 0.366)	0.335	0.822 (0.816 - 0.830)
D statistic	1.056 (0.979 - 1.136)	1.055	0.925 (0.850 - 1.000)	0.924	0.769 (0.695 - 0.847)
R^2_D (%)	21.0 (18.6 - 23.5)	21.000	17.0 (14.7 - 19.3)	16.900	12.4 (10.3 - 14.6)
Scaled Brier Score (%)	0.34	0.340	0.29	0.280	0.2
Calibration Slope	-	0.996	-	0.996	0.805 (0.724 - 0.899)

Values are performance metrics withs 95% confidence intervals. HR per SD – adjusted hazard ratio per standard deviation. Pairwise comparisons of performance metrics were all significantly different P<0.001. *modelled using multiple fractional polynomial and therefore not presented.

Table 6.3: Apparent performance of QCancer-10+LDP, QCancer-10+GWS, and QCancer-10, excluding CRC cases diagnosed within 2 years of enrolment

	QCancer-10+LDP	QCancer-10+GWS	QCancer-10
Males			
C statistic	0.733 (0.721 - 0.745)	0.721 (0.708 - 0.731)	0.693 (0.682 - 0.704)
Somers' D_{xy}	0.867 (0.861 - 0.872)	0.860 (0.854 - 0.865)	0.847 (0.841 - 0.852)
D statistic	1.289 (1.219 - 1.358)	1.224 (1.155 - 1.290)	1.058 (0.987 - 1.121)
R^2_D (%)	28.4 (26.2 - 30.6)	26.3 (24.1 - 28.4)	21.1 (18.9 - 23.1)
Calibration Slope	1.013 (0.943 - 1.077)	1.019 (0.950 - 1.086)	0.995 (0.914 - 1.063)
Females			
C statistic	0.683 (0.666 - 0.698)	0.663 (0.646 - 0.677)	0.645 (0.631 - 0.659)
Somers' D_{xy}	0.841 (0.833 - 0.849)	0.831 (0.823 - 0.838)	0.822 (0.816 - 0.830)
D statistic	1.036 (0.946 - 1.124)	0.900 (0.805 - 0.985)	0.769 (0.695 - 0.847)
R^2_D (%)	20.4 (17.6 - 23.2)	16.2 (13.4 - 18.8)	12.4 (10.3 - 14.6)
Calibration Slope	0.981 (0.892 - 1.068)	0.969 (0.855 - 1.070)	0.805 (0.724 - 0.899)

Kaplan Meier cumulative incidence curves (Figure 6.7), plotting incidence across four risk groups defined by integrated models compared to QCancer-10, also indicates better discrimination when PRS are included, with greater separation of risk groups. Again, the improvement was greatest with the genome-wide PRS.

A sensitivity analysis in which I excluded CRC cases occurring within two years of enrolment, suggested that reverse causality did not have a major impact (Table 6.3).

6.5.1 Subgroup analysis

Table 6.4 shows expected/observed ratios for each model in subgroup analysis of those with a first degree family history, and in minority ethnic and white participants separately. In individuals with a first-degree family history of CRC, all models appear to be well calibrated for men. However, models for women tend to over-predict risk, particularly in the highest risk groups (Figure 6.8).

In participants of minority ethnic background, all models underpredict risk but this is exacerbated by the addition of PRS, to a greater degree with the genome-wide model. The caveat to these results are the low number of incident cases in this group, and I did not plot calibration plots for minority ethnic participants for this reason. Calibration for white participants is excellent.

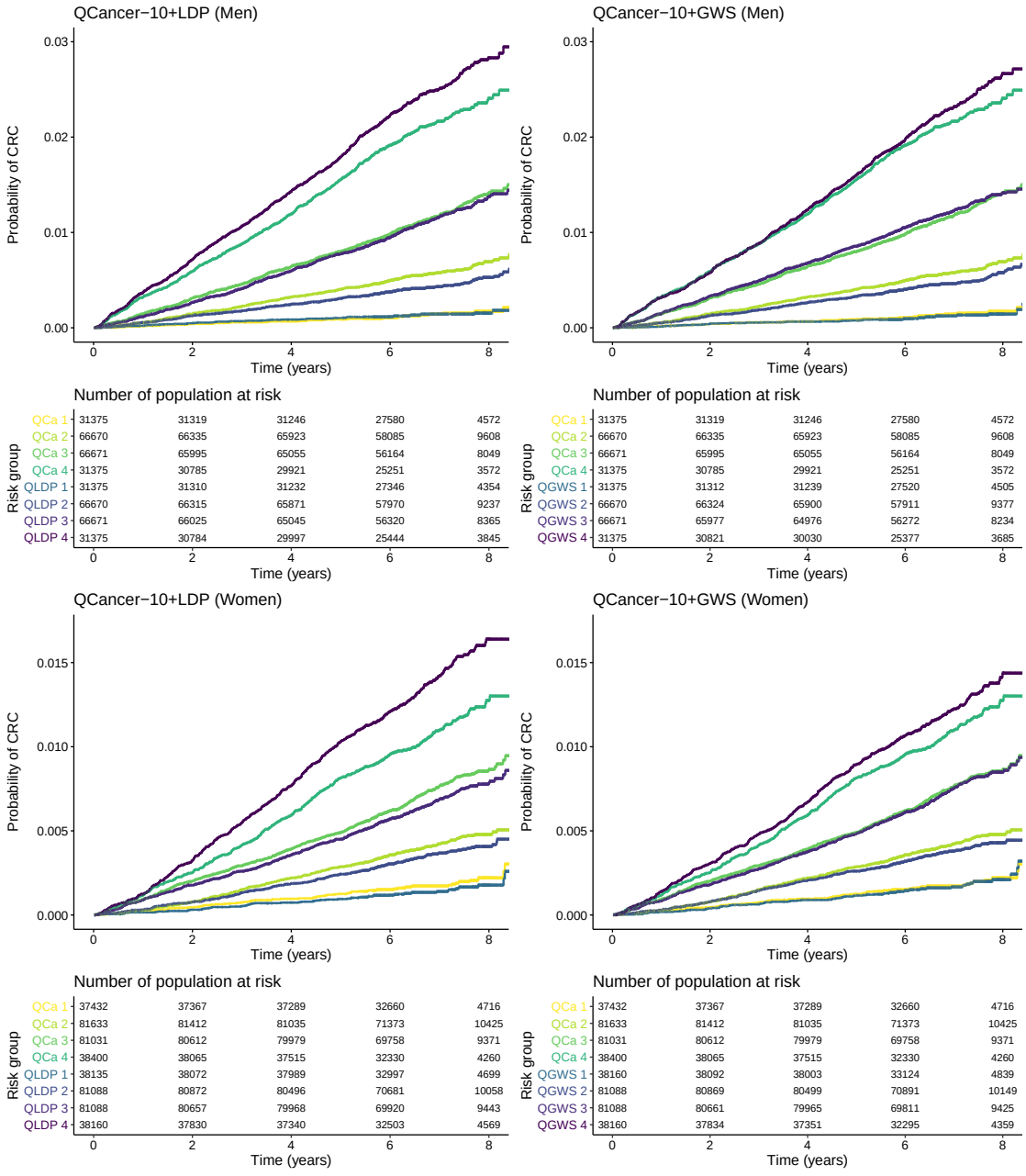

Figure 6.7: Kaplan Meier cumulative incidence curves for integrated models compared to QCancer-10

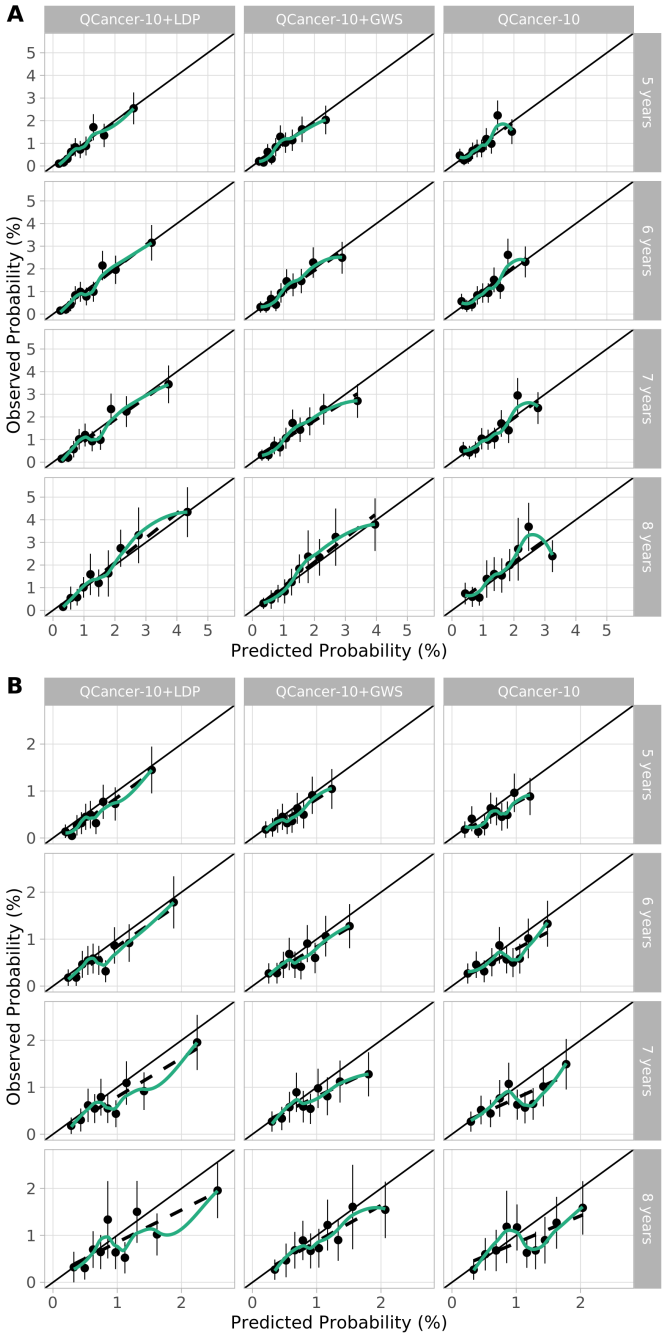

Figure 6.8: Calibration of prediction models in individuals with a first degree family history of CRC. Plots show predicted and observed CRC probability in males (A) and females (B) by tenths of predicted risk

Table 6.4: Expected/observed risk ratio for QCancer-10+LDP, QCancer-10+GWS, and QCancer-10 models in subgroup analysis

	Follow-up (years)	QCancer-10+LDP	QCancer-10+GWS	QCancer-10
Family history				
Male	5	1.06	1.05	1.02
	6	1.04	1.02	0.99
	7	1.08	1.06	1.04
	8	0.97	0.95	0.93
Female	5	1.28	1.25	1.31
	6	1.22	1.19	1.25
	7	1.26	1.23	1.30
	8	1.19	1.16	1.23
Minority Ethnicity				
Male	5	0.50	0.60	0.73
	6	0.57	0.68	0.82
	7	0.60	0.71	0.86
	8	0.70	0.83	1.01
Female	5	0.65	0.73	0.74
	6	0.58	0.65	0.65
	7	0.54	0.61	0.62
	8	0.49	0.55	0.56
White British Ethnicity				
Male	5	1.02	1.02	1.01
	6	1.02	1.02	1.01
	7	1.02	1.02	1.01
	8	1.02	1.02	1.02
Female	5	1.01	1.01	1.01
	6	1.02	1.02	1.02
	7	1.02	1.02	1.02
	8	1.02	1.02	1.02

Initial evaluation of model performance by age (Figure 6.9) shows that each model is well calibrated, with slight under-prediction of risk in women in the top age bracket.

I subsequently undertook a post-hoc analysis of performance by age, prompted by the finding that calibration by age in the PRS scores (see Figure 5.19) appeared to be miscalibrated in the middle age groups. As shown in Table 6.5, performance metrics are lower in all age bands compared to model performance overall. Each model performs best in the <50 age group, particularly for proportion of explained variation (R_2D). The improvement in performance offered by the addition of PRS is also greatest in this age group. Of note, the calibration slope is particularly poor in women over the age of 60, in both QCancer-10 and the integrated models.

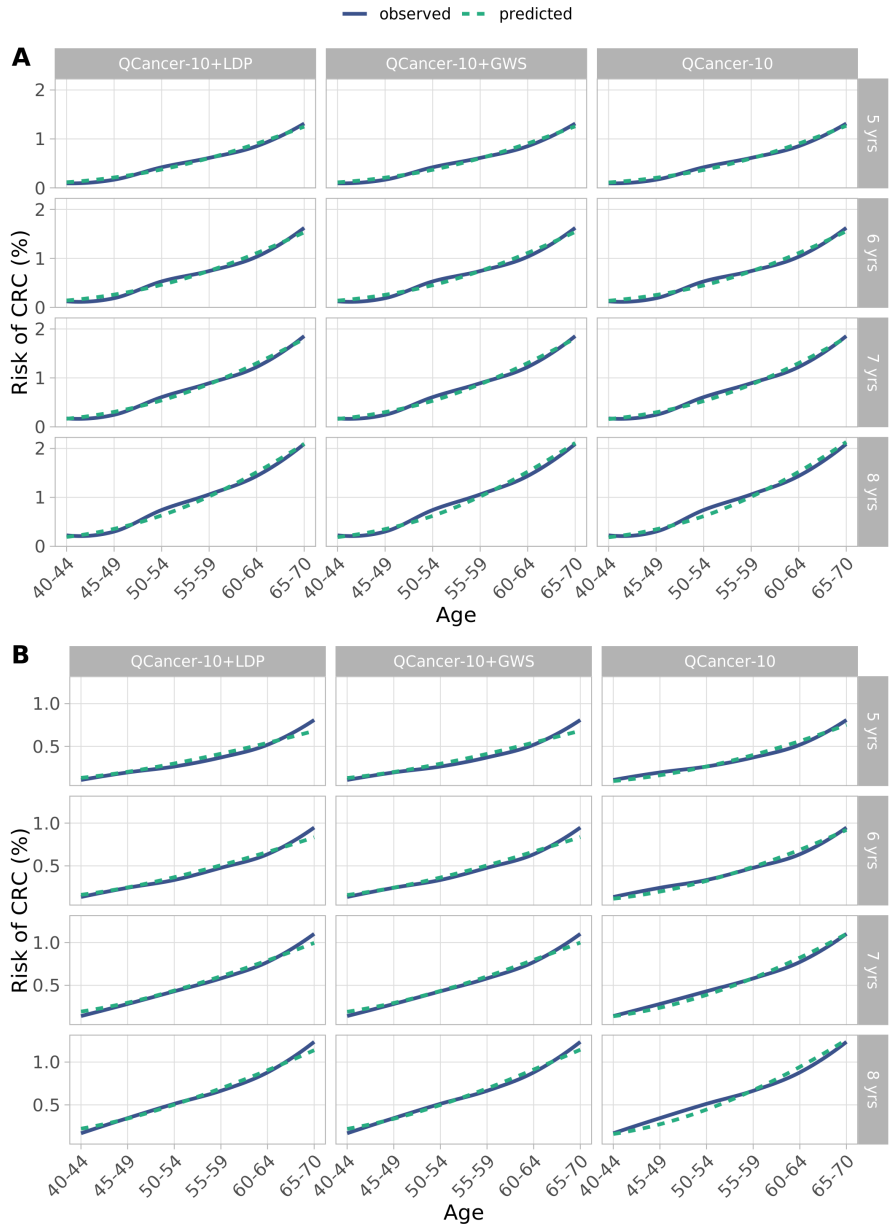

Figure 6.9: Calibration of prediction models by age in males (A) and females (B)

6.5.2 Model sensitivity and specificity

Sensitivity and specificity across the top 25 risk centiles of absolute risk are given in Tables 6.6, 6.7, and 6.8. The addition of PRS improves the sensitivity of the model, with QCancer-10+LDP having the greatest sensitivity. As an example, using a threshold of 20%, the sensitivity for QCancer-10+LDP is 47.6% in men and 42.4% in women, compared to 44.8% and 38.6% for QCancer-10+GWS, and

Table 6.5: Performance of QCancer-10+LDP, QCancer-10+GWS and QCancer-10 by age group

Age (years)	Index	QCancer-10+LDP	QCancer-10+GWS	QCancer-10
Males				
	C statistic	0.671 (0.615 - 0.723)	0.694 (0.649 - 0.747)	0.621 (0.559 - 0.676)
	Somers' D_{xy}	0.835 (0.807 - 0.862)	0.847 (0.825 - 0.874)	0.810 (0.779 - 0.838)
<50	D statistic	1.088 (0.760 - 1.417)	1.110 (0.822 - 1.440)	0.678 (0.370 - 0.979)
	R^2_D (%)	22.0 (12.1 - 32.4)	22.7 (13.9 - 33.1)	9.9 (3.2 - 18.6)
	Calibration Slope	0.989 (0.686 - 1.293)	1.120 (0.832 - 1.445)	0.865 (0.461 - 1.269)
	C statistic	0.672 (0.649 - 0.694)	0.646 (0.622 - 0.669)	0.590 (0.565 - 0.615)
	Somers' D_{xy}	0.836 (0.825 - 0.847)	0.823 (0.811 - 0.835)	0.795 (0.782 - 0.807)
50-59	D statistic	1.010 (0.870 - 1.142)	0.831 (0.694 - 0.966)	0.516 (0.382 - 0.655)
	R^2_D (%)	19.6 (15.3 - 23.7)	14.2 (10.3 - 18.2)	6.0 (3.4 - 9.3)
	Calibration Slope	1.049 (0.898 - 1.192)	0.969 (0.803 - 1.126)	0.849 (0.623 - 1.089)
	C statistic	0.656 (0.641 - 0.672)	0.638 (0.622 - 0.654)	0.608 (0.592 - 0.623)
	Somers' D_{xy}	0.828 (0.821 - 0.836)	0.819 (0.811 - 0.827)	0.804 (0.796 - 0.812)
>60	D statistic	0.859 (0.776 - 0.943)	0.776 (0.694 - 0.874)	0.578 (0.490 - 0.664)
	R^2_D (%)	15.0 (12.6 - 17.5)	12.6 (10.3 - 15.4)	7.4 (5.4 - 9.5)
	Calibration Slope	0.980 (0.883 - 1.080)	1.020 (0.911 - 1.147)	1.163 (0.986 - 1.327)
Females				
	C statistic	0.674 (0.627 - 0.724)	0.640 (0.596 - 0.688)	0.594 (0.540 - 0.644)
	Somers' D_{xy}	0.837 (0.813 - 0.862)	0.820 (0.798 - 0.844)	0.797 (0.770 - 0.822)
<50	D statistic	1.055 (0.765 - 1.343)	0.755 (0.510 - 1.004)	0.483 (0.189 - 0.773)
	R^2_D (%)	21.0 (12.3 - 30.1)	12.0 (5.8 - 19.4)	5.3 (0.8 - 12.5)
	Calibration Slope	1.185 (0.885 - 1.480)	1.032 (0.698 - 1.374)	0.729 (0.289 - 1.158)
	C statistic	0.623 (0.592 - 0.649)	0.629 (0.602 - 0.653)	0.573 (0.547 - 0.599)
	Somers' D_{xy}	0.811 (0.796 - 0.825)	0.815 (0.801 - 0.826)	0.787 (0.773 - 0.800)
50-59	D statistic	0.770 (0.581 - 0.937)	0.705 (0.549 - 0.846)	0.438 (0.280 - 0.597)
	R^2_D (%)	12.4 (7.5 - 17.3)	10.6 (6.7 - 14.6)	4.4 (1.8 - 7.8)
	Calibration Slope	0.966 (0.751 - 1.151)	1.049 (0.816 - 1.269)	0.800 (0.500 - 1.095)
	C statistic	0.626 (0.606 - 0.643)	0.595 (0.575 - 0.616)	0.549 (0.530 - 0.569)
	Somers' D_{xy}	0.813 (0.803 - 0.822)	0.797 (0.788 - 0.808)	0.775 (0.765 - 0.785)
>60	D statistic	0.696 (0.588 - 0.801)	0.509 (0.406 - 0.619)	0.257 (0.158 - 0.365)
	R^2_D (%)	10.4 (7.6 - 13.3)	5.8 (3.8 - 8.4)	1.5 (0.6 - 3.1)
	Calibration Slope	0.906 (0.777 - 1.029)	0.847 (0.674 - 1.032)	0.563 (0.328 - 0.804)

41.2% and 34.6% with QCancer-10. Detection rates followed the same pattern, with rates of 0.46%, 0.43%, 0.40% in men for QCancer-10+LDP, QCancer-10+GWS, and QCancer-10 respectively, and 0.26%, 0.24% and 0.21% for women.

We can see that, as with other performance measures, the incremental improvement is greater in women - an improvement in sensitivity of 8.2% for women compared to 6.4% for men, and a 1.24-fold increase in detection rate compared to 1.15-fold respectively. Differences in specificity and false positive rates between the models were minimal.

Table 6.6: Sensitivity and specificity of QCancer-10+LDP models for CRC diagnosis for males and females across the top 25 centiles of absolute risk

Centile	Population per centile	Absolute 5-year risk threshold (%)	Cases per centile	Cumulative % cases based on absolute risk(sensitivity)	Specificity	DR (%)	FPR (%)
Males							
1	1960	2.75	64	3.4	99.0	0.03	1.0
2	1961	2.34	61	6.6	98.0	0.06	2.0
3	1961	2.10	70	10.3	97.1	0.10	2.9
4	1961	1.94	56	13.3	96.1	0.13	3.9
5	1961	1.81	55	16.2	95.1	0.16	4.9
6	1961	1.71	57	19.2	94.1	0.19	5.9
7	1961	1.63	60	22.4	93.1	0.22	6.9
8	1961	1.55	41	24.6	92.2	0.24	7.8
9	1961	1.49	41	26.8	91.2	0.26	8.8
10	1961	1.43	48	29.3	90.2	0.28	9.8
11	1961	1.38	48	31.8	89.2	0.31	10.8
12	1960	1.33	35	33.6	88.2	0.32	11.8
13	1961	1.29	44	35.9	87.2	0.35	12.8
14	1961	1.25	31	37.5	86.2	0.36	13.8
15	1961	1.21	35	39.3	85.2	0.38	14.8
16	1961	1.17	33	41.0	84.2	0.40	15.8
17	1961	1.14	35	42.8	83.3	0.42	16.7
18	1961	1.11	32	44.5	82.3	0.43	17.7
19	1961	1.08	30	46.1	81.3	0.45	18.7
20	1961	1.05	29	47.6	80.3	0.46	19.7
21	1961	1.02	25	48.9	79.3	0.47	20.7
22	1961	1.00	32	50.6	78.3	0.49	21.7
23	1960	0.97	37	52.6	77.3	0.51	22.7
24	1961	0.95	27	54.0	76.3	0.52	23.7
25	1961	0.93	24	55.3	75.3	0.54	24.7
Females							
1	2384	1.53	58	4.0	99.0	0.02	1.0
2	2385	1.27	50	7.4	98.0	0.05	2.0
3	2385	1.13	48	10.7	97.0	0.07	3.0
4	2385	1.04	37	13.2	96.1	0.08	3.9
5	2385	0.97	37	15.7	95.1	0.10	4.9
6	2385	0.91	31	17.8	94.1	0.11	5.9
7	2385	0.87	39	20.5	93.1	0.13	6.9
8	2385	0.83	33	22.8	92.1	0.14	7.9
9	2385	0.80	29	24.8	91.1	0.15	8.9
10	2385	0.77	40	27.5	90.1	0.17	9.9
11	2385	0.74	26	29.3	89.1	0.18	10.9
12	2385	0.72	25	31.0	88.1	0.19	11.9
13	2385	0.70	19	32.3	87.1	0.20	12.9
14	2385	0.68	24	33.9	86.1	0.21	13.9
15	2385	0.66	22	35.4	85.1	0.22	14.9
16	2385	0.64	24	37.0	84.1	0.23	15.9
17	2385	0.63	14	38.0	83.1	0.23	16.9
18	2385	0.61	23	39.6	82.1	0.24	17.9
19	2385	0.60	22	41.1	81.1	0.25	18.9
20	2385	0.59	19	42.4	80.1	0.26	19.9
21	2385	0.57	17	43.6	79.1	0.27	20.9
22	2385	0.56	19	44.9	78.1	0.28	21.9
23	2385	0.55	20	46.3	77.1	0.28	22.9
24	2385	0.54	29	48.3	76.1	0.30	23.9
25	2385	0.53	15	49.3	75.1	0.30	24.9

DR - Detection Rate; FPR - False Positive Rate

Table 6.7: Sensitivity and specificity of QCancer-10+GWS models for CRC diagnosis for males and females across the top 25 centiles of absolute risk

Centile	Population per centile	Absolute 5-year risk threshold (%)	Cases per centile	Cumulative % cases based on absolute risk(sensitivity)	Specificity	DR (%)	FPR (%)
Males							
1	1960	2.46	74	3.9	99.0	0.04	1.0
2	1961	2.15	65	7.3	98.1	0.07	1.9
3	1961	1.97	44	9.6	97.1	0.09	2.9
4	1961	1.83	54	12.4	96.1	0.12	3.9
5	1961	1.72	47	14.9	95.1	0.14	4.9
6	1961	1.63	39	17.0	94.1	0.16	5.9
7	1961	1.56	53	19.8	93.1	0.19	6.9
8	1961	1.50	37	21.8	92.1	0.21	7.9
9	1961	1.44	43	24.1	91.1	0.23	8.9
10	1961	1.39	47	26.6	90.2	0.26	9.8
11	1961	1.34	36	28.5	89.2	0.27	10.8
12	1960	1.30	42	30.7	88.2	0.30	11.8
13	1961	1.27	27	32.1	87.2	0.31	12.8
14	1961	1.23	47	34.6	86.2	0.33	13.8
15	1961	1.20	30	36.2	85.2	0.35	14.8
16	1961	1.17	28	37.7	84.2	0.36	15.8
17	1961	1.14	37	39.7	83.2	0.38	16.8
18	1961	1.11	22	40.9	82.2	0.39	17.8
19	1961	1.08	35	42.7	81.2	0.41	18.8
20	1961	1.06	39	44.8	80.2	0.43	19.8
21	1961	1.03	24	46.1	79.2	0.44	20.8
22	1961	1.01	31	47.7	78.2	0.46	21.8
23	1960	0.99	31	49.3	77.3	0.48	22.7
24	1961	0.96	32	51.0	76.3	0.49	23.7
25	1961	0.94	32	52.7	75.3	0.51	24.7
Females							
1	2384	1.18	37	2.5	99.0	0.02	1.0
2	2385	1.06	39	5.2	98.0	0.03	2.0
3	2385	0.98	35	7.6	97.0	0.05	3.0
4	2385	0.93	31	9.7	96.0	0.06	4.0
5	2385	0.88	28	11.6	95.0	0.07	5.0
6	2385	0.85	30	13.7	94.0	0.08	6.0
7	2385	0.82	29	15.7	93.1	0.10	6.9
8	2385	0.79	34	18.0	92.1	0.11	7.9
9	2385	0.77	24	19.6	91.1	0.12	8.9
10	2385	0.75	25	21.3	90.1	0.13	9.9
11	2385	0.73	30	23.4	89.1	0.14	10.9
12	2385	0.71	35	25.8	88.1	0.16	11.9
13	2385	0.70	28	27.7	87.1	0.17	12.9
14	2385	0.68	20	29.1	86.1	0.18	13.9
15	2385	0.67	22	30.6	85.1	0.19	14.9
16	2385	0.65	21	32.0	84.1	0.20	15.9
17	2385	0.64	25	33.7	83.1	0.21	16.9
18	2385	0.63	22	35.2	82.1	0.22	17.9
19	2385	0.62	26	37.0	81.1	0.23	18.9
20	2385	0.61	23	38.6	80.1	0.24	19.9
21	2385	0.60	27	40.5	79.1	0.25	20.9
22	2385	0.59	26	42.3	78.1	0.26	21.9
23	2385	0.58	22	43.8	77.1	0.27	22.9
24	2385	0.57	20	45.2	76.1	0.28	23.9
25	2385	0.56	21	46.6	75.1	0.29	24.9

DR - Detection Rate; FPR - False Positive Rate

Table 6.8: Sensitivity and specificity of QCancer-10 models for CRC diagnosis for males and females across the top 25 centiles of absolute risk

Centile	Population per centile	Absolute 5-year risk threshold (%)	Cases per centile	Cumulative % cases based on absolute risk(sensitivity)	Specificity	DR (%)	FPR (%)
Males							
1	1960	1.90	49	2.6	99.0	0.02	1.0
2	1961	1.71	44	4.9	98.0	0.05	2.0
3	1961	1.60	51	7.6	97.0	0.07	3.0
4	1961	1.53	48	10.1	96.1	0.10	3.9
5	1961	1.47	51	12.8	95.1	0.12	4.9
6	1961	1.42	43	15.1	94.1	0.15	5.9
7	1961	1.38	35	16.9	93.1	0.16	6.9
8	1961	1.34	52	19.7	92.1	0.19	7.9
9	1961	1.31	36	21.6	91.1	0.21	8.9
10	1961	1.28	41	23.7	90.1	0.23	9.9
11	1961	1.25	37	25.7	89.1	0.25	10.9
12	1960	1.22	42	27.9	88.2	0.27	11.8
13	1961	1.20	38	29.9	87.2	0.29	12.8
14	1961	1.18	28	31.4	86.2	0.30	13.8
15	1961	1.16	39	33.5	85.2	0.32	14.8
16	1961	1.14	28	34.9	84.2	0.34	15.8
17	1961	1.12	36	36.8	83.2	0.36	16.8
18	1961	1.10	33	38.6	82.2	0.37	17.8
19	1961	1.08	27	40.0	81.2	0.39	18.8
20	1961	1.07	23	41.2	80.2	0.40	19.8
21	1961	1.05	27	42.6	79.2	0.41	20.8
22	1961	1.03	31	44.3	78.2	0.43	21.8
23	1960	1.01	25	45.6	77.2	0.44	22.8
24	1961	1.00	36	47.5	76.2	0.46	23.8
25	1961	0.98	22	48.7	75.2	0.47	24.8
Females							
1	2336	1.10	24	1.6	99.0	0.01	1.0
2	2344	0.98	38	4.3	98.1	0.03	1.9
3	2364	0.91	22	5.8	97.1	0.04	2.9
4	2422	0.86	21	7.2	96.1	0.04	3.9
5	2375	0.82	34	9.5	95.1	0.06	4.9
6	2203	0.80	18	10.8	94.1	0.07	5.9
7	2598	0.78	24	12.4	93.1	0.08	6.9
8	1827	0.76	25	14.1	92.3	0.09	7.7
9	2991	0.75	24	15.8	91.0	0.10	9.0
10	900	0.74	8	16.3	90.7	0.10	9.3
11	3846	0.72	38	18.9	89.1	0.12	10.9
12	2392	0.71	22	20.4	88.1	0.12	11.9
13	1543	0.70	20	21.8	87.4	0.13	12.6
14	2417	0.69	35	24.2	86.4	0.15	13.6
15	3216	0.68	33	26.5	85.1	0.16	14.9
16	2294	0.67	26	28.3	84.1	0.17	15.9
17	2328	0.66	23	29.8	83.1	0.18	16.9
18	2499	0.65	16	30.9	82.1	0.19	17.9
19	2306	0.64	23	32.5	81.1	0.20	18.9
20	2434	0.63	30	34.6	80.1	0.21	19.9
21	2418	0.62	15	35.6	79.1	0.22	20.9
22	2058	0.61	27	37.4	78.2	0.23	21.8
23	2388	0.60	16	38.5	77.2	0.24	22.8
24	2542	0.59	19	39.8	76.2	0.24	23.8
25	2539	0.58	27	41.7	75.1	0.25	24.9

DR - Detection Rate; FPR - False Positive Rate

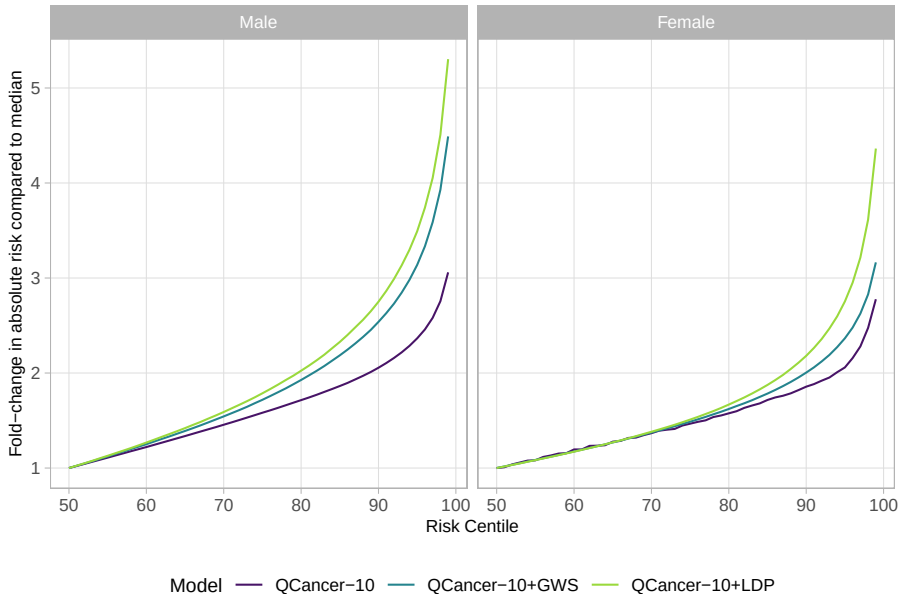

Figure 6.10: Change in absolute risk compared to the median, across the top 50 centiles of risk, for QCancer-10+LDP, QCancer-10+GWAS, and QCancer-10 models

In looking at centiles of relative risk we see a similar pattern of greater sensitivity with QCancer-10+LDP (Table 6.9) compared to QCancer-10+GWS (Table 6.10). I have not presented the results for relative risk for QCancer-10 alone, as the model did not produce sufficient variation in risk to divide the population into centile groupings.

Figure 6.10 shows the increase in predicted absolute risk threshold relative to the 50th centile for those at elevated risk. This demonstrates that it is in those at very high risk where the addition of PRS has the greatest potential to increase risk discrimination, with the QCancer-10+LDP model able to identify men at 5-fold increased risk, and women at 4-fold increase risk in the highest risk centile, compared to an approximately 3-fold increase discerned using QCancer-10. Table 6.11 gives the respective values for the 99th, 95th, and 80th centiles.

However, the corresponding changes in absolute risk between models are small. For the 95th centile, the difference in 5-year absolute risk predicted by QCancer-10+LDP compared to QCancer-10 is approximately 0.3% for men, and 0.15% for women, whilst there is no difference at the 80th centile threshold (Table 6.12).

As a further illustration of clinical utility, in current clinical practice in the UK individuals with more than one first degree relative with CRC, equating to an

Table 6.9: Sensitivity and specificity of QCancer-10+LDP models for CRC diagnosis for males and females across the top 25 centiles of relative risk

Centile	Population per centile	Absolute 5-year risk threshold (%)	Cases per centile	Cumulative % cases based on relative risk(sensitivity)	Specificity	DR (%)	FPR (%)
Males							
1	1960	4.81	55	2.9	99.0	0.03	1.0
2	1961	4.14	49	5.5	98.0	0.05	2.0
3	1961	3.78	36	7.4	97.0	0.07	3.0
4	1961	3.51	47	9.9	96.1	0.10	3.9
5	1961	3.32	32	11.6	95.1	0.11	4.9
6	1961	3.15	35	13.4	94.1	0.13	5.9
7	1961	3.01	38	15.4	93.1	0.15	6.9
8	1961	2.89	29	16.9	92.1	0.16	7.9
9	1961	2.79	40	19.0	91.1	0.18	8.9
10	1961	2.70	31	20.6	90.1	0.20	9.9
11	1961	2.62	36	22.5	89.1	0.22	10.9
12	1960	2.55	35	24.3	88.1	0.24	11.9
13	1961	2.48	40	26.4	87.1	0.26	12.9
14	1961	2.42	27	27.8	86.1	0.27	13.9
15	1961	2.36	27	29.2	85.1	0.28	14.9
16	1961	2.31	27	30.6	84.1	0.30	15.9
17	1961	2.26	22	31.8	83.1	0.31	16.9
18	1961	2.21	31	33.4	82.2	0.32	17.8
19	1961	2.17	26	34.8	81.2	0.34	18.8
20	1961	2.12	29	36.3	80.2	0.35	19.8
21	1961	2.08	28	37.8	79.2	0.37	20.8
22	1961	2.04	33	39.5	78.2	0.38	21.8
23	1960	2.01	30	41.1	77.2	0.40	22.8
24	1961	1.97	29	42.6	76.2	0.41	23.8
25	1961	1.94	29	44.1	75.2	0.43	24.8
Females							
1	2199	3.89	46	3.2	99.1	0.02	0.9
2	2570	3.26	34	5.5	98.0	0.03	2.0
3	2384	2.90	41	8.3	97.0	0.05	3.0
4	2385	2.66	28	10.2	96.0	0.06	4.0
5	2386	2.48	29	12.2	95.0	0.07	5.0
6	2385	2.34	29	14.2	94.1	0.09	5.9
7	2385	2.23	25	15.9	93.1	0.10	6.9
8	2384	2.14	22	17.4	92.1	0.11	7.9
9	2386	2.06	26	19.2	91.1	0.12	8.9
10	2385	1.99	27	21.1	90.1	0.13	9.9
11	2385	1.92	26	22.9	89.1	0.14	10.9
12	2384	1.86	33	25.2	88.1	0.15	11.9
13	2385	1.81	21	26.6	87.1	0.16	12.9
14	2385	1.77	17	27.8	86.1	0.17	13.9
15	2386	1.72	14	28.8	85.1	0.18	14.9
16	2385	1.69	15	29.8	84.1	0.18	15.9
17	2385	1.65	17	31.0	83.1	0.19	16.9
18	2384	1.62	20	32.4	82.1	0.20	17.9
19	2386	1.58	19	33.7	81.1	0.21	18.9
20	2384	1.55	25	35.4	80.1	0.22	19.9
21	2385	1.52	18	36.6	79.1	0.22	20.9
22	2386	1.50	20	38.0	78.1	0.23	21.9
23	2385	1.47	21	39.4	77.1	0.24	22.9
24	2385	1.45	21	40.8	76.1	0.25	23.9
25	2385	1.42	19	42.1	75.1	0.26	24.9

DR - Detection Rate; FPR - False Positive Rate

Table 6.10: Sensitivity and specificity of QCancer-10+GWS models for CRC diagnosis for males and females across the top 25 centiles of relative risk

Centile	Population per centile	Absolute 5-year risk threshold (%)	Cases per centile	Cumulative % cases based on relative risk(sensitivity)	Specificity	DR (%)	FPR (%)
Males							
1	1960	4.08	36	1.9	99.0	0.02	1.0
2	1961	3.54	45	4.3	98.0	0.04	2.0
3	1961	3.26	47	6.8	97.0	0.07	3.0
4	1961	3.06	41	9.0	96.0	0.09	4.0
5	1961	2.90	33	10.7	95.1	0.10	4.9
6	1961	2.78	33	12.4	94.1	0.12	5.9
7	1961	2.68	33	14.1	93.1	0.14	6.9
8	1961	2.58	36	16.0	92.1	0.16	7.9
9	1961	2.50	38	18.0	91.1	0.17	8.9
10	1961	2.43	43	20.3	90.1	0.20	9.9
11	1961	2.37	30	21.9	89.1	0.21	10.9
12	1960	2.31	26	23.3	88.1	0.22	11.9
13	1961	2.26	23	24.5	87.1	0.24	12.9
14	1961	2.21	26	25.9	86.1	0.25	13.9
15	1961	2.16	23	27.1	85.1	0.26	14.9
16	1961	2.12	18	28.0	84.1	0.27	15.9
17	1961	2.08	36	29.9	83.1	0.29	16.9
18	1961	2.04	25	31.2	82.1	0.30	17.9
19	1961	2.01	25	32.5	81.1	0.31	18.9
20	1961	1.98	28	34.0	80.1	0.33	19.9
21	1961	1.94	25	35.3	79.1	0.34	20.9
22	1961	1.91	25	36.6	78.1	0.35	21.9
23	1960	1.89	26	38.0	77.1	0.37	22.9
24	1961	1.85	24	39.3	76.1	0.38	23.9
25	1961	1.83	21	40.4	75.2	0.39	24.8
Females							
1	2383	2.63	19	1.3	99.0	0.01	1.0
2	2386	2.36	35	3.7	98.0	0.02	2.0
3	2384	2.23	22	5.2	97.0	0.03	3.0
4	2386	2.12	24	6.8	96.0	0.04	4.0
5	2385	2.02	20	8.2	95.0	0.05	5.0
6	2385	1.95	22	9.7	94.0	0.06	6.0
7	2384	1.89	19	11.0	93.0	0.07	7.0
8	2385	1.84	20	12.4	92.0	0.08	8.0
9	2385	1.79	17	13.6	91.0	0.08	9.0
10	2385	1.75	33	15.9	90.0	0.10	10.0
11	2386	1.72	26	17.7	89.0	0.11	11.0
12	2385	1.68	20	19.1	88.0	0.12	12.0
13	2385	1.65	18	20.3	87.0	0.12	13.0
14	2385	1.62	21	21.7	86.0	0.13	14.0
15	2385	1.60	25	23.4	85.1	0.14	14.9
16	2385	1.57	19	24.7	84.1	0.15	15.9
17	2385	1.55	22	26.2	83.1	0.16	16.9
18	2385	1.53	19	27.5	82.1	0.17	17.9
19	2385	1.51	32	29.7	81.1	0.18	18.9
20	2384	1.49	21	31.1	80.1	0.19	19.9
21	2386	1.47	21	32.5	79.1	0.20	20.9
22	2384	1.45	21	33.9	78.1	0.21	21.9
23	2386	1.43	17	35.1	77.1	0.22	22.9
24	2384	1.41	19	36.4	76.1	0.22	23.9
25	2386	1.40	13	37.3	75.1	0.23	24.9

DR - Detection Rate; FPR - False Positive Rate

Table 6.11: Fold-increase in absolute risk compared to the 50th centile, for QCancer-10+LDP, QCancer-10+GWS, and QCancer-10 models

Risk Centile	QCancer-10+LDP	QCancer-10+GWS	QCancer-10
Males			
99	5.30	4.49	3.06
95	3.49	3.14	2.37
80	2.02	1.93	1.71
Females			
99	4.36	3.16	2.78
95	2.75	2.37	2.06
80	1.67	1.62	1.58

Table 6.12: 5-year absolute risk predicted by QCancer-10+LDP, QCancer-10+GWS, and QCancer-10 models

Risk Centile	QCancer-10+LDP	QCancer-10+GWS	QCancer-10
Males			
99	2.75	2.46	1.90
95	1.81	1.72	1.47
80	1.05	1.06	1.07
Females			
99	1.53	1.18	1.10
95	0.97	0.88	0.82
80	0.59	0.61	0.63

approximately 2.2-fold risk increase [486], are often offered enhanced colonoscopic screening. Table 6.13 shows the proportion of the study population with RR $>$ 2.2, showing that the addition of PRS improves identification of these individuals.

6.5.3 Decision curve analysis

Decision curve analyses demonstrate that all models outperform the default strategies across most relevant thresholds, and that the QCancer-LDP model has the highest net benefit (NB) across most risk thresholds considered (Figure 6.11). However, the NB is quite small. At a threshold probability of 0.5% (i.e. willing to perform colonoscopy for 200 individuals to detect 1 cancer, or NNS of 200) the NB for QCancer-10+LDP in men is 0.00701 true positives, or 0.7 net detected cancers without unnecessary colonoscopies per 100 patients. At a threshold of 1% NB is 0.00429, or 0.4 true positives per 100 screened. In women, NB at threshold probability of 0.5% is 0.00271 (0.2 true positives per 100 screened), and at a threshold probability of 1% this was 0.00090, or 0.09 true positives per 100

Table 6.13: Percentage of study population and CRC cases with relative risk >2.2 for QCancer-10+LDP, QCancer-10+GWS, and QCancer-10 models

	QCancer-10+LDP		QCancer-10+GWS		QCancer-10	
	Males	Females	Males	Females	Males	Females
% population with RR > 2.2	18.2	7.2	14.2	3.2	4.1	1.2
% of individuals with RR > 2.2 without FDRCRC	75.9	69.6	70.8	44.8	29.4	30.3
% cases with RR > 2.2	34.0	16.5	26.3	6.0	4.9	1.6

RR - relative risk; FDRCRC - first-degree relative with colorectal cancer

screened. In comparison to these, the default strategy of colonoscopy for all has a lower NB of 0.00620 in men and 0.00186 in women at a probability threshold of 0.5%, and the NB is negative for both sexes at a threshold probability of 1%, indicating that this strategy would be harmful.

The slightly lower net benefit than the default treat all strategy at the lowest threshold probabilities (below 0.05% in men and 0.1% in women) represents slight miscalibration of the model at this level, as do the curves falling below no intervention at the very highest risk thresholds (Van Calster et al. [479] note that models cannot be harmful, i.e. perform worse than the default strategy, unless they are miscalibrated).

At a threshold probability of 1.0%, the difference in net benefit (ΔNB) for men with the addition of the LDPred2 PRS is 0.00067, with a test trade-off of 1504. If we are happy to accept doing PRS on 1504 people for one additional cancer detected, then the added utility of the addition of PRS is worth the cost. In women, at a threshold probability of 1.0% ΔNB was 0.00058, and test trade-off 1731. Values at thresholds of 0.5% to 2% are presented in Table 6.14.

One may also look at the unnecessary interventions avoided to consider whether the addition of PRS is worthwhile (Figure 6.11). At a threshold of 1%, using the QCancer-10 model avoids approximately 24.2 colonoscopies per 100 men compared to a colonoscopy for all strategy. The addition of the LDpred2 PRS results in 30.8 avoided procedures. Thus an additional 6.6 colonoscopies are saved per 100 men at this threshold. Values for other thresholds are shown in Table 6.15. The additional interventions avoided by adding the LDpred2 PRS, above using QCancer-10 alone, range from 3.1 to 6.6 in men and 0.9 to 5.7 in women, depending on threshold chosen.

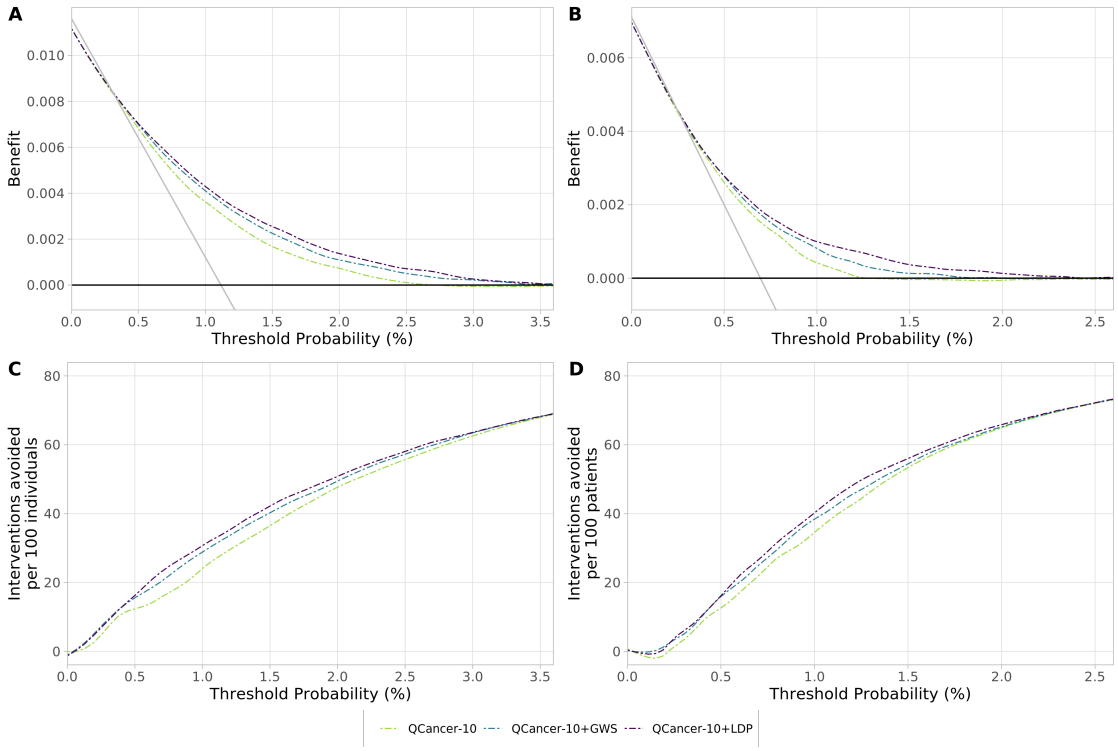

Figure 6.11: Decision curve analysis and interventions saved, calculated at 8 years of follow-up

Table 6.14: Net benefit and test trade-off for QCancer-10+PRS and QCancer-10 models

Threshold Probability (%)	Net Benefit			ΔNB with addition of LDP PRS	Test trade-off
	QCancer-10 +LDP	QCancer-10 +GWS	QCancer-10		
Males					
0.5	0.00701	0.00695	0.00682	0.00020	5076
1.0	0.00429	0.00408	0.00362	0.00067	1503
1.5	0.00252	0.00224	0.00168	0.00084	1185
2.0	0.00137	0.00110	0.00075	0.00062	1604
Females					
0.5	0.00271	0.00270	0.00253	0.00018	5493
1.0	0.00099	0.00079	0.00042	0.00058	1731
1.5	0.00037	0.00012	-0.00004	0.00041	2451
2.0	0.00013	0.00001	-0.00006	0.00019	5228

Table 6.15: Unnecessary interventions avoided QCancer-10+PRS and QCancer-10 models

Threshold Probability (%)	Interventions avoided			Δ with addition of LDP PRS
	QCancer-10 +LDP	QCancer-10 +GWS	QCancer-10	
Males				
0.5	16.2	14.9	12.3	3.9
1.0	30.8	28.7	24.2	6.6
1.5	42.1	40.2	36.5	5.5
2.0	50.9	49.5	47.8	3.1
Females				
0.5	16.6	16.3	13.0	3.5
1.0	40.6	38.6	34.9	5.7
1.5	56.0	54.4	53.4	2.7
2.0	66.0	65.4	65.1	0.9

6.6 Discussion

In this chapter, I report the first risk models for colorectal cancer to incorporate genome-wide PRS with non-genetic risk predictors. I demonstrate that integrating PRS with the QCancer-10 non-genetic model modestly improves model performance, and that genome-wide models developed using LDpred2 conferred the best performance. Models combining non-genetic risk predictors with GWAS-significant PRS have previously shown that these outperform either PRS or non-genetic risk models used in isolation [215, 238]. The work presented here shows a step-wise improvement on this with the use of genome-wide PRS in integrated models.

The sensitivities achieved with QCancer-10+PRS models exceeded those of QCancer-10 alone, and of previously published evaluations of integrated models in UKB [215]. For the top 20% at risk, sensitivity with QCancer-10+LDP is 47.6% for men and 42.4% for women, very close to the sensitivity of faecal immunochemical testing (FIT) currently in use in the BCSP (47.8% at a FIT threshold of 120μgm/g) [487]. For the QCancer-10 model, the same sensitivity is reached by including the top ~25% at risk.

In addition to this, decision curve analysis, in terms of both net benefit and the number of unnecessary interventions avoided, showed that QCancer-10+LDP was the superior model across a range of probability thresholds in both men and

women. In simplistic terms, in decision curve analysis the model with the greatest net benefit would be the model of choice.

However, for most metrics the improvements in performance are small. The increase in C-statistic obtained by adding PRS to QCancer-10 is just 0.04 with the addition of the LDpred2 model, though this is metric is well recognised as difficult to increase, even if new predictors with large effect sizes are introduced to a model [488]. The improvement seen in explained variation with QCancer-10+LDP, of 7% for males and 8.5% for females, suggests that the QCancer-10+PRS models better represent CRC risk.

Absolute risk, which is generally key to clinical decision making, is an important consideration [489]. The differences in absolute risk identified by the models is small, with a 0.15% difference in 5-year absolute risk in the top 5% risk threshold between QCancer10+LDP and QCancer-10 for women, and a 0.3% difference for men. Detection rates were also modestly improved with PRS.

Small incremental improvements in risk prediction afforded by the PRS may not be worth the potential costs (in the broadest sense) of implementing PRS in bowel cancer screening. Decision curve analysis can be useful in considering the clinical utility of a risk-prediction model, particularly when complex modelling and cost-effectiveness analysis is not possible. Test trade-off analysis showed that, assuming a simple scenario of colonoscopy based screening, between ~1180 and ~5500 LDpred2 PRS tests (varying by preferred probability threshold and sex) would need to be performed to identify one additional cancer over 8 years of follow-up. These numbers may be considered too great to be worth implementing testing. As an alternative measure, the number of colonoscopies saved by the addition of LDpred2 PRS to QCancer-10 was between 0.9 and 6.6 per 100 screened individuals, which may be considered too few.

The DCA analysis has limitations. Bowel cancer screening entails multiple rounds of screening, with the PRS undertaken only at the first screening round but providing benefit over successive rounds, and thus this is a fairly simplified approximation. The difference in "costs" of implementing PRS above QCancer-10

will also vary depending on implementation. QCancer-10 could be relatively easily implemented using linked primary care data. Roll-out of PRS specifically for bowel cancer screening would be quite costly, requiring significant changes to testing and infrastructural change within the screening programme. If, however, PRS were to become a standard part of national health records, as is the aim of initiatives in the UK such as "Our Future Health" [490], this may ease utilisation and reduce costs of a combined score. Full evaluation of the financial costs, environmental impact, and effects on screening participation would need to be evaluated to assess whether PRS-implementation is "worth it".

Several other observations around the results presented in this chapter are worthy of further discussion. With regard to QCancer-10, performance estimates on external validation in my Integrated Modelling Cohort were in line with previous external validation in the UKB cohort as a whole [491]. However, in both studies model performance was considerably lower than the external validation performed in the original QCancer-10 study, where the ROC and R^2 were 0.84 and 46.3% for women, and 0.851 and 49.1% for men [182]. This is likely to be in a large part due to the much narrower age range available in UKB, compared to the original study which included 25-84 year olds. Age is a particularly strong predictor of CRC risk, relative to other predictors in the model, and thus this will have an outsized impact on model performance. In addition, the distribution of other predictors are also narrower in UKB due to the healthy cohort effect [266]. Of note, an alternative QCancer (Colorectal) score developed for symptomatic individuals using the QResearch primary care dataset [492] maintained performance compared to the derivation study on external validation in another primary care dataset [493]. Despite these observations, the narrower age range of UKB more closely reflects that of bowel screening participants, and thus model performance here may be more relevant for risk-stratified screening, though one might expect that these are underestimates of likely model performance in a truly population-based cohort of the same age due to the healthy cohort effect.

The weak interaction between PRS and QCancer-10 scores in this analysis indicate a smaller effect of QCancer-10 at higher PRS scores. This could indicate that with a higher genetic predisposition, environmental exposures may have a lesser impact on CRC risk. Prior studies have found limited evidence for gene-environment interactions between individual lifestyle factors (such as smoking, alcohol, BMI, dietary intake, and aspirin use) and genetic risk loci [494–496]. Other studies have examined interaction between summative PRS and individual non-genetic risk factors and found no evidence to support these [217, 223, 497]. Further studies have looked at the interaction between PRS and composite scores of environmental exposures for CRC, with mixed results. Choi et al. [498] developed a "Healthy lifestyle score" (HLS, incorporating BMI, waist-to-hip ratio, physical activity, sedentary time, processed and red meat intake, vegetable and fruit intake, and alcohol and tobacco exposures) and used a 95-variant PRS to evaluate relationships between these and CRC risk in UKB. They categorised the HLS and PRS into 3 groups, and counter to my results found a greater RR reduction between higher and lower HLS at higher levels of PRS, concluding that lifestyle changes might be of greater benefit to those at high genetic risk. On the other hand, Carr et al. [499] and Erben et al. [500] found no variation in the relative risk of a healthy lifestyle score incorporating similar lifestyle exposures with genetic risk for CRC or colorectal adenoma in the German population. Thus the evidence for interactions between environmental and genetic risk scores appears to be contradictory, and was not strongly supported in this study. Potential interactions ought to be considered in future modelling work.

The poorer performance seen across all metrics in subgroup analysis by age, compared to model performance overall, reflects the very strong effect of age on risk, the impact of which is minimised when examined across a narrower age bracket. Integrated models performed best in under 50 year olds (as did PRS, discussed in Chapter 5). Genetic risk likely accounts for a greater proportion of an individual's risk at this age, relative to environmental and health exposures, many of which

accrue over time. Poor calibration of models in older women could be due to the healthy cohort effect, which may be more pronounced in older participants.

As noted for PRS, models for women performed less well than those for men. The healthy volunteer bias seen in UKB is more pronounced in women. Fry et al. [266] noted that the reduction in total cancer incidence and all-cause mortality in UKB compared to the general population was larger in women than men. As noted in Chapter 4 CRC case numbers for women also fell slightly short of sample size requirements, and consequently the risk estimates for women presented here might be less precise. Extension of this analysis with the updated follow-up duration now available in UKB (see Section 4.7) would increase the power available and accuracy of performance estimates for women. External validation of model performance would be essential prior to implementation.

Models also performed less well in individuals from minority ethnic backgrounds. The scope for my analysis here was limited by low numbers of incident cases in this group (46 cases in men, 58 in women). As discussed in Chapter 5, demographic differences in minority ethnic participants may have exacerbated this, including a younger mean age contributing to lower incidence of CRC cases, and greater missingness exacerbating under-representation in my Integrated Modelling Cohort. Given the low case numbers, I restricted my evaluation to examination of E/O ratios across 5-8 years follow-up, which showed that models were less well calibrated in self-identified minority ethnic individuals, with E/O of 0.70 for QCancer-10+LDP, and 0.83 for QCancer-10+GWS at 8 years in men, compared to 1.01 for QCancer-10 alone (and suggesting that the majority of this discrepancy arose from the PRS). Interestingly in women of minority ethnicities, E/O worsened over extended follow-up. For example, at 5 years follow-up, E/O was 0.65 for QCancer-10+LDP and 0.74 for Qcancer-10 alone; at 8 years follow-up these values were 0.49 and 0.56. The reason for this is unclear, but should be interpreted with caution given the low case numbers. In contrast, for both men and women of White ethnicity, E/O was ~1 at all periods of follow-up. A careful evaluation of integrated model performance in minority ethnic screenees would be essential prior to implementation.

6. Intergrated-risk-models

Considerable overlap might be anticipated between family history and genetic risk, which are modelled separately in this study. In addition to shared genetics, family history incorporates shared environmental exposures which may not be captured by other predictors in the model. Weigl et al. [501] evaluated the relationship between family history and a 58-SNP PRS in over 2,400 individuals, modelling family history and PRS both separately and jointly in logistic regression models, with adjustment for a number of confounders. PRS and family history were essentially independent predictors, with joint modelling improving risk prediction compared to either predictor alone. Jenkins et al. [502] also found PRS and family history to be uncorrelated in a population-based case control study using a PRS of 45 SNPs. They modelled the potential differences in screening age onset based on PRS and family history, assuming screening at a threshold of 0.3% five year risk. For those without a positive family history, screening began 4 years earlier for women and 5 years earlier for men in those in the highest quintile of PRS risk. With two first-degree relatives with CRC, individuals additionally in the highest PRS quintile would begin screening 16 years earlier. Thus the inclusion of both family history and PRS in the integrated models appears reasonable. Coding of family history in this modelling is quite crude - in QCancer-10 development it is a binary yes/no query, and in UKB only first degree relatives were queried. In a high risk CRC clinic, a more refined assessment of family history is used to assess risk. Zheng et al. [503] have recently shown improved discrimination of risk with a more complex assessment of family history, a 'family risk profile' (incorporating detailed family history including relationships and age at diagnosis, probability of carrying mutations in MMR and MUTYH genes, and a modelled polygenic component), compared to binary assessment of history in a first degree relative. The data capture for this is considerably more complex, and would be harder to implement in a population screening programme. Family history appears to have a stronger effect on CRC risk at younger ages [182, 504], which is incorporated into the QCancer-10 score as an interaction term.

6. Intergrated-risk-models

A further limitation of my UKB modelling work is the lack of assessment of advanced adenoma as an outcome. This was restricted by the lack of detail on adenoma diagnosis and pathology available in UKB. Prediction of advanced adenoma is highly relevant to improving bowel cancer screening - as previously noted, part of the benefit of which is derived from removing pre-cancerous lesions. Increasing polygenic risk scores have previously been shown to be associated with increasing advanced adenoma prevalence, but not of increased non-advanced adenoma prevalence [505]. Another study found PRS to be associated with adenomatous polyp diagnosis at screening colonoscopy, though when evaluating advanced adenomas they did not find a significant association, likely due to lack of power [506]. In addition FIT is not as sensitive in predicting advanced adenoma as it is in predicting CRC, and in implementing a risk score alongside FIT, it is possible that the incremental benefit of risk stratification might be greater for advanced adenoma. In the future, UKB plans to link to Bowel Cancer Screening Data, which would be a valuable resource in exploring this further.

The modelling work presented here and in Chapter 5 also does not take into account participation in bowel cancer screening (though prior diagnosis of polyps is included in the QCancer-10 score), which is an important predictor of future bowel cancer risk in those of screening age. Participation in bowel screening prior to enrolment in UKB would reduce the likelihood of CRC in individuals with higher risk. Guo et al. [507] found that there was a reduction in AUROC for PRS predicting CRC risk in those who had previously participated in colonoscopy based screening (AUROC 0.568 compared to 0.622 in non-screenees). The impact on this study is likely to be less marked as UKB participants would have undergone FOBT-based screening (with FOBT less sensitive for adenomas), but this could impact on performance estimates to a degree. However this effect would have been the same for all risk models. Were I to conduct this study again I would evaluate the impact of participation in bowel screening in a sensitivity analysis, and inclusion of screening participation would also be important for longitudinal risk estimates throughout the duration of an individual's screening eligibility.

6. Intergrated-risk-models

A further approach which I would have considered had time allowed is competing risks analysis, which incorporates consideration of the chance that non-CRC related death may occur prior to a CRC diagnosis into modelling [508]. This is particularly relevant in an ageing population, and in an age-related condition such as CRC. Competing risks can result in over-prediction of risk in Cox proportional hazards models.

The risk scores developed here predict that ~10% of individuals aged ~40-70 have sufficiently high relative risks of CRC to justify enhanced screening, based on current British Society for Gastroenterology guidelines for those with higher familial risk [65]. These guidelines recommend increased surveillance for individuals with "moderate risk" familial history, which is expected to confer relative risk of 2-6-fold above the general population. In the context of a family history of CRC, an individual is considered at moderate risk with one first degree relative (FDR) with CRC below 50 years of age, or two FDRs in first degree kinship with CRC at any age, of whom the individual must be a FDR of at least one. Individuals with moderate familial risk are recommended to undergo one-off colonoscopy at the age of 55, with subsequent follow-up determined by post-polypectomy guidance. This age recommendation is guided by the observation in cohort studies that individuals with family history of CRC have very little increase in adenoma rates prior to 50 years of age [509, 510]. The authors note that with the increasing incidence of young-onset CRC, the age recommendations may need to be reviewed.

Of the individuals identified in this study as having a level of risk which would justify enhanced screening, ~70% (using the QCancer-10+LDP model) did not have a family history of CRC, and thus would not be identified through high risk clinics. This suggests that screening as it currently stands is insufficient for a relatively large number of individuals. Indeed, given that it is those with very high risk in whom the addition of PRS is beneficial (as evidenced by the large increase in risk discerned in those above the 90th centile of risk with the additional of LDP described in Section 6.5.2), a large part of the utility of PRS may be in identifying those at highest risk for enhanced screening in some form.

Overall, the results presented in this chapter demonstrate that whilst the addition of PRS to QCancer-10 does improve performance, the impact of this is modest at best. I would postulate that in the general population, the relative added benefit of PRS to QCancer-10 may be even smaller, as QCancer-10 itself is likely to perform better outside of the "healthy population" of UKB, giving a higher "baseline". The results presented here suggest that integrated models might have a higher chance of exacerbating inequalities, due to poorer performance in minority ethnic groups, and in women compared to men. I discussion this issue further in the final chapter.

Discussion

My thesis explores genetic risk for colorectal cancer, and the potential of polygenic risk scores to improve prediction of colorectal cancer. At present, prevention and early diagnosis is a key healthcare focus in England [58], and refined risk prediction has been heralded by many as playing a significant role. In Professor Sir Mike Richards' Independent Review of Screening in 2019 (commissioned by NHS England in November 2018 following serious incidents in breast and cervical screening) he emphasised the increasing importance of targeted screening over the next decade, anticipating implementation of PRS and other stratifying tests, which would become more feasible and affordable [154]. Politically this has been an area of intense interest, with the Genome UK report highlighting the use of genomics in stratified prevention as a key focus [511]. In addition, commercially a number of companies have been quick to cash-in on this aspect of the "genomics revolution", at times with relatively limited test validation (leading to 23andMe being forced to withdraw its initial health testing by the FDA [512]).

In this final chapter, I will briefly summarise the results presented in this thesis, and then explore some of the wider narrative around risk prediction and polygenic risk scores. Finally, I will consider future directions of research in this field.

I began in Chapter 3 with a search for new CRC risk loci. My GWAS demonstrated a high type 1 error rate, and my investigation of this highlighted some fundamental issues with GWAS, including the identification of an appropriate control set, and approaches to imputation. The GWAS meta-analysis identified 31 new loci for CRC, 8 new SNPs at previously identified loci, and validated 9 SNPs previously reported in Asian populations. This work provided evidence, alongside functional investigation, of genes and pathways of particular relevance for future study of

CRC pathogenesis. In my linkage analysis, I identified several new possible loci and candidate genes, the most interesting of which was a missense variant in *GFI1*. I replicated previously identified linkage loci at 4q13.1 and 11q23, where I identified a 3'-UTR for *C2CD2L* and an intronic variant in *MCAM*. Two identified loci (3q26.2 and 5p15.33) replicated GWAS loci, and at the former I found a 3'-UTR variant of *EIF5A2*. All of these genes could plausibly be involved in CRC pathogenesis based on function and previous studies. Notably, none of the linkage peaks reached nominal significance, likely due to the size of the families included in the analysis.

In Chapter 4, I described the UKB dataset. In the context of CRC in UKB I found, as expected, evidence of a healthy cohort bias. The lower CRC incidence noted in UKB means that the prediction models developed within it, including those developed in subsequent chapters here, would need recalibrating for the general population. The small numbers of minority ethnic participants in UKB also had implications for my research and highlights the need for more diverse recruitment into these high value, high definition datasets. The consequent low numbers of CRC cases precluded thorough evaluation of my combined risk models in minority groups, an issue which I shall explore in more detail later in this discussion.

In Chapter 5 I developed and evaluated a number of different PRS for CRC. In excluding the UKB GWAS from my base dataset, I avoided bias often evident in PRS studies as a result of overlap between the base and modelling datasets. Genome-wide models developed using LDpred2 performed best, with the sparse model containing ~600k SNPs performing as well as the non-sparse model with ~1.2million. All PRS performed less well in women compared to men, and were poor predictors of risk in minority ethnic participants. Models underpredicted risk in individuals with a family history of CRC, emphasising the need to evaluate family history in addition to PRS (and I would argue, by extension, screen for Mendelian variants in a population setting).

In Chapter 6 I present novel research which developed risk prediction models integrating genome-wide PRS with phenotypic risk scores. Integrated models outperformed QCancer-10 alone across all metrics, with genome-wide PRS giving

greatest improvement. However the absolute improvements in performance were relatively small, and decision curve analysis showed only modest improvements with PRS. Some discrepancies in subgroup performance remained: though the analysis was by necessity simple, QCancer-10 was well calibrated for male minority ethnic participants but this was worsened considerably with the addition of PRS. All models performed less well in women (though these results should be interpreted with caution as the analysis is slightly underpowered); this may be in part due to the characteristics of UKB, with a more pronounced healthy volunteer bias in women. In contrast to PRS alone, QCancer-10+PRS models were well calibrated in men with a family history of CRC; in women, however, all models continued to under-predict risk. Overall, the small benefits in predictive performance gained by adding PRS may not currently justify their implementation, particularly given these desparities in performance, and the relative ease of implementation of QCancer-10.

The recently convened Polygenic Risk Score Task Force (part of the International Common Disease Alliance), advocate for "responsible use" of PRS, which they define as "use of a PRS where there are clear benefits that outweigh risks, and where effort is taken towards a goal of equitable benefit for all" [513]. Whether PRS might be used "responsibly" in bowel cancer screening, now or in future, and areas for further development of risk-stratified bowel screening, will be the subject of the rest of this discussion.

Implementation of risk-stratified screening

In terms of the options for implementing risk-stratified cancer screening, a number of approaches are possible. Depending on risk, one might begin screening earlier, modify screening intervals, or change modality (for example colonoscopy instead of FIT). As an example, the recommended 10-yearly colonoscopic screening interval could potentially be extended for individuals with low PRS risk following negative colonoscopies [514]. In the case of FIT-based screening, one could also change the positivity threshold. Evidence suggests that reducing screening for low-risk groups may not be socially acceptable [515, 516]. The optimal approach, and risk

thresholds, may vary in different populations and healthcare systems, depending for example on current screening modality, disease prevalence, acceptability, and cost.

Sampling and genotyping might be performed once for a specific purpose (for example for bowel screening at the first screening round), or as part of a multi-PRS test, potentially retaining the sample and/or data over many years for repeated use. The timing of risk assessment would vary depending on approach. Given the steadily increasing rates of early onset CRC, and the seemingly stronger effect of PRS and family history at younger ages, it is potentially in improving cancer prevention for younger age groups that PRS would have the greatest benefit. Although the absolute risk of CRC in under 50's is low (UK incidence rates are 22 cases per 100,000 women and 24 cases per 100,000 men aged 45-49, compared to 119 and 198 per 100,000 women and men in 65-69 year olds [517]), some screening recommendations now advise screening under 50's (The US Preventive Services Task Force 2021 recommendations suggest screening is offered from 45 years [30]). Microsimulation modelling from the Cancer Intervention and Surveillance Modelling Network (CISNET) demonstrated that compared to screening from 50, screening from 40 could moderately increase life-years gained and reduce CRC cases and mortality [518]. Notably, as population risk factors and disease incidence changes over time, any implemented model would need to be regularly updated and recalibrated. If screening for high and moderate penetrance cancer susceptibility genes were also incorporated, testing as early as 20-30 might be needed to maximise opportunities for cancer prevention and risk reduction [519]. Some have advocated introducing genotyping into the Newborn Screening Programme, but this would potentially have significant ethical and legal ramifications [520].

Projects such as the 100,000 Genomes project in the UK are evaluating the governance and data infrastructure required for population-scale genomic testing [79]. Significant infrastructural change will be needed within screening programmes to implement risk-stratified screening, including urgent updating of outdated IT systems (already being overseen by the NHS digital transformation directorate, NHSX), facilitating more data-driven screening pathways. The optimal approach to

risk-stratified screening needs to be carefully considered, both in terms of logistical feasibility, cost-effectiveness, and acceptability. However, a strong argument for the use of QCancer-10 above PRS-based methods is the potential to use linked primary care data to calculate an individuals' risk score, rather than requiring extra testing. This places less burden on screening participants, which would likely reduce impacts on uptake.

Clinical trials are needed to evaluate the feasibility of risk prediction models within the UK healthcare system, and assess whether these models have clinical utility, for which they must address an unmet need and improve on current pathways [521]. Richards [154] noted that although the NHS and national screening programmes put UK researchers at the forefront of screening research, screening processes, data infrastructure, and difficulties in accessing this data, have historically hindered these efforts considerably. Concerted effort is now being made to change this [159]. Several trials have evaluated whether adding non-genetic risk predictors to FIT might improve population screening, as discussed in Section 1.4.1. Results are awaited for a randomised trial of non-genetic risk-based screening compared to standard FIT screening in 23,000 Dutch screening participants [522].

There are, to my knowledge, no current trials evaluating PRS-based stratification in bowel cancer. However, in prostate cancer, the BARCODE-1 study is enrolling 5,000 men to evaluate a community based screening programme, in which men aged 55-69 (of European ancestry) are invited to provide a saliva sample for PRS calculation, with those in the top 10% of risk invited for MRI and biopsy. The pilot feasibility study demonstrated uptake of just 26%, 17 of 25 men at high risk invited for MRI underwent the procedure, and the 7 cancers detected were all low risk [523]. Though a pilot, this reinforces a number of potential issues with PRS approaches, including restriction to Europeans, the identification of indolent disease, and poor uptake. In breast cancer, myPeBS randomised screening trial is recruiting 85,000 women from Belgium, France, Italy, Spain, the UK, and Israel, to compare risk stratification based on the MammoRisk® tool, incorporating PRS

and other risk factors [524]. The WISDOM trial is also evaluating risk-informed breast screening compared to standard of care [525].

Effective risk communication is also essential, particularly at the scale of population screening, and the standardisation of this process with clear guidelines would be beneficial [513]. The difficulties of this were clearly demonstrated by MP Matt Hancock's public pronouncement that PRS testing had "saved my life", and his commitment to undertake (unevidenced) prostate cancer screening based on his predicted 15% lifetime risk of prostate cancer - fractionally higher than the average [526]. Discussion is needed of both the estimated level of risk, and the uncertainty around these estimates [513, 527]. The optimal approach to accessible and culturally appropriate risk explanation is an area of current research. Whilst some would advocate for expert support, others, particularly in the context of the profusion of direct-to-consumer testing, feel this is unnecessary [528]. Adequate training of healthcare professionals will be required to understand the implications of PRS-based risk stratification - efforts are already underway to improve this for cancer susceptibility genes, for example through the CanGene-CanVar project [529], and could be extended for PRS.

Evaluating risk-stratified screening

The oversight of screening programmes in the UK is currently undergoing significant change. Following Professor Sir Mike Richards' 2019 recommendations, the National Screening Committee (NSC) now has oversight of population, targeted, and risk-stratified screening approaches (targeted screening previously having been under the remit of the National Institute for Health and Care Excellence (NICE)). The new NSC remit states that they will ensure that screening recommendations are embedded in a robust ethical framework and reduce inequalities [159].

Though there are parallels with monogenic testing, the ethical, legal and social implications (ELSI) that PRS raise pose additional challenges, which will vary depending on the approach taken to testing. Monogenic testing is a diagnostic test, whilst PRS are prognostic. Adding PRS to population screening expands screenings

remit from cancer prevention and early diagnosis to include prognostication, which has wider implications for personal health [519]. The implications for family members of an individual with high polygenic risk are unclear, but lower than for a monogenic condition, and will vary by trait [520, 528]. Where PRS are used in combination with non-genetic risk predictors to define risk then the implications for family (who may share some environmental exposures) are even less clear.

The ethical implications of PRS-based screening (and of any personalised screening) can be considered according to the four principles of medical ethics: autonomy, beneficence, non-maleficence, and justice. Upholding individual autonomy is commonly felt to be supported through informed consent or decision making. There is an argument that more refined risk information may improve this - "empowerment" is a concept often used when discussing genetic information and autonomy [528]. Upholding autonomy might also require screening participants to be able to opt-out of knowing their risk [520].

Another commonly cited motivation for PRS use is that risk information could stimulate behaviour change, facilitating primary prevention. The evidence for this in the context of PRS limited. A literature review in 2016 (including 18 studies) suggested there was little evidence to support this, though study quality was generally low [530]. A later meta-analysis of direct-to-consumer testing indicated some positive effect on behaviours [531]. The evidence is probably strongest for cardiovascular disease, where disclosure of PRS based risk led to increased information seeking and positive health behaviours [513]. PRS could also be used to select individuals for chemoprevention, such as regular aspirin to lower CRC risk.

Purported harms of population genetic testing include anxiety and stress. Whilst research looking at delivering monogenic risk information to individuals generally shows little evidence of negative psychosocial impact [528, 530, 532], a few studies do point to harm [513]. PRS generally deal with more common traits, though for those at the extreme of the risk distribution, risk may be equivalent to monogenic conditions. In the context of risk-stratified screening more broadly, given the amount

of data held by or transferred to screening programmes, legal and ethical issues of data storage and confidentiality need to be considered.

A further potential harm of risk-stratified screening is the risk of over-diagnosis - a greater issue in some other cancer types, particularly prostate and breast, than necessarily in colorectal cancer. This is an issue particularly where risk is used to identify who to screen, rather than modulating screening modality or age range [533]. Many risk prediction models proposed for use in screening programmes have been developed to predict incident cancers, rather than cancer mortality, and so will over-represent more indolent cancers unless the PRS also predicts mortality [533]. Studies in prostate cancer demonstrated that addition of PRS to PSA did not improve risk stratification for lethal prostate cancer [534].

The potential for genetic discrimination is also commonly raised in relation to genetic testing. This led to the Genetic Information Non-discrimination Act (GINA) in the USA in 2008, which protects against discrimination for health and employment, but does not cover other forms of insurance, including life insurance. In the UK, the open-ended Code on Genetic Testing and Insurance protects individuals' predictive genetic test result from affecting their insurance [535], though in risk-stratified screening, the screening pathway could feasibly be used by insurers as a proxy [520].

A key issue in PRS implementation is of poor risk prediction, and thus inequality in care provision, across demographic groups within the population (that is, distributive justice). Most GWAS to date have been performed in populations of European ancestry, with identification of SNPs and effect sizes that do not necessarily transfer to other populations. Martin et al. [474] explored the transferability of eight PRS constructed from published GWAS summary statistics to underrepresented populations. They found discrepancies in the direction of effects for all of the scores in new populations, and showed that genetic drift biases scores from European GWAS when used in other populations. The direction of bias was unpredictable.

Thus, the Eurocentric bias in GWAS inevitably leads to poor performance when used in other population of non-European ancestries. These populations are also under-represented in large datasets such as UKB, which further impedes

development and assessment of PRS in diverse populations. Minority populations already have higher CRC mortality and lower uptake of screening [445, 536, 537], although there have been some improvements in these discrepancies in recent years. Implementation of PRS in their current form has the potential to exacerbate current inequalities in healthcare, as individuals of European descent, already the most privileged in global healthcare, will derive greatest benefit [538]. The Inverse Equity Hypothesis shows that when new interventions are implemented, inequality tends to widen [539]. Access to genomic testing in cancer treatment is poorer for Black and Hispanic populations, and new technologies tend to reach socioeconomically poorer communities later [540]. A multitude of barriers to access for underprivileged groups means that PRS use is more likely to benefit wealthier white populations, both within and between countries.

Of note, the under-representation of non-white groups is pervasive not only in GWAS datasets and PRS studies, but in vast swathes of genetic research (and indeed medical research more broadly). The current Human Reference Genome is also based on individuals of European ancestry, and indeed is largely based on the sequence of one individual. Efforts are underway to improve representation here - for example the Human Pangenome Project aims to develop a reference panel representative of diverse populations [541].

Options for dealing with these discrepancies in PRS performance in the near term, if PRS were felt to have clinical utility in European populations, could include implementing them only in ancestrally European populations, implementing for all but with caveats given to the results in non-European groups, or choosing not to implement them at all given the risk of widening the health gap. None of these options seems to be ideal in the context of population screening, but the choice needs to be informed by the views of those adversely affected [528].

A further relatively unexplored issue is the poorer predictive performance of PRS in women, noted in my own study and in others. Possible genetic explanations for this included differing genetic architecture of traits by sex, sex-chromosome effects, and gene-environment interactions [542]. Generally, SNP-based studies

support sex-linked variability in heritability for only a limited number of phenotypes [543]. There do not appear to be differences in frequency of common alleles between sexes [544], but effect size could differ by sex, with differing loci contributing to the same total heritability [542, 545]. GWAS studies are often sex-agnostic (and were underpowered for sex-specific analysis), however with increased sample size sex-stratified GWAS, and sex-specific PRS, might help to improve on this.

An increasingly highlighted and important additional ethical consideration is the potential impact of our current healthcare choices on the environment. In the UK, healthcare contributes to 4% of greenhouse gas emissions (10% in the USA) and the NHS has committed to reaching NetZero by 2045 [47, 546]. Current healthcare emissions contribute to global heating which is already affecting population health, disproportionately in countries which already have relatively poor access to healthcare. Whilst this results in intra-generational injustice, environmental impacts will inevitably affect the health of future generations, leading to concerns around inter-generational justice.

Future research directions

Considerable resources have recently been allocated to evaluating PRS-driven healthcare. The UK Research Innovation (UKRI) Industrial Strategy Challenge Fund (ISCF) has invested £79 million (with further funding from industrial partners) in a programme to accelerate disease detection, led by the new "Our Future Health" research organisation. Up to 5 million individuals will donate biological samples, and permit linking of health-related data, with one of their explicit aims to "improve risk prediction, early detection and early intervention". The study will use genetic data and other health information to calculate risk scores for a range of diseases and evaluate how these risk scores might be put into practice. There is no clear standardisation currently as to how risk scores should be reported or interpreted by clinicians. Studies such as Our Future Health and the eMERGE network in the US will contribute to this, and to evaluating broader ELSI considerations [490, 547].

We have likely reached the ceiling of European GWAS sample size, the latest GWAS representing the combination of most internationally available datasets for colorectal cancer, and so the power to detect new variants or improve on effect size estimation is limited [232, 489]. Within PRS research, expansion of GWAS to under-represented populations will likely provide the next step-change in PRS performance.

Sharing of ancestry-specific GWAS summary statistics will help with this, as will the development of new multi-ancestry methodologies [513]. Expanding datasets and studies in non-white populations, for example through the work of the H3Africa consortium, will improve the power of risk prediction in under-represented populations [548]. PRS-CSx, a multi-ancestry extension of Bayesian PRS software PRS-CS, improves prediction across diverse populations, though the extent of this improvement varies across traits, and considerable performance gap persists between European and non-European populations [549]. Incorporating functionally-informed fine mapping also improved trait prediction across different populations [550]. Improving prediction in recently admixed populations, for whom appropriate LD reference panels are generally unavailable, also remains a challenge [551].

Both Our Future Health and eMERGE initiatives have committed to diversifying PRS, as did the UK government in the broader context of genomics its Genome UK report [511]. Our Future Health state that their resource will be "truly representative" of the UK population, ensuring inclusion of minority ethnic groups and individuals with low incomes. Notably, as evidenced in my own research, absolute numbers (rather than representative proportions) of minority groups are needed to permit well-powered studies within these groups, especially for lower incidence conditions. Although the Our Future Health dataset will be ten times the size of UKB, efforts must be made to ensure that the resource serves these communities sufficiently.

Beyond this, the sex-based differences in model performance discussed above also need to be explored, though this disparity is notably far smaller than for non-European populations. We need to ensure that PRS-based models perform well across diverse populations, and to minimise the potential to exacerbate existing inequalities.

There are many risk-prediction models for CRC, heterogeneous in design and, given the varied approaches taken to bowel screening internationally, also in context [552]. Relatively few of these have been externally validated, and this, rather than development of more models, should be a priority. However, there are several areas of potential additional interest. Additional biomarkers, such as the faecal microbiome, may further improve risk prediction in cancer screening [553]. Expansion of models to integrate longitudinal screening data such as prior FIT results, or adenoma detection in prior screening rounds, is likely to also improve predictive performance.

With increasingly large datasets, there is inevitable interest in "big data" machine learning approaches to risk prediction. Particular benefits of these approaches are the greater ability to analyse and model non-linear effects. Critics of machine learning note that these approaches are opaque, resource intensive, and dependent on the quality of the data training the models. A recent review of supervised machine learning prediction models across healthcare found them to be generally of poor quality and at high risk of bias [554]. Thomas et al. [228] used machine learning approaches (elastic net, ridge, and lasso penalized regression, and gradient boosted decision trees) to train PRS for colorectal cancer and found them inferior to LDpred. Min et al. [555] found no difference in the performance of deep neural networks compared to logistic regression models for CRC risk on external validation.

The National Screening Committee notes that a screening test,

> *"should be a simple test that has evidence of suitable accuracy and technical performance derived from studies in the population in which the test is being used."*

There is a need now to focus on more extensive evaluation of risk stratification in clinical practice with prospective trials and, in the context of cancer screening, of evaluating the impact on cancer mortality and quality of life, rather than test metrics such as detection rates.

Risk-stratified approaches to screening must also be shown to be cost-effective, and this is to date relatively understudied. In a recent systematic review of the literature on the cost-effectiveness of PRS-based population cancer screening, just

three studies of CRC screening were identified, and only ten studies were found across all cancer types [556].

The three bowel screening studies do not support PRS as a cost-effective approach. One study was undertaken in the context of the English BCSP [557]. This compared biennial FIT beginning at an age determined by risk stratification at 40 years of age, with biennial FIT at a standard age for all, using an individual patient microsimulation model, Microsimulation Model in Cancer of the Bowel (MiMiC-Bowel). They modelled participants from 30 years old with a lifetime horizon, in the context of NHS care. Notably, they did not assign costs to the use of the PRS, and instead undertook a justifiable costs analysis. They compared several risk prediction models, the best of which had an AUROC of 0.72.

Using a FIT threshold of 120 ug/g, and screening from 60 years of age, they found that risk-based screening would produce 418 QALYs, and save 218 cancer cases, and 156 cancer deaths, at a cost of £247,000 per 100,000 individuals with a similar number of FIT invitees. They estimated the net monetary benefit of this approach to be £8.1 million per 100,000 individuals. Up to £114 could be spent per person and still be cost effective. Lower benefit was obtained when comparing to FIT from 50 years of age, with net monetary benefit of £2-£46, and justifiable costs of £4-£65. The risk stratified approaches also had lesser benefit for women, as observed in my own study.

Neither of the other bowel screening studies found risk stratified approaches to be cost-effective, compared to colonoscopy based screening in the US (based on a PRS with an AUROC of 0.6, and price per PRS of $200) [558], or annual FIT (assuming costs of $240 for risk determination) [559].

There are a number of limitations to the cost-effectiveness analyses published to date [556]. The modelled performance of PRS are derived from cohort studies and do not reflect performance in a screening population. No paper described how the genetic data would be ascertained at scale, or how services might need to be redesigned to accommodate implementation. Costing of the PRS process was basic in all studies, nor did they conduct any sensitivity analyses around different

ways of modelling the PRS. No paper considered differing performance of PRS by ancestry, assuming instead that the European-based PRS used would apply to the whole population with equal efficacy; cost-effectiveness is highly likely to vary between individuals and populations as a result of this.

Beyond evaluating cost-effectiveness, it is likely over the coming decade that assessments of the environmental impact of new tests and pathways will be required prior to implementation, in line with NHS and government commitments to reduce carbon emissions [47]. It would seem logical that a PRS-based approach would have a higher footprint than using non-genetic data, but this could potentially be balanced by carbon savings of fewer colonoscopies, or more cancers prevented, when the whole screening pathway is considered. Carbon footprinting of healthcare, evaluation of broader effects such as water use and biodiversity impact, and explorations of the ways in which environmental harms are weighed against health, economic, and social harms and benefits, will be an important area of future research.

Conclusion

My thesis demonstrates that whilst PRS may give an incremental improvement in risk prediction over non-genetic risk models, these gains are small. At present, the significant ethical, social, and logistical implications of implementing PRS-based stratified screening almost certainly outweigh these benefits. This may change as issues such as increasing diversity are addressed (though this is likely to take some time), as the use of our genetic data becomes more widely employed in healthcare, and as costs reduce further.

Risk-stratification in some form is likely to have a role in cancer screening in the near future, though historically the speed of change within UK screening programmes has been slow. Although faecal immunochemical testing was proposed for consideration by the UK NSC in 2003, it was not recommended by NSC until 2013, and finally implemented in England in 2019. Commitments have recently been made to improve horizon-scanning and research links within screening programmes which will accelerate progress [159]. It remains to be seen whether

the evidence-based cautions against PRS implementation are sufficient to stall their implementation, given the enthusiasm and commitment to genomics already expressed by politicians and policymakers.

Appendices

A

SNPs inclusion list for 97-SNP PRS

Table A.1: SNPs included in 97-SNP PRS modelling from the normal distribution, described in Chapter 5

Locus	rsID	Risk Allele	RAF	OR	Corrected OR
1p32.3	rs12143541	G	0.153	1.10	1.098559
1q25.3	rs4546885	G	0.591	1.09	1.087321
1p34.3	rs61776719	C	0.445	1.07	1.068527
1q41	rs6658977	T	0.369	1.08	1.077774
1p36.12	rs72647484	T	0.908	1.13	1.127217
2q11.2	rs11692435	G	0.901	1.12	1.118929
2q33.1	rs11893063	A	0.467	1.07	1.069339
2q35	rs13020391	C	0.630	1.09	1.087549
2q24.2	rs448513	C	0.334	1.05	1.049724
2q33.1	rs7593422	T	0.547	1.07	1.068319
3q22.2	rs10049390	A	0.739	1.06	1.059638
3q13.2	rs12635946	C	0.623	1.08	1.077770
3q26.2	rs35446936	G	0.757	1.07	1.069513
3p22.1	rs35470271	G	0.156	1.09	1.088715
3p21.1	rs9831861	G	0.588	1.07	1.068735
4q24	rs17035289	T	0.830	1.10	1.097994
4q31.21	rs75686861	A	0.095	1.12	1.118373
5p13.1	rs1445011	C	0.284	1.11	1.106041
5p15.33	rs2735940	A	0.502	1.16	1.156871
5q31.1	rs639933	C	0.381	1.07	1.069238
5p15.33	rs77776598	C	0.063	1.16	1.156871
6p21.2	rs1321310	C	0.238	1.09	1.087951
6p21.31	rs16878812	A	0.887	1.11	1.108467
6p24.1	rs2070699	T	0.475	1.07	1.069130
6p21.33	rs3131043	G	0.431	1.07	1.069558
6p12.1	rs62404966	C	0.755	1.08	1.078983
6q21	rs6928864	C	0.909	1.13	1.128883
6p21.1	rs6933790	T	0.830	1.10	1.098513
6p21.32	rs9271770	A	0.807	1.08	1.079526

Table A.1: SNPs included in 97-SNP PRS modelling from the normal distribution, described in Chapter 5 *(continued)*

Locus	rsID	Risk Allele	RAF	OR	Corrected OR
7p12.3	rs10951878	C	0.485	1.06	1.059457
7p13	rs12672022	T	0.831	1.07	1.069604
7p12.3	rs3801081	G	0.675	1.08	1.077833
8q23.3	rs16892766	C	0.089	1.26	1.253448
8q24.21	rs6983267	G	0.529	1.11	1.108225
9q31.3	rs10980628	C	0.212	1.07	1.069752
9p21.3	rs1412834	T	0.498	1.08	1.077768
9q22.33	rs34405347	T	0.908	1.09	1.089747
10q25.2	rs12255141	G	0.102	1.11	1.107922
10q24.2	rs2193352	G	0.194	1.11	1.106575
10q22.3	rs704017	G	0.596	1.10	1.096633
10p14	rs7894531	G	0.686	1.13	1.126435
11q22.1	rs2186607	T	0.511	1.05	1.049927
11q23.1	rs3087967	T	0.300	1.15	1.146394
11q13.4	rs3824999	G	0.508	1.28	1.270318
11q12.2	rs4246215	G	0.641	1.06	1.059647
11p15.4	rs4450168	C	0.172	1.10	1.099187
11q13.4	rs4944940	G	0.960	1.28	1.270318
11q13.4	rs57796856	T	0.505	1.09	1.087253
12p13.31	rs10849438	G	0.122	1.12	1.116986
12q13.12	rs11169572	C	0.404	1.09	1.087524
12p13.32	rs12818766	A	0.180	1.10	1.098255
12p13.32	rs3217810	T	0.137	1.10	1.098255
12q24.12	rs597808	G	0.522	1.09	1.087250
12q24.21	rs7315438	T	0.590	1.08	1.077798
12q13.3	rs7398375	C	0.723	1.09	1.088815
13q13.3	rs12427600	C	0.242	1.09	1.087899
13q22.3	rs1330889	C	0.870	1.11	1.108279
13q22.1	rs45597035	A	0.639	1.08	1.078286
13q34	rs7993934	T	0.646	1.08	1.077973
13q13.2	rs9537521	G	0.646	1.08	1.079188
14q22.2	rs1570405	G	0.310	1.07	1.069653
14q23.1	rs17094983	G	0.882	1.09	1.089636
14q22.2	rs35107139	C	0.403	1.09	1.087528
14q22.2	rs4444235	C	0.465	1.07	1.069653
15q23	rs10152518	G	0.191	1.08	1.079396
15q13.3	rs16959063	A	0.013	1.33	1.325863
15q13.3	rs16969681	T	0.089	1.21	1.204484
15q13.3	rs17816465	A	0.199	1.12	1.116322
15q22.31	rs4776316	A	0.732	1.08	1.079158
15q22.33	rs56324967	C	0.660	1.07	1.069624
15q13.3	rs73376930	G	0.209	1.33	1.325863
15q26.1	rs7495132	T	0.117	1.11	1.108316
16q24.1	rs2696839	G	0.514	1.09	1.088003
16q23.2	rs61336918	A	0.294	1.09	1.087498
16q24.1	rs899244	T	0.214	1.09	1.088003
16q22.1	rs9939049	A	0.715	1.06	1.059631
17p12	rs1078643	A	0.768	1.09	1.087865
17p13.3	rs73975588	A	0.871	1.10	1.099004
17q25.3	rs75954926	G	0.660	1.09	1.089400

Table A.1: SNPs included in 97-SNP PRS modelling from the normal distribution, described in Chapter 5 *(continued)*

Locus	rsID	Risk Allele	RAF	OR	Corrected OR
17q24.3	rs983318	A	0.248	1.06	1.059645
18q21.1	rs7226855	A	0.541	1.21	1.206655
19q13.33	rs12979278	T	0.535	1.07	1.068893
19p13.11	rs285245	T	0.110	1.11	1.109115
19q13.11	rs73039434	T	0.953	1.30	1.289962
19q13.43	rs73068325	T	0.184	1.07	1.069670
19q13.2	rs9797885	G	0.708	1.08	1.078439
20q13.33	rs1741640	C	0.775	1.16	1.155932
20q13.13	rs1810502	C	0.561	1.08	1.078694
20q13.12	rs2179593	A	0.721	1.07	1.069153
20q13.33	rs3787089	C	0.322	1.07	1.069498
20q13.13	rs4811050	A	0.181	1.09	1.087600
20p12.3	rs6055286	A	0.146	1.11	1.107128
20q13.13	rs6066825	A	0.653	1.10	1.097015
20p12.3	rs6085661	T	0.388	1.09	1.087478
20q13.13	rs6091213	C	0.260	1.08	1.078694
20p12.3	rs961253	A	0.360	1.09	1.087478
Xp22.2	rs2732875	C	0.200	1.18	1.174658

B

Model specifications

Logistic regression PRS models

LDpred2-inf

$$\text{Prob}\{\text{pheno} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = \; & \\ & -9.776021 + 1.021227\,\text{PRS} \\ & -0.4011441[\text{Female}] + 0.08623154\,\text{age} \\ & +0.05306746[\text{UKBL}] - 0.01143121\,\text{PC}_1 - 0.0004654591\,\text{PC}_2 \\ & -0.00674316\,\text{PC}_3 - 0.01210423\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise. Following internal validation,

$$
\begin{aligned}
X\hat{\beta} = \\
&-9.7554910720 + 1.0176460857\ \text{PRS} \\
&-0.3997374587[\text{Female}] + 0.0859291598\ \text{age} \\
&+0.0528813798[\text{UKBL}] - 0.0113911291\ \text{PC}_1 - 0.0004638269\ \text{PC}_2 \\
&-0.0067195150\ \text{PC}_3 - 0.0120617855\ \text{PC}_4
\end{aligned}
$$

LDpred2-grid

$$\text{Prob}\{\text{pheno} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

$$
\begin{aligned}
X\hat{\beta} = \\
&-9.244175 + 1.04631\ \text{PRS} \\
&-0.4042047[\text{Female}] + 0.08638471\ \text{age} \\
&+0.05274771[\text{UKBL}] - 0.01191899\ \text{PC}_1 - 0.000458451\ \text{PC}_2 \\
&-0.007074939\ \text{PC}_3 - 0.009255497\ \text{PC}_4
\end{aligned}
$$

and $[c] = 1$ if subject is in group c, 0 otherwise. Following internal validation,

$$
\begin{aligned}
X\hat{\beta} = \\
&-9.2275069333 + 1.0430568922\ \text{PRS} \\
&-0.4029480042[\text{Female}] + 0.0861161216\ \text{age} \\
&+0.0525837127[\text{UKBL}] - 0.0118819318\ \text{PC}_1 - 0.0004570255\ \text{PC}_2 \\
&-0.0070529425\ \text{PC}_3 - 0.0092267211\ \text{PC}_4
\end{aligned}
$$

LDpred-grid-sp

$$\text{Prob}\{\text{pheno} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = & \\ & -9.323984 + 0.985573\,\text{PRS} \\ & -0.4036001[\text{Female}] + 0.08635244\,\text{age} \\ & +0.05231831[\text{UKBL}] - 0.01171954\,\text{PC}_1 - 0.0004695418\,\text{PC}_2 \\ & -0.007014255\,\text{PC}_3 - 0.01011505\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise. Following internal validation,

$$\begin{aligned} X\hat{\beta} = & \\ & -9.3105253498 + 0.9831336820\,\text{PRS} \\ & -0.4026011918[\text{Female}] + 0.0861387053\,\text{age} \\ & +0.0521888184[\text{UKBL}] - 0.0116905372\,\text{PC}_1 - 0.0004683795\,\text{PC}_2 \\ & -0.0069968945\,\text{PC}_3 - 0.0100900178\,\text{PC}_4 \end{aligned}$$

SCT

$$\text{Prob}\{\text{pheno} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = & \\ & -4.163311 + 1.21179\,\text{PRS} \\ & -0.4047076[\text{Female}] + 0.08630363\,\text{age} \\ & +0.05792523[\text{UKBL}] - 0.0123641\,\text{PC}_1 + 0.0001318505\,\text{PC}_2 \\ & -0.006999591\,\text{PC}_3 - 0.002657923\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise. Following internal validation,

$$
\begin{aligned}
X\hat{\beta} = \\
&-4.1624637504 + 1.2074970414\,\text{PRS} \\
&-0.4032739068[\text{Female}] + 0.0859978959\,\text{age} \\
&+0.0577200345[\text{UKBL}] - 0.0123203013\,\text{PC}_1 - 0.0001313836\,\text{PC}_2 \\
&-0.0069747950\,\text{PC}_3 - 0.0026485069\,\text{PC}_4
\end{aligned}
$$

C+T

$$\text{Prob}\{\text{pheno} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

$$
\begin{aligned}
X\hat{\beta} = \\
&-9.068643 + 0.07412662\,\text{PRS} \\
&-0.4029091[\text{Female}] + 0.08625292\,\text{age} \\
&+0.05362289[\text{UKBL}] - 0.01242532\,\text{PC}_1 + 1.63068\times10^{-5}\text{PC}_2 \\
&-0.008724144\,\text{PC}_3 - 0.006594728\,\text{PC}_4
\end{aligned}
$$

and $[c] = 1$ if subject is in group c, 0 otherwise. Following internal validation,

$$
\begin{aligned}
X\hat{\beta} = \\
&-9.04511500 + 0.07378766\,\text{PRS} \\
&-0.4010668[\text{Female}] + 0.08585851\,\text{age} \\
&+0.05337770[\text{UKBL}] - 0.0236850\,\text{PC}_1 + 1.623235\times10^{-5}\text{PC}_2 \\
&-0.008684252\,\text{PC}_3 - 0.006564573\,\text{PC}_4
\end{aligned}
$$

B. Model specifications

GWAS-sig

$$\text{Prob}\{\text{pheno} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = \\ & -13.0741 + 0.7126475\,\text{PRS} \\ & -0.4052543[\text{Female}] + 0.08610265\,\text{age} \\ & +0.05611813[\text{UKBL}] - 0.01428634\,\text{PC}_1 - 0.001213305\,\text{PC}_2 \\ & -0.009382505\,\text{PC}_3 + 0.008451566\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise.

following internal validation,

$$\begin{aligned} X\hat{\beta} = \\ & -13.063331950 + 0.711808118\,\text{PRS} \\ & -0.404777025[\text{Female}] + 0.086001227\,\text{age} \\ & +0.056052028[\text{UKBL}] - 0.014269518\,\text{PC}_1 - 0.001211876\,\text{PC}_2 \\ & -0.009371454\,\text{PC}_3 + 0.008441612\,\text{PC}_4 \end{aligned}$$

Cox PRS models

LDpred2-inf

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = \\ &-4.837914 + 0.8874252\,\text{PRS} \\ &-0.4574084[\text{Female}] + 0.07944168\,\text{age} \\ &+0.07447489[\text{UKBL}] - 0.01183015\,\text{PC}_1 - 0.0002170615\,\text{PC}_2 \\ &-0.0066799\,\text{PC}_3 - 0.007724787\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise

t	$S_0(t)$
0	1.0000000
1	0.9993397
2	0.9985263
3	0.9976918
4	0.9968184
5	0.9958934
6	0.9949866
7	0.9940408
8	0.9930979

Following internal validation, $X\hat{\beta}$ was multiplied by a global correction factor of 0.9920289, and $S_0(t)$ reestimated

t	$S_0(t)$
0	1.0000000
1	0.9993371
2	0.9985205
3	0.9976827
4	0.9968059
5	0.9958774
6	0.9949671
7	0.9940177
8	0.9930714

LDpred2-grid

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = & \\ & -4.319668 + 1.016773\,\text{PRS} \\ & -0.460077[\text{Female}] + 0.07956688\,\text{age} \\ & +0.07353158[\text{UKBL}] - 0.01182863\,\text{PC}_1 - 0.0002922827\,\text{PC}_2 \\ & -0.006630462\,\text{PC}_3 - 0.007400685\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise

t	$S_0(t)$
0	1.0000000
1	0.9993725
2	0.9985994
3	0.9978061
4	0.9969757
5	0.9960961
6	0.9952335
7	0.9943330
8	0.9934351

Following internal validation, $X\hat{\beta}$ was multiplied by a global correction factor of 0.9938484, and $S_0(t)$ reestimated

t	$S_0(t)$
0	1.0000000
1	0.9993702
2	0.9985943
3	0.9977981
4	0.9969646
5	0.9960819
6	0.9952162
7	0.9943126
8	0.9934116

LDpred2-grid-sp

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = & \\ & -4.401747 + 0.9494384\,\text{PRS} \\ & -0.4595375[\text{Female}] + 0.07953966\,\text{age} \\ & +0.07319098[\text{UKBL}] - 0.01163802\,\text{PC}_1 - 0.0002867862\,\text{PC}_2 \\ & -0.006627901\,\text{PC}_3 - 0.008062116\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise

t	$S_0(t)$
0	1.0000000
1	0.9993692
2	0.9985920
3	0.9977945
4	0.9969596
5	0.9960754
6	0.9952083
7	0.9943032
8	0.9934010

Following internal validation, $X\hat{\beta}$ was multiplied by a global correction factor of 0.9953671, and $S_0(t)$ reestimated

t	$S_0(t)$
0	1.0000000
1	0.9993675
2	0.9985882
3	0.9977885
4	0.9969514
5	0.9960649
6	0.9951954
7	0.9942880
8	0.9933834

B. Model specifications

SCT

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = \\ & 0.3031138 + 1.115351\,\text{PRS} \\ & -0.4605902[\text{Female}] + 0.07954427\,\text{age} \\ & +0.07795189[\text{UKBL}] - 0.01249099\,\text{PC}_1 + 0.0002968488\,\text{PC}_2 \\ & -0.006862566\,\text{PC}_3 - 0.0002282657\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise

t	$S_0(t)$
0	1.0000000
1	0.9993420
2	0.9985315
3	0.9976999
4	0.9968295
5	0.9959078
6	0.9950044
7	0.9940620
8	0.9931192

Following internal validation, $X\hat{\beta}$ was multiplied by a global correction factor of 0.9939586, and $S_0(t)$ reestimated

t	$S_0(t)$
0	1.0000000
1	0.9993400
2	0.9985270
3	0.9976930
4	0.9968200
5	0.9958956
6	0.9949894
7	0.9940444
8	0.9930989

B. Model specifications

C+T

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \quad \text{where}$$

$$X\hat{\beta} =$$

$$-4.20198 + 0.07001151\,\text{PRS}$$

$$-0.4586389[\text{Female}] + 0.07948435\,\text{age}$$

$$+0.07422172[\text{UKBL}] - 0.01244809\,\text{PC}_1 + 0.0002402616\,\text{PC}_2$$

$$-0.008344321\,\text{PC}_3 - 0.004250053\,\text{PC}_4$$

and $[c] = 1$ if subject is in group c, 0 otherwise

t	$S_0(t)$
0	1.0000000
1	0.9993447
2	0.9985374
3	0.9977091
4	0.9968422
5	0.9959242
6	0.9950241
7	0.9940850
8	0.9931462

Following internal validation, $X\hat{\beta}$ was multiplied by a global correction factor of 0.9916302, and $S_0(t)$ reestimated

t	$S_0(t)$
0	1.0000000
1	0.9993419
2	0.9985312
3	0.9976994
4	0.9968289
5	0.9959070
6	0.9950032
7	0.9940603
8	0.9931177

B. Model specifications

GWAS-sig

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = & \\ & -8.095997 + 0.6928628\,\text{PRS} \\ & -0.4610349[\text{Female}] + 0.07936184\,\text{age} \\ & +0.07622241[\text{UKBL}] - 0.01417734\,\text{PC}_1 - 0.0009608898\,\text{PC}_2 \\ & -0.008870042\,\text{PC}_3 + 0.009949456\,\text{PC}_4 \end{aligned}$$

and $[c] = 1$ if subject is in group c, 0 otherwise %latex.default(s, file = file, append = TRUE, rowlabel = "t", rowlabel.just = "r", dec = dec, table.env = FALSE)%

t	$S_0(t)$
0	1.0000000
1	0.9993421
2	0.9985317
3	0.9977003
4	0.9968301
5	0.9959087
6	0.9950054
7	0.9940622
8	0.9931147

Following internal validation, $X\hat{\beta}$ was multiplied by a global correction factor of 0.9918869, and $S_0(t)$ reestimated

t	$S_0(t)$
0	1.0000000
1	0.9993394
2	0.9985257
3	0.9976910
4	0.9968173
5	0.9958922
6	0.9949854
7	0.9940385
8	0.9930874

C

Plots of Schoenfeld Residuals for Cox models

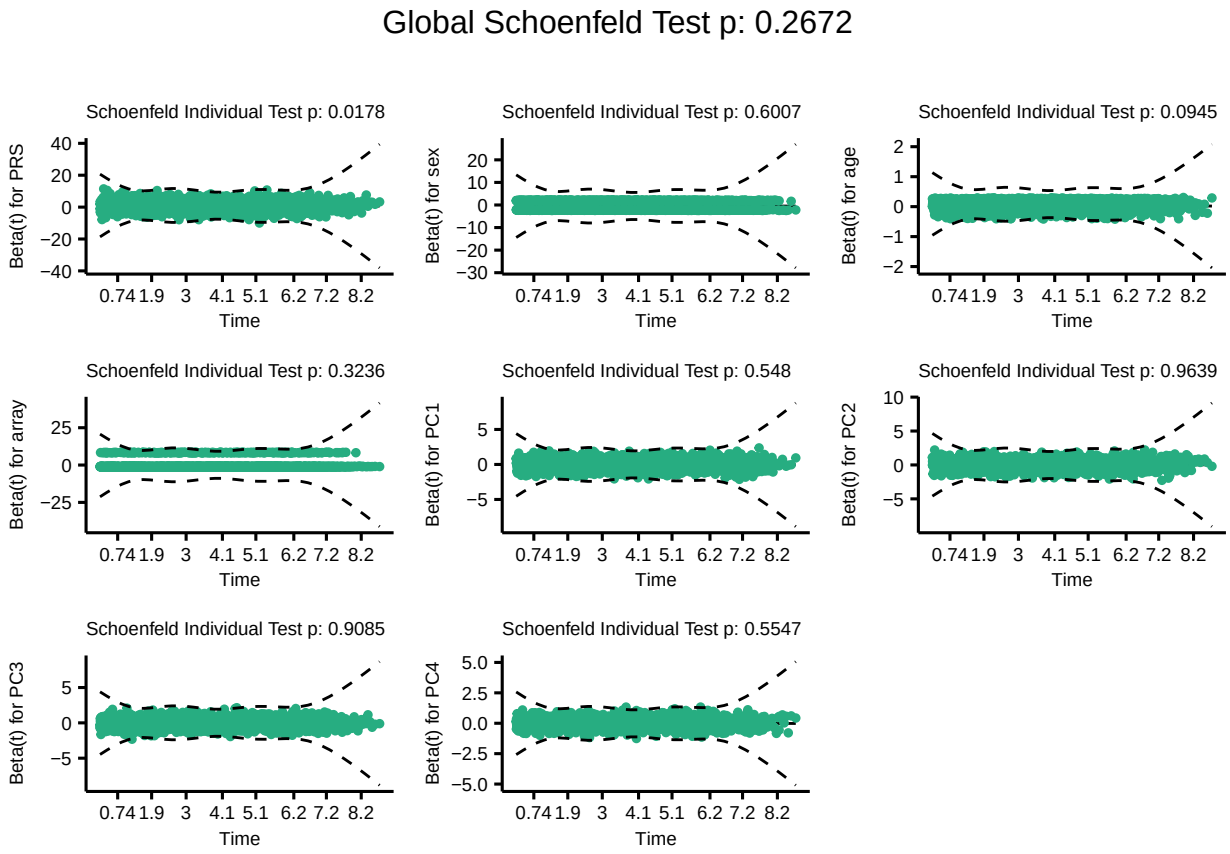

Figure C.1: Plots of Schoenfeld residuals for LDpred2-inf Cox model

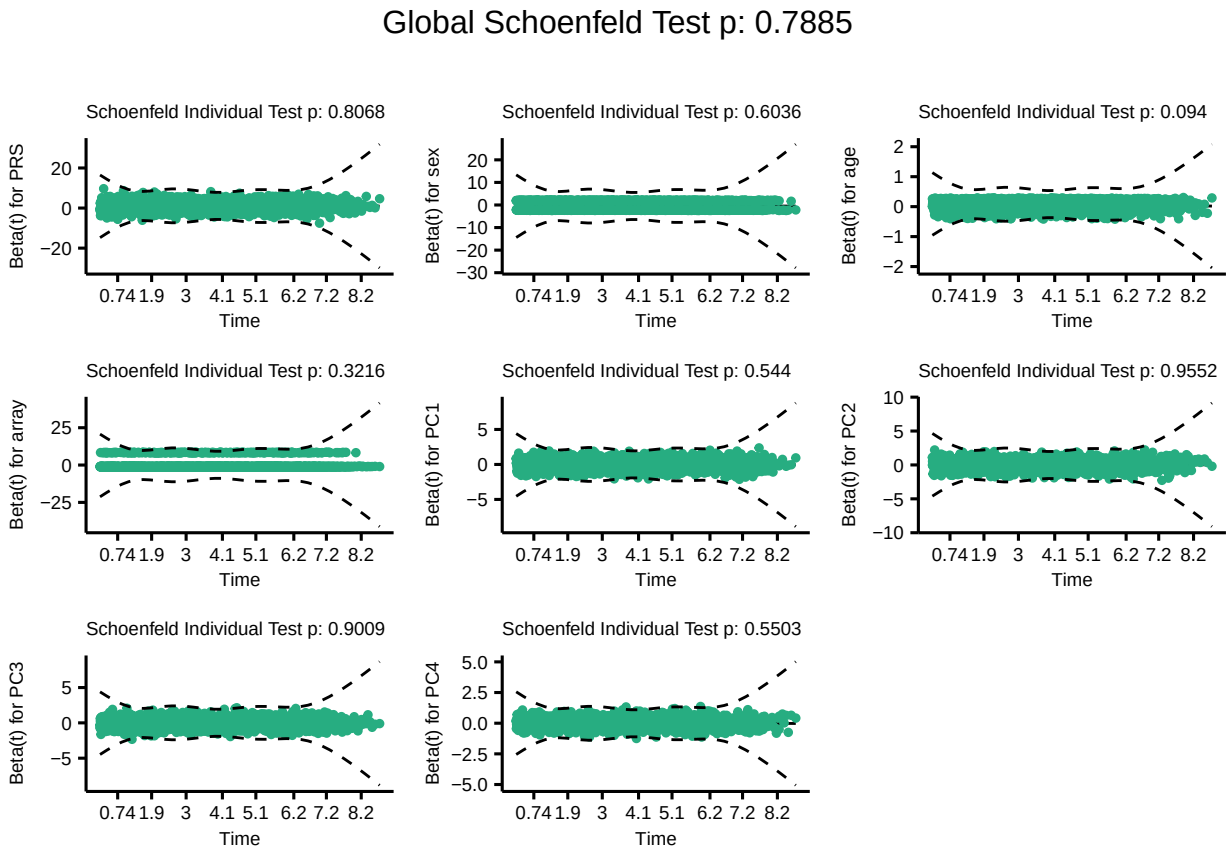

Figure C.2: Plots of Schoenfeld residuals for LDpred2-grid Cox model

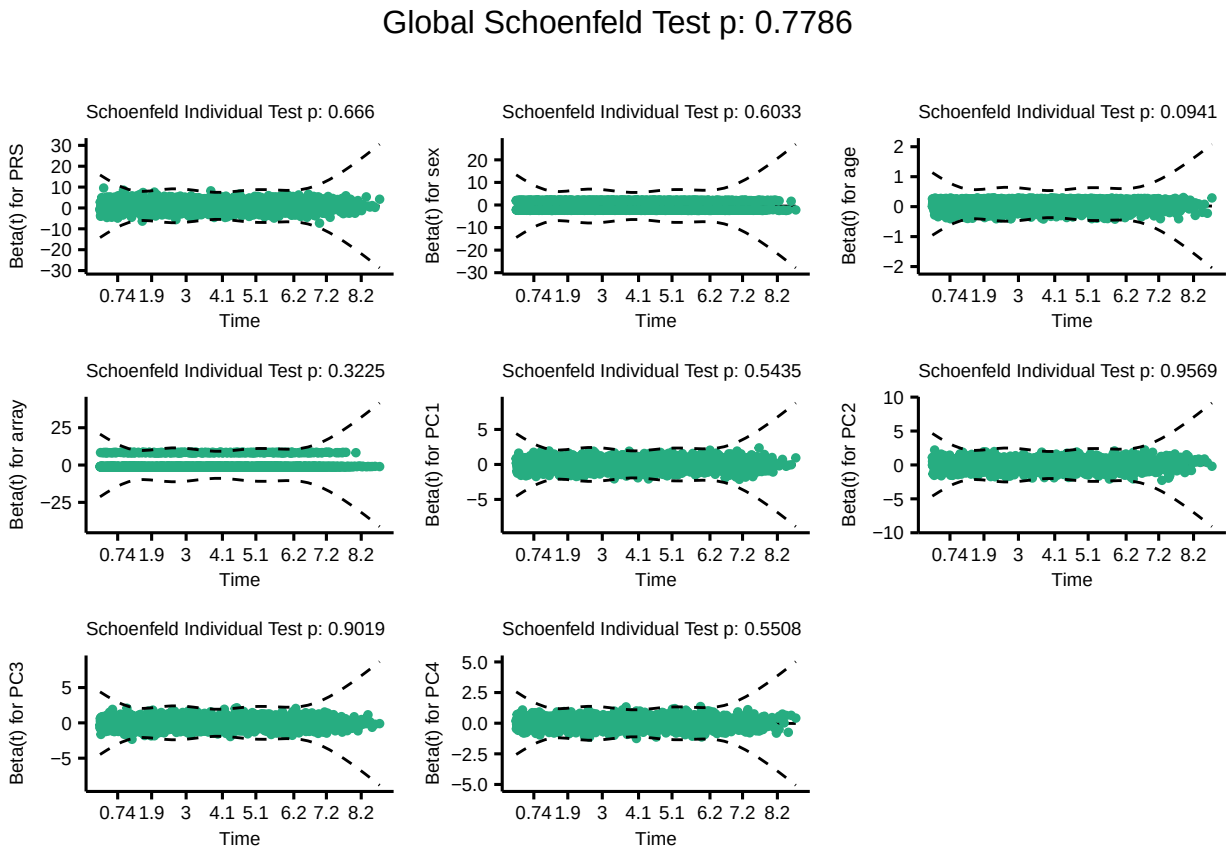

Figure C.3: Plots of Schoenfeld residuals for LDpred2-grid-sp Cox model

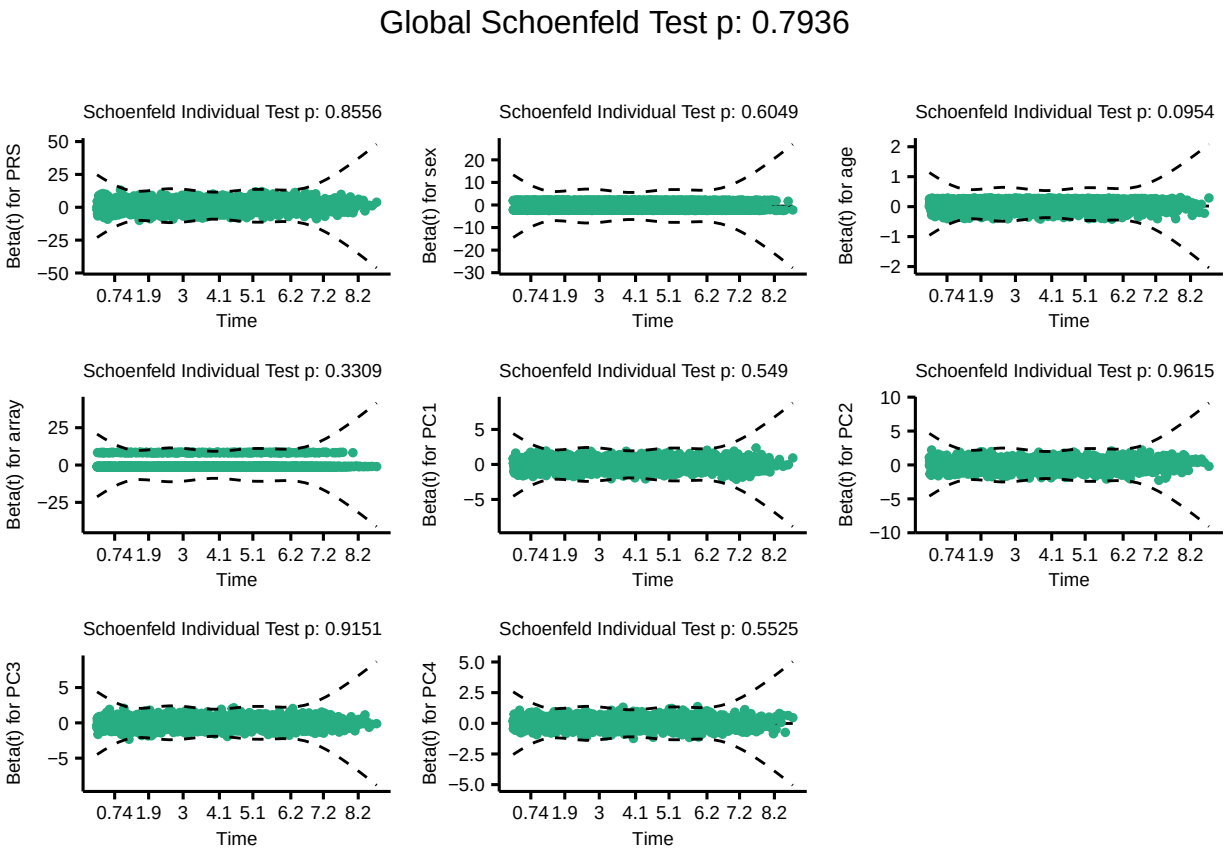

Figure C.4: Plots of Schoenfeld residuals for SCT Cox model

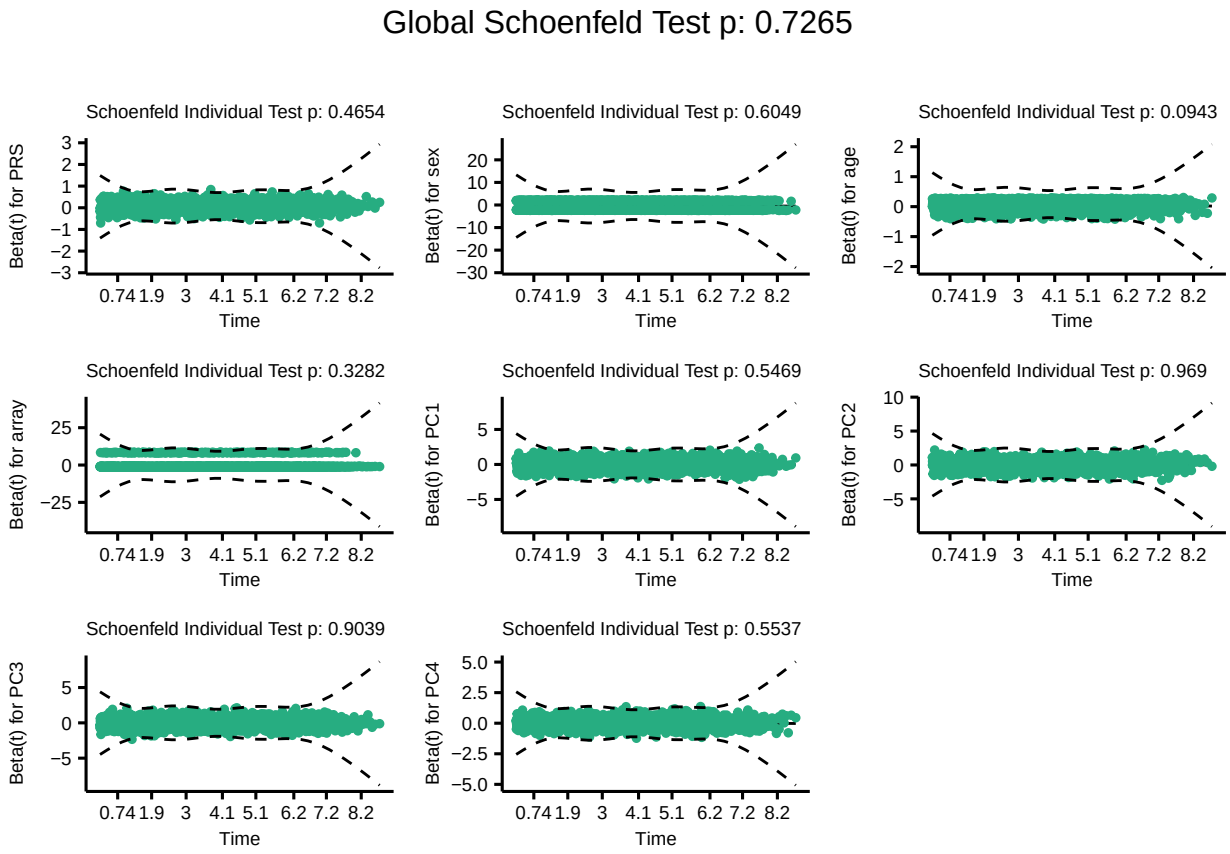

Figure C.5: Plots of Schoenfeld residuals for C+T Cox model

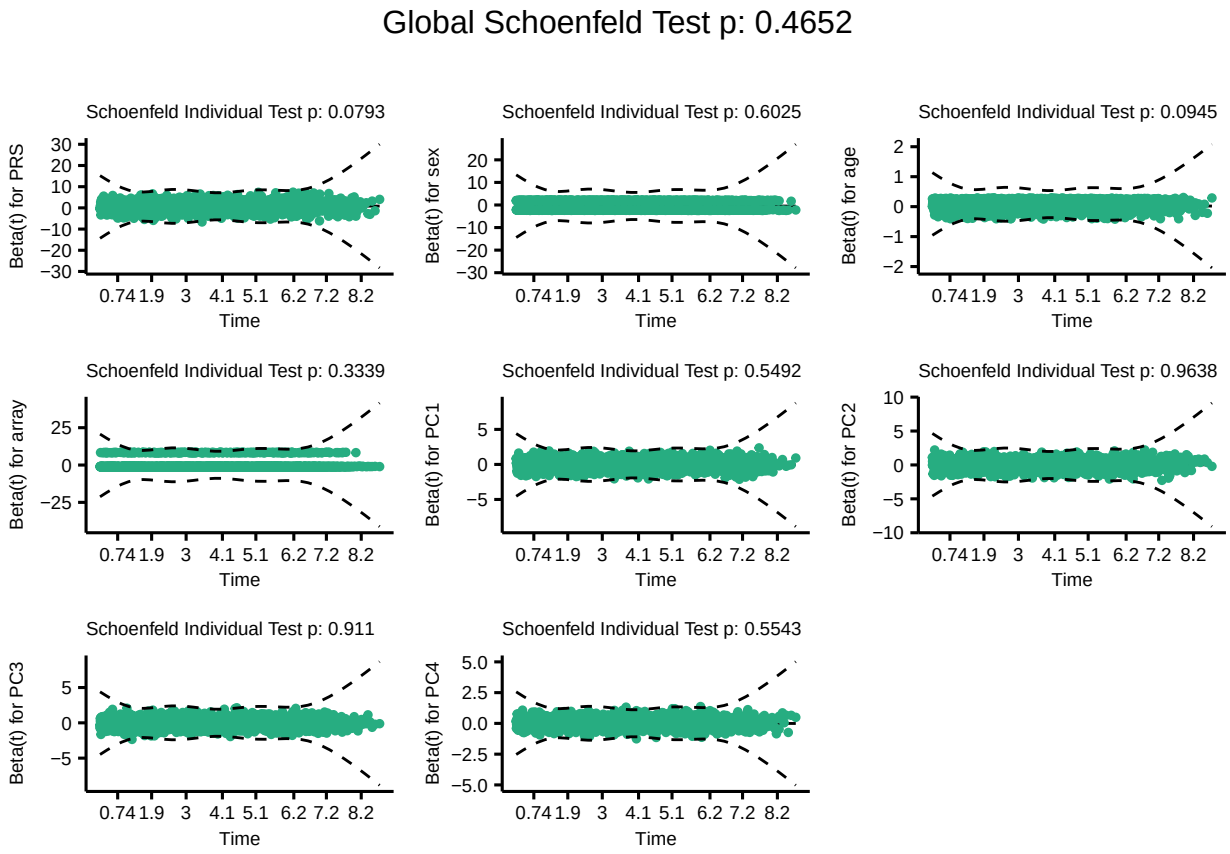

Figure C.6: Plots of Schoenfeld residuals for GWAS-sig Cox model

References

[1] Cancer Research UK. Bowel cancer statistics. 2022. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer, [Accessed 08-07-2022].

[2] GBD 2017 Colorectal Cancer Collaborators. The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. *Lancet Gastroenterology & Hepatology* 4.12 (2019), 913–933.

[3] N. Keum and E. Giovannucci. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology & Hepatology* 16.12 (2019), 713–732.

[4] R. L. Siegel, L. A. Torre, I. Soerjomataram, R. B. Hayes, F. Bray, T. K. Weber, and A. Jemal. Global patterns and trends in colorectal cancer incidence in young adults. *Gut* 68.12 (2019), 2179–2185.

[5] C. A. Doubeni. Early-onset colorectal cancer: what reported statistics can and cannot tell us and their implications. *Cancer* 125.21 (2019), 3706–3708.

[6] J. Lockhart-Mummery and C. Dukes. The precancerous changes in the rectum and colon. *Surgery, Gynecology & Obstetrics* 36 (1928), 591–596.

[7] R. J. Jackman and C. W. Mayo. The adenoma-carcinoma sequence in cancer of the colon. *Surgery, Gynecology & Obstetrics* 93.3 (1951), 327–30.

[8] S. J. Stryker, B. G. Wolff, C. E. Culp, S. D. Libbe, D. M. Ilstrup, and R. L. MacCarty. Natural history of untreated colonic polyps. *Gastroenterology* 93.5 (1987), 1009–13.

[9] I. Tomlinson. The mendelian colorectal cancer syndromes. *Annals of Clinical Biochemistry* 52.Pt 6 (2015), 690–2.

[10] B. Vogelstein, E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, Y. Nakamura, R. White, A. M. Smits, and J. L. Bos. Genetic alterations during colorectal-tumor development. *New England Journal of Medicine* 319.9 (1988), 525–32.

[11] E. R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell* 61.5 (1990), 759–67.

[12] M. Schmitt and F. R. Greten. The inflammatory pathogenesis of colorectal cancer. *Nature Reviews Immunology* 21.10 (2021), 653–667.

[13] G. Poulogiannis, K. Ichimura, R. A. Hamoudi, F. Luo, S. Y. Leung, S. T. Yuen, D. J. Harrison, A. H. Wyllie, and M. J. Arends. Prognostic relevance of dna copy number changes in colorectal cancer. *Journal of Pathology* 220.3 (2010), 338–47.

[14] D. J. Weisenberger, K. D. Siegmund, M. Campan, J. Young, T. I. Long, M. A. Faasse, G. H. Kang, M. Widschwendter, D. Weener, D. Buchanan, H. Koh, L. Simms, M. Barker, B. Leggett, J. Levine, M. Kim, A. J. French, S. N. Thibodeau, J. Jass, R. Haile, and P. W. Laird. Cpg island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with braf mutation in colorectal cancer. *Nature Genetics* 38.7 (2006), 787–93.

[15] J. Bogaert and H. Prenen. Molecular genetics of colorectal cancer. *Annals of Gastroenterology* 27.1 (2014), 9–14.

[16] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487.7407 (2012), 330–7.

[17] J. Guinney, R. Dienstmann, X. Wang, A. de Reynies, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, B. M. Bot, J. S. Morris, I. M. Simon, S. Gerster, E. Fessler, E. M. F. De Sousa, E. Missiaglia, H. Ramay, D. Barras, K. Homicsko, D. Maru, G. C. Manyam, B. Broom, V. Boige, B. Perez-Villamil, T. Laderas, R. Salazar, J. W. Gray, D. Hanahan, J. Tabernero, R. Bernards, S. H. Friend, P. Laurent-Puig, J. P. Medema, A. Sadanandam, L. Wessels, M. Delorenzi, S. Kopetz, L. Vermeulen, and S. Tejpar. The consensus molecular subtypes of colorectal cancer. *Nature Medicine* 21.11 (2015), 1350–6.

[18] E. Domingo, L. Freeman-Mills, E. Rayner, M. Glaire, S. Briggs, L. Vermeulen, E. Fessler, J. P. Medema, A. Boot, H. Morreau, T. van Wezel, G. J. Liefers, R. A. Lothe, S. A. Danielsen, A. Sveen, A. Nesbakken, I. Zlobec, A. Lugli, V. H. Koelzer, M. D. Berger, S. Castellvi-Bel, J. Munoz, c. Epicolon, M. de Bruyn, H. W. Nijman, M. Novelli, K. Lawson, D. Oukrif, E. Frangou, P. Dutton, S. Tejpar, M. Delorenzi, R. Kerr, D. Kerr, I. Tomlinson, and D. N. Church. Somatic pole proofreading domain mutation, immune response, and prognosis in colorectal cancer: a retrospective, pooled biomarker study. *Lancet Gastroenterology & Hepatology* 1.3 (2016), 207–216.

[19] J. Wilson and J. Jungner. Principles and practice of screening for disease. *Public Health Paper Number 34. Geneva: WHO* (1968).

[20] H. Brenner, M. Hoffmeister, C. Stegmaier, G. Brenner, L. Altenhofen, and U. Haug. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. *Gut* 56.11 (2007), 1585–9.

[21] R. E. Hertz, M. R. Deddish, and E. Day. Value of periodic examinations in detecting cancer of the rectum and colon. *Postgraduate Medicine* 27.3 (1960), 290–294.

[22] V. A. Gilbertsen and J. M. Nelms. The prevention of invasive cancer of the rectum. *Cancer* 41.3 (1978), 1137–1139.

[23] S. J. Winawer. The history of colorectal cancer screening: a personal perspective. *Digestive Diseases and Sciences* 60.3 (2015), 596–608.

[24] J. D. Hardcastle, J. O. Chamberlain, M. H. Robinson, S. M. Moss, S. S. Amar, T. W. Balfour, P. D. James, and C. M. Mangham. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *The Lancet* 348.9040 (1996), 1472–1477.

[25] J. S. Mandel, J. H. Bond, T. R. Church, D. C. Snover, G. M. Bradley, L. M. Schuman, and F. Ederer. Reducing mortality from colorectal cancer by screening for fecal occult blood. *New England Journal of Medicine* 328.19 (1993), 1365–1371. PMID: 8474513.

[26] O. Kronborg, C. Fenger, J. Olsen, O. D. Jørgensen, and O. Søndergaard. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *The Lancet* 348.9040 (1996), 1467–1471.

[27] A. Shaukat, S. J. Mongin, M. S. Geisser, F. A. Lederle, J. H. Bond, J. S. Mandel, and T. R. Church. Long-term mortality after screening for colorectal cancer. *New England Journal of Medicine* 369.12 (2013), 1106–1114. PMID: 24047060.

[28] J. E. Allison, I. S. Tekawa, L. J. Ransom, and A. L. Adrain. A comparison of fecal occult-blood tests for colorectal-cancer screening. *New England Journal of Medicine* 334.3 (1996), 155–160. PMID: 8531970.

[29] T. F. Imperiale, D. F. Ransohoff, S. H. Itzkowitz, B. A. Turnbull, and M. E. Ross. Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. *New England Journal of Medicine* 351.26 (2004), 2704–2714. PMID: 15616205.

[30] US Preventive Services Task Force. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *Journal of the American Medical Association* 325.19 (2021), 1965–1977.

[31] P. Hewitson, P. Glasziou, E. Watson, B. Towler, and L. Irwig. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. *American Journal of Gastroenterology* 103.6 (2008), 1541–9.

[32] T. Morikawa, J. Kato, Y. Yamaji, R. Wada, T. Mitsushima, and Y. Shiratori. A comparison of the immunochemical fecal occult blood test and total colonoscopy in the asymptomatic population. *Gastroenterology* 129.2 (2005), 422–8.

[33] L. Hol, M. E. van Leerdam, M. van Ballegooijen, A. J. van Vuuren, H. van Dekken, J. C. I. Y. Reijerink, A. C. M. van der Togt, J. D. F. Habbema, and E. J. Kuipers. Screening for colorectal cancer: randomised trial comparing guaiac-based and immunochemical faecal occult blood testing and flexible sigmoidoscopy. *Gut* 59.01 (2010), 62–68.

[34] H. Brenner and S. Tao. Superior diagnostic performance of faecal immunochemical tests for haemoglobin in a head-to-head comparison with guaiac based faecal occult blood test among 2235 participants of screening colonoscopy. *European Journal of Cancer* 49.14 (2013), 3049–54.

[35] T. F. Imperiale, D. F. Ransohoff, S. H. Itzkowitz, T. R. Levin, P. Lavin, G. P. Lidgard, D. A. Ahlquist, and B. M. Berger. Multitarget stool DNA testing for colorectal-cancer screening. *New England Journal of Medicine* 370.14 (2014), 1287–97.

[36] S. J. Winawer, A. G. Zauber, M. N. Ho, M. J. O'Brien, L. S. Gottlieb, S. S. Sternberg, J. D. Waye, M. Schapiro, J. H. Bond, J. F. Panish, and et al. Prevention of colorectal cancer by colonoscopic polypectomy. the national polyp study workgroup. *New England Journal of Medicine* 329.27 (1993), 1977–81.

[37] A. G. Zauber, S. J. Winawer, M. J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B. F. Hankey, W. Shi, J. H. Bond, M. Schapiro, J. F. Panish, E. T. Stewart, and J. D. Waye. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. *New England Journal of Medicine* 366.8 (2012), 687–96.

[38] M. Bretthauer, M. F. Kaminski, M. Løberg, A. G. Zauber, J. Regula, E. J. Kuipers, M. A. Hern'an, E. McFadden, A. Sunde, M. Kalager, E. Dekker, I. Lansdorp-Vogelaar, K. Garborg, M. Rupinski, M. C. W. Spaander, M. Bugajski, O. Høie, T. Stefansson, G. Hoff, H.-O. Adami, and for the Nordic-European Initiative on Colorectal Cancer (NordICC) Study Group. Population-Based Colonoscopy Screening for Colorectal Cancer: A Randomized Clinical Trial. *JAMA Internal Medicine* 176.7 (2016), 894–902.

[39] E. Quintero, A. Castells, L. Bujanda, J. Cubiella, D. Salas, A. Lanas, M. Andreu, F. Carballo, J. D. Morillas, C. Hernandez, R. Jover, I. Montalvo, J. Arenas, E. Laredo, V. Hernandez, F. Iglesias, E. Cid, R. Zubizarreta, T. Sala, M. Ponce, M. Andres, G. Teruel, A. Peris, M. P. Roncales, M. Polo-Tomas, X. Bessa, O. Ferrer-Armengou, J. Grau, A. Serradesanferm, A. Ono, J. Cruzado, F. Perez-Riquelme, I. Alonso-Abreu, M. de la Vega-Prieto, J. M. Reyes-Melian, G. Cacho, J. Diaz-Tasende, A. Herreros-de-Tejada, C. Poves, C. Santander, A. Gonzalez-Navarro, and C. S. Investigators. Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. *New England Journal of Medicine* 366.8 (2012), 697–706.

[40] E. H. Schreuders, A. Ruco, L. Rabeneck, R. E. Schoen, J. J. Sung, G. P. Young, and E. J. Kuipers. Colorectal cancer screening: a global overview of existing programmes. *Gut* 64.10 (2015), 1637–49.

[41] L. Kaalby, U. Deding, M. Kobaek-Larsen, A. V. Havshoi, E. Zimmermann-Nielsen, M. K. Thygesen, R. Kroeijer, T. Bjørsum-Meyer, and G. Baatrup. Colon capsule endoscopy in colorectal cancer screening: a randomised controlled trial. *BMJ Open Gastroenterology* 7.1 (2020), e000411.

[42] C. Spada, C. Hassan, B. Barbaro, F. Iafrate, P. Cesaro, L. Petruzziello, L. Minelli Grazioli, C. Senore, G. Brizi, I. Costamagna, G. Alvaro, M. Iannitti, M. Salsano, M. Ciolina, A. Laghi, L. Bonomo, and G. Costamagna. Colon capsule versus ct colonography in patients with incomplete colonoscopy: a prospective, comparative trial. *Gut* 64.2 (2015), 272–81.

[43] I. Lansdorp-Vogelaar, A. B. Knudsen, and H. Brenner. Cost-effectiveness of colorectal cancer screening. *Epidemiol Rev* 33 (2011), 88–100.

[44] M. McLeod, G. Kvizhinadze, M. Boyd, J. Barendregt, D. Sarfati, N. Wilson, and T. Blakely. Colorectal cancer screening: how health gains and cost-effectiveness vary by ethnic group, the impact on health inequalities, and the optimal age range to screen. *Cancer Epidemiology, Biomarkers & Prevention* 26.9 (2017), 1391–1400.

[45] A. G. Zauber. Cost-effectiveness of colonoscopy. *Gastrointest Endosc Clin N Am* 20.4 (2010), 751–70.

[46] G. M. Ginsberg, J. A. Lauer, S. Zelle, S. Baeten, and R. Baltussen. Cost effectiveness of strategies to combat breast, cervical, and colorectal cancer in sub-saharan africa and south east asia: mathematical modelling study. *British Medical Journal* 344 (2012), e614.

[47] NHS England and NHS Improvement. *Delivering a 'Net Zero' National Health Service*. Report. 2020.

[48] World Health Organisation. WHO and NHS to work together on decarbonization of health care systems across the world. 2022. `https://www.who.int/news-room/feature-stories/detail/who-and-nhs-to-work-together-on-decarbonization-of-health-care-systems-across-the-world`, [Accessed 15-09-2022].

[49] K. Siau, B. Hayee, and S. Gayam. Endoscopy's current carbon footprint. *Techniques and Innovations in Gastrointestinal Endoscopy* 23.4 (2021), 344–352.

[50] M. Sharma, S. Walpole, and K. Shah. Spotlight environmental sustainability: a strategic priority for nice. *Journal of Public Health* (2022).

[51] A. Miles, J. Cockburn, R. A. Smith, and J. Wardle. A perspective from countries using organized screening programs. *Cancer* 101.5 Suppl (2004), 1201–13.

[52] R. J. Steele, R. Parker, J. Patnick, J. Warner, C. Fraser, N. A. Mowat, J. Wilson, F. E. Alexander, J. G. Paterson, and United Kingdom Colorectal Screening Pilot Group. A demonstration pilot trial for colorectal cancer screening in the united kingdom: a new concept in the introduction of healthcare strategies. *Journal of Medical Screening* 8.4 (2001), 197–202.

[53] U.K. Colorectal Cancer Screening Pilot Group. Results of the first round of a demonstration pilot of screening for colorectal cancer in the united kingdom. *British Medical Journal* 329.7458 (2004), 133.

[54] R. F. A. Logan, J. Patnick, C. Nickerson, L. Coleman, M. D. Rutter, and o. von Wagner Christian. Outcomes of the bowel cancer screening programme (bcsp) in england after the first 1 million tests. *Gut* 61.10 (2012), 1439–1446.

[55] W. S. Atkin, R. Edwards, I. Kralj-Hans, K. Wooldrage, A. R. Hart, J. M. Northover, D. M. Parkin, J. Wardle, S. W. Duffy, J. Cuzick, and U. K. F. S. T. Investigators. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet* 375.9726 (2010), 1624–33.

[56] S. Moss, C. Mathews, T. J. Day, S. Smith, H. E. Seaman, J. Snowball, and S. P. Halloran. Increased uptake and improved outcomes of bowel cancer screening with a faecal immunochemical test: results from a pilot study within the national screening programme in england. *Gut* 66.9 (2017), 1631–1644.

[57] J. Murphy, S. Halloran, and A. Gray. Cost-effectiveness of the faecal immunochemical test at a range of positivity thresholds compared with the guaiac faecal occult blood test in the nhs bowel cancer screening programme in england. *BMJ Open* 7.10 (2017), e017186.

[58] NHS. Nhs long term plan. 2019. `https://www.longtermplan.nhs.uk/`, [Accessed December 2021].

[59] Bowel Cancer UK. Bowel scope screening is stopping in england. what is the future of bowel cancer screening? 2021. `https://www.bowelcanceruk.org.uk/news-and-blogs/research-blog/bowel-scope-screening-is-stopping-in-england/`, [Accessed 16-07-2022].

[60] Official Journal of the European Union. Council recommendation of 2 December 2003 on cancer screening. 2003. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:327:0034:0038:EN:PDF`, [Accessed 17-07-2022].

[61] Digestive Cancers Europe. White paper: colorectal screening in europe. 2019. `https://www.digestivecancers.eu/wp-content/uploads/2020/02/466-Document-DiCEWhitePaper2019.pdf`, [Accessed 17-07-2022].

[62] T. Byers, B. Levin, D. Rothenberger, G. D. Dodd, R. A. Smith, A. C. S. Detection, and T. A. G. on Colorectal Cancer). American cancer society guidelines for screening and surveillance for early detection of colorectal polyps and cancer: update 1997. *CA: a cancer journal for clinicians* 47.3 (1997), 154–160.

[63] S. G. Patel, F. P. May, J. C. Anderson, C. A. Burke, J. A. Dominitz, S. A. Gross, B. C. Jacobson, A. Shaukat, and D. J. Robertson. Updates on age to start and stop colorectal cancer screening: recommendations from the u.s. multi-society task force on colorectal cancer. *American Journal of Gastroenterology* 117.1 (2022), 57–69.

[64] R. Cardoso, F. Guo, T. Heisser, M. Hackl, P. Ihle, H. De Schutter, N. Van Damme, Z. Valerianova, T. Atanasov, O. Majek, J. Muzik, M. C. Nilbert, A. J. Tybjerg, K. Innos, M. Magi, N. Malila, A. M. Bouvier, V. Bouvier, G. Launoy, A. S. Woronoff, M. Cariou, M. Robaszkiewicz, P. Delafosse, F. Poncet, A. Katalinic, P. M. Walsh, C. Senore, S. Rosso, I. Vincerzevskiene, V. Lemmens, M. A. G. Elferink, T. B. Johannesen, H. Korner, F. Pfeffer, M. J. Bento, J. Rodrigues, F. Alves da Costa, A. Miranda, V. Zadnik, T. Zagar, A. Lopez de Munain Marques, R. Marcos-Gragera, M. Puigdemont, J. Galceran, M. Carulla, M. D. Chirlaque, M. Ballesta, K. Sundquist, J. Sundquist, M. Weber, A. Jordan, C. Herrmann, M. Mousavi, A. Ryzhov, M. Hoffmeister, and H. Brenner. Colorectal cancer incidence, mortality, and stage distribution in european countries in the colorectal cancer screening era: an international population-based study. *Lancet Oncol* 22.7 (2021), 1002–1013.

[65] K. J. Monahan, N. Bradshaw, S. Dolwani, B. Desouza, M. G. Dunlop, J. E. East, M. Ilyas, A. Kaur, F. Lalloo, A. Latchford, M. D. Rutter, I. Tomlinson, H. J. W. Thomas, and J. Hill. Guidelines for the management of hereditary colorectal cancer from the british society of gastroenterology (bsg)/association of coloproctology of great britain and ireland (acpgbi)/united kingdom cancer genetics group (ukcgg). *Gut* 69.3 (2020), 411–444.

[66] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from sweden, denmark, and finland. *New England Journal of Medicine* 343.2 (2000), 78–85.

[67] R. E. Graff, S. Moller, M. N. Passarelli, J. S. Witte, A. Skytthe, K. Christensen, Q. Tan, H. O. Adami, K. Czene, J. R. Harris, E. Pukkala, J. Kaprio, E. L. Giovannucci, L. A. Mucci, and J. B. Hjelmborg. Familial risk and heritability of colorectal cancer in the nordic twin study of cancer. *Clinical Gastroenterology and Hepatology* 15.8 (2017), 1256–1264.

[68] K. W. Jasperson, T. M. Tuohy, D. W. Neklason, and R. W. Burt. Hereditary and familial colon cancer. *Gastroenterology* 138.6 (2010), 2044–2058. Colon Cancer: An Update and Future Directions.

[69] M. B. Yurgelun, M. H. Kulke, C. S. Fuchs, B. A. Allen, H. Uno, J. L. Hornick, C. I. Ukaegbu, L. K. Brais, P. G. McNamara, R. J. Mayer, D. Schrag, J. A. Meyerhardt, K. Ng, J. Kidd, N. Singh, A.-R. Hartman, R. J. Wenstrup, and S. Syngal. Cancer susceptibility gene mutations in individuals with colorectal cancer. *Journal of Clinical Oncology* 35.10 (2017), 1086–1095. PMID: 28135145.

[70] E. M. Stoffel, P. B. Mangu, S. B. Gruber, S. R. Hamilton, M. F. Kalady, M. W. Y. Lau, K. H. Lu, N. Roach, and P. J. Limburg. Hereditary colorectal cancer syndromes: american society of clinical oncology clinical practice guideline endorsement of the familial-risk colorectal cancer: european society for medical oncology clinical practice guidelines. *Journal of Clinical Oncology* 33.2 (2015), 209–217.

[71] M. Daca Alvarez, I. Quintana, M. Terradas, P. Mur, F. Balaguer, and L. Valle. The inherited and familial component of early-onset colorectal cancer. *Cells* 10.3 (2021).

[72] M. S. DeRycke, S. Gunawardena, J. R. Balcom, A. M. Pickart, L. A. Waltman, A. J. French, S. McDonnell, S. M. Riska, Z. C. Fogarty, M. C. Larson, S. Middha, B. W. Eckloff, Y. W. Asmann, M. J. Ferber, R. W. Haile, S. Gallinger, M. Clendenning, C. Rosty, A. K. Win, D. D. Buchanan, J. L. Hopper, P. A. Newcomb, L. Le Marchand, E. L. Goode, N. M. Lindor, and S. N. Thibodeau. Targeted sequencing of 36 known or putative colorectal cancer susceptibility genes. *Molecular Genetics & Genomic Medicine* 5.5 (2017), 553–569.

[73] E. M. Stoffel, E. Koeppe, J. Everett, P. Ulintz, M. Kiel, J. Osborne, L. Williams, K. Hanson, S. B. Gruber, and L. S. Rozek. Germline genetic features of young individuals with colorectal cancer. *Gastroenterology* 154.4 (2018), 897–905.e1.

[74] W. F. Bodmer, C. J. Bailey, J. Bodmer, H. J. Bussey, A. Ellis, P. Gorman, F. C. Lucibello, V. A. Murday, S. H. Rider, P. Scambler, D. Sheer, E. Solomon, and N. K. Spurr. Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature* 328.6131 (1987), 614–6.

[75] P. Peltomåki, L. A. Aaltonen, P. Sistonen, L. Pylkkånen, J.-P. Mecklin, H. Järvinen, J. S. Green, J. R. Jass, J. L. Weber, F. S. Leach, G. M. Petersen, S. R. Hamilton, A. de la Chapelle, and B. Vogelstein. Genetic mapping of a locus predisposing to human colorectal cancer. *Science* 260.5109 (1993), 810–812.

[76] F. S. Leach, N. C. Nicolaides, N. Papadopoulos, B. Liu, J. Jen, R. Parsons, P. Peltomaki, P. Sistonen, L. A. Aaltonen, M. Nystrom-Lahti, and et al. Mutations of a muts homolog in hereditary nonpolyposis colorectal cancer. *Cell* 75.6 (1993), 1215–25.

[77] A. Lindblom, P. Tannergård, B. Werelius, and M. Nordenskjöld. Genetic mapping of a second locus predisposing to hereditary non–polyposis colon cancer. *Nature Genetics* 5.3 (1993), 279–282.

[78] N. Papadopoulos, N. C. Nicolaides, Y. F. Wei, S. M. Ruben, K. C. Carter, C. A. Rosen, W. A. Haseltine, R. D. Fleischmann, C. M. Fraser, M. D. Adams, and et al. Mutation of a mutl homolog in hereditary colon cancer. *Science* 263.5153 (1994), 1625–9.

[79] C. Turnbull, A. Sud, and R. S. Houlston. Cancer genetics, precision prevention and a call to action. *Nature Genetics* 50.9 (2018), 1212–1218.

[80] A. Hemminki, I. Tomlinson, D. Markie, H. Jarvinen, P. Sistonen, A. M. Bjorkqvist, S. Knuutila, R. Salovaara, W. Bodmer, D. Shibata, A. de la Chapelle, and L. A. Aaltonen. Localization of a susceptibility locus for peutz-jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis. *Nature Genetics* 15.1 (1997), 87–90.

[81] J. R. Howe, J. L. Bair, M. G. Sayed, M. E. Anderson, F. A. Mitros, G. M. Petersen, V. E. Velculescu, G. Traverso, and B. Vogelstein. Germline mutations of the gene encoding bone morphogenetic protein receptor 1a in juvenile polyposis. *Nature Genetics* 28.2 (2001), 184–7.

[82] E. D. Lynch, E. A. Ostermeyer, M. K. Lee, J. F. Arena, H. Ji, J. Dann, K. Swisshelm, D. Suchard, P. M. MacLeod, S. Kvinnsland, B. T. Gjertsen, K. Heimdal, H. Lubs, P. Moller, and M. C. King. Inherited mutations in pten that are associated with breast cancer, cowden disease, and juvenile polyposis. *American Journal of Human Genetics* 61.6 (1997), 1254–60.

[83] D. Liaw, D. J. Marsh, J. Li, P. L. Dahia, S. I. Wang, Z. Zheng, S. Bose, K. M. Call, H. C. Tsou, M. Peacocke, C. Eng, and R. Parsons. Germline mutations of the pten gene in cowden disease, an inherited breast and thyroid cancer syndrome. *Nature Genetics* 16.1 (1997), 64–7.

[84] E. E. Jaeger, K. L. Woodford-Richens, M. Lockett, A. J. Rowan, E. J. Sawyer, K. Heinimann, P. Rozen, V. A. Murday, S. C. Whitelaw, A. Ginsberg, W. S. Atkin, H. T. Lynch, M. C. Southey, H. Debinski, C. Eng, W. F. Bodmer, I. C. Talbot, S. V. Hodgson, H. J. Thomas, and I. P. Tomlinson. An ancestral ashkenazi haplotype at the hmps/crac1 locus on 15q13-q14 is associated with hereditary mixed polyposis syndrome. *American Journal of Human Genetics* 72.5 (2003), 1261–7.

[85] E. Jaeger, E. Webb, K. Howarth, L. Carvajal-Carmona, A. Rowan, P. Broderick, A. Walther, S. Spain, A. Pittman, Z. Kemp, K. Sullivan, K. Heinimann, S. Lubbe, E. Domingo, E. Barclay, L. Martin, M. Gorman, I. Chandler, J. Vijayakrishnan, W. Wood, E. Papaemmanuil, S. Penegar, M. Qureshi, members of the CORGI Consortium, S. Farrington, A. Tenesa, J.-B. Cazier, D. Kerr, R. Gray, J. Peto, M. Dunlop, H. Campbell, H. Thomas, R. Houlston, and I. Tomlinson. Common genetic variants at the crac1 (hmps) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature Genetics* 40.1 (2007), 26–8.

[86] C. Palles, J. B. Cazier, K. M. Howarth, E. Domingo, A. M. Jones, P. Broderick, Z. Kemp, S. L. Spain, E. Guarino, I. Salguero, A. Sherborne, D. Chubb, L. G. Carvajal-Carmona, Y. Ma, K. Kaur, S. Dobbins, E. Barclay, M. Gorman, L. Martin, M. B. Kovac, S. Humphray, C. Consortium, W. G. S. Consortium, A. Lucassen, C. C. Holmes, D. Bentley, P. Donnelly, J. Taylor, C. Petridis, R. Roylance, E. J. Sawyer, D. J. Kerr, S. Clark, J. Grimes, S. E. Kearsey, H. J. Thomas, G. McVean, R. S. Houlston, and I. Tomlinson. Germline mutations

affecting the proofreading domains of pole and pold1 predispose to colorectal adenomas and carcinomas. *Nature Genetics* 45.2 (2013), 136–44.

[87] N. Al-Tassan, N. H. Chmiel, J. Maynard, N. Fleming, A. L. Livingston, G. T. Williams, A. K. Hodges, D. R. Davies, S. S. David, J. R. Sampson, and J. P. Cheadle. Inherited variants of myh associated with somatic g:c–>t:a mutations in colorectal tumors. *Nature Genetics* 30.2 (2002), 227–32.

[88] M. Miyaki, M. Konishi, K. Tanaka, R. Kikuchi-Yanoshita, M. Muraoka, M. Yasuno, T. Igari, M. Koike, M. Chiba, and T. Mori. Germline mutation of msh6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature Genetics* 17.3 (1997), 271–272.

[89] P. Broderick, S. E. Dobbins, D. Chubb, B. Kinnersley, M. G. Dunlop, I. Tomlinson, and R. S. Houlston. Validation of recently proposed colorectal cancer susceptibility gene variants in an analysis of families and patients—a systematic review. *Gastroenterology* 152.1 (2017), 75–77.e4.

[90] M. Terradas, G. Capella, and L. Valle. Dominantly inherited hereditary nonpolyposis colorectal cancer not caused by mmr genes. *J Clin Med* 9.6 (2020).

[91] C. Palles, H. D. West, E. Chew, S. Galavotti, C. Flensburg, J. E. Grolleman, E. A. M. Jansen, H. Curley, L. Chegwidden, E. H. Arbe-Barnes, N. Lander, R. Truscott, J. Pagan, A. Bajel, K. Sherwood, L. Martin, H. Thomas, D. Georgiou, F. Fostira, Y. Goldberg, D. J. Adams, S. A. M. van der Biezen, M. Christie, M. Clendenning, L. E. Thomas, C. Deltas, A. J. Dimovski, D. Dymerska, J. Lubinski, K. Mahmood, R. S. van der Post, M. Sanders, J. Weitz, J. C. Taylor, C. Turnbull, L. Vreede, T. van Wezel, C. Whalley, C. Arnedo-Pac, G. Caravagna, W. Cross, D. Chubb, A. Frangou, A. J. Gruber, B. Kinnersley, B. Noyvert, D. Church, T. Graham, R. Houlston, N. Lopez-Bigas, A. Sottoriva, D. Wedge, C. Genomics England Research, C. Consortium, W. G. S. Consortium, M. A. Jenkins, R. P. Kuiper, A. W. Roberts, J. P. Cheadle, M. J. L. Ligtenberg, N. Hoogerbrugge, V. H. Koelzer, A. D. Rivas, I. M. Winship, C. R. Ponte, D. D. Buchanan, D. G. Power, A. Green, I. P. M. Tomlinson, J. R. Sampson, I. J. Majewski, and R. M. de Voer. Germline mbd4 deficiency causes a multi-tumor predisposition syndrome. *American Journal of Human Genetics* 109.5 (2022), 953–960.

[92] G. J. Mendel. Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn* 4 (1865), 3–47.

[93] W. Bateson, E. Saunders, and R. Punnett. Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society* 2 (1905), 1–55, 80–99.

[94] T. H. Morgan. Sex-limited inheritance in drosophila. *Science* 132 (1910), 120–122.

[95] F. A. Janssens. La theorie de la chiasmatypie. *Cellule* 25 (1909), 387–411.

[96] T. H. Morgan. Random segregation versus coupling in mendelian inheritance. *Science* 34.873 (1911), 384.

[97] I. Lobo and K. Shaw. Discovery and types of genetic linkage. *Nature Education* 1.1 (2008), 139.

[98] Z. Kemp, L. Carvajal-Carmona, S. Spain, E. Barclay, M. Gorman, L. Martin, E. Jaeger, N. Brooks, D. T. Bishop, H. Thomas, I. Tomlinson, E. Papaemmanuil, E. Webb, G. S. Sellick, W. Wood, G. Evans, A. Lucassen, E. R. Maher, R. S. Houlston, and C. ColoRectal tumour Gene Identification Study. Evidence for a colorectal cancer susceptibility locus on chromosome 3q21-q24 from a high-density snp genome-wide linkage scan. *Human Molecular Genetics* 15.19 (2006), 2903–10.

[99] J. Skoglund, T. Djureinovic, X. L. Zhou, J. Vandrovcova, E. Renkonen, L. Iselius, M. L. Bisgaard, P. Peltomaki, and A. Lindblom. Linkage analysis in a large swedish family supports the presence of a susceptibility locus for adenoma and colorectal cancer on chromosome 9q22.32-31.1. *J Med Genet* 43.2 (2006), e7.

[100] D. W. Neklason, R. A. Kerber, D. B. Nilson, H. Anton-Culver, A. G. Schwartz, C. A. Griffin, J. T. Lowery, J. M. Schildkraut, J. P. Evans, G. E. Tomlinson, L. C. Strong, A. R. Miller, J. E. Stopfer, D. M. Finkelstein, P. M. Nadkarni, C. H. Kasten, G. P. Mineau, and R. W. Burt. Common familial colorectal cancer linked to chromosome 7q31: a genome-wide analysis. *Cancer Research* 68.21 (2008), 8993–7.

[101] G. L. Wiesner, D. Daley, S. Lewis, C. Ticknor, P. Platzer, J. Lutterbaugh, M. MacMillen, B. Baliner, J. Willis, R. C. Elston, and S. D. Markowitz. A subset of familial colorectal neoplasia kindreds linked to chromosome 9q22.2-31.2. *Proceedings of the National Academy of Sciences USA* 100.22 (2003), 12961–5.

[102] S. Picelli, J. Vandrovcova, S. Jones, T. Djureinovic, J. Skoglund, X. L. Zhou, V. E. Velculescu, B. Vogelstein, and A. Lindblom. Genome-wide linkage scan for colorectal cancer susceptibility genes supports linkage to chromosome 3q. *BMC Cancer* 8 (2008), 87.

[103] C. Gray-McGuire, K. Guda, I. Adrianto, C. P. Lin, L. Natale, J. D. Potter, P. Newcomb, E. M. Poole, C. M. Ulrich, N. Lindor, E. L. Goode, B. L. Fridley, R. Jenkins, L. Le Marchand, G. Casey, R. Haile, J. Hopper, M. Jenkins, J. Young, D. Buchanan, S. Gallinger, M. Adams, S. Lewis, J. Willis, R. Elston, S. D. Markowitz, and G. L. Wiesner. Confirmation of linkage to and localization of familial colon cancer risk haplotype on chromosome 9q22. *Cancer Research* 70.13 (2010), 5409–18.

[104] A. Middeldorp, S. C. Jagmohan-Changur, H. M. van der Klift, M. van Puijenbroek, J. J. Houwing-Duistermaat, E. Webb, R. Houlston, C. Tops, H. F. Vasen, P. Devilee, H. Morreau, T. van Wezel, and J. Wijnen. Comprehensive genetic analysis of seven large families with mismatch repair proficient colorectal cancer. *Genes Chromosomes & Cancer* 49.6 (2010), 539–48.

[105] E. Sanchez-Tome, B. Rivera, J. Perea, G. Pita, D. Rueda, F. Mercadillo, A. Canal, A. Gonzalez-Neira, J. Benitez, and M. Urioste. Genome-wide linkage analysis and tumoral characterization reveal heterogeneity in familial colorectal cancer type x. *J Gastroenterol* 50.6 (2015), 657–66.

[106] M. S. Cicek, J. M. Cunningham, B. L. Fridley, D. J. Serie, W. R. Bamlet, B. Diergaarde, R. W. Haile, L. Le Marchand, T. G. Krontiris, H. B. Younghusband, S. Gallinger, P. A. Newcomb, J. L. Hopper, M. A. Jenkins, G. Casey, F. Schumacher, Z. Chen, M. S. DeRycke, A. S. Templeton, I. Winship, R. C. Green, J. S. Green, F. A. Macrae, S. Parry, G. P. Young, J. P. Young,

D. Buchanan, D. C. Thomas, D. T. Bishop, N. M. Lindor, S. N. Thibodeau, J. D. Potter, E. L. Goode, and C. F. R. Colon. Colorectal cancer linkage on chromosomes 4q21, 8q13, 12q24, and 15q22. *PLoS One* 7.5 (2012), e38175.

[107] I. W. Saunders, J. Ross, F. Macrae, G. P. Young, I. Blanco, J. Brohede, G. Brown, D. Brookes, T. Lockett, P. L. Molloy, V. Moreno, G. Capella, and G. N. Hannan. Evidence of linkage to chromosomes 10p15.3-p15.1, 14q24.3-q31.1 and 9q33.3-q34.3 in non-syndromic colorectal cancer families. *European Journal of Human Genetics* 20.1 (2012), 91–6.

[108] V. Kontham, S. von Holst, and A. Lindblom. Linkage analysis in familial non-lynch syndrome colorectal cancer families from sweden. *PLOS ONE* 8.12 (2013), e83936.

[109] S. von Holst, X. Jiao, W. Liu, V. Kontham, J. Thutkawkorapin, J. Ringdahl, P. Bryant, and A. Lindblom. Linkage analysis revealed risk loci on 6p21 and 18p11.2-q11.2 in familial colon and rectal cancer, respectively. *European Journal of Human Genetics* 27.8 (2019), 1286–1295.

[110] C. Toma, M. Diaz-Gay, S. Franch-Exposito, C. Arnau-Collell, B. Overs, J. Munoz, L. Bonjoch, Y. Soares de Lima, T. Ocana, M. Cuatrecasas, A. Castells, L. Bujanda, F. Balaguer, J. Cubiella, T. Caldes, J. M. Fullerton, and S. Castellvi-Bel. Using linkage studies combined with whole-exome sequencing to identify novel candidate genes for familial colorectal cancer. *International Journal of Cancer* 146.6 (2020), 1568–1577.

[111] T. Djureinovic, J. Skoglund, J. Vandrovcova, X.-L. Zhou, A. Kalushkova, L. Iselius, and A. Lindblom. A genome wide linkage analysis in swedish families with hereditary non-familial adenomatous polyposis/non-hereditary non-polyposis colorectal cancer. *Gut* 55.3 (2006), 362–366.

[112] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science* 273.5281 (1996), 1516–7.

[113] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Philos. Trans. R. Soc. Edinb.* 52 (1918), 399–433.

[114] C. Fernandez-Rozadilla, M. Timofeeva, Z. Chen, P. Law, M. Thomas, S. Schmit, V. Díez-Obrero, L. Hsu, J. Fernandez-Tajes, C. Palles, K. Sherwood, S. Briggs, V. Svinti, K. Donnelly, S. Farrington, J. Blackmur, P. Vaughan-Shaw, X.-o. Shu, J. Long, Q. Cai, X. Guo, Y. Lu, P. Broderick, J. Studd, J. Huyghe, T. Harrison, D. Conti, C. Dampier, M. Devall, F. Schumacher, M. Melas, G. Rennert, M. Obón-Santacana, V. Martín-Sánchez, F. Moratalla-Navarro, J. H. Oh, J. Kim, S. H. Jee, K. J. Jung, S.-S. Kweon, M.-H. Shin, A. Shin, Y.-O. Ahn, D.-H. Kim, I. Oze, W. Wen, K. Matsuo, K. Matsuda, C. Tanikawa, Z. Ren, Y.-T. Gao, W.-H. Jia, J. Hopper, M. Jenkins, A. K. Win, R. Pai, J. Figueiredo, R. Haile, S. Gallinger, M. Woods, P. Newcomb, D. Duggan, J. Cheadle, R. Kaplan, T. Maughan, R. Kerr, D. Kerr, I. Kirac, J. Böhm, L.-P. Mecklin, P. Jousilahti, P. Knekt, L. Aaltonen, H. Rissanen, E. Pukkala, J. Eriksson, T. Cajuso, U. Hänninen, J. Kondelin, K. Palin, T. Tanskanen, L. Renkonen-Sinisalo, B. Zanke, S. Männistö, D. Albanes, S. Weinstein, E. Ruiz-Narvaez, J. Palmer, D. Buchanan, E. Platz, K. Visvanathan, C. Ulrich, E. Siegel, S. Brezina, A. Gsur, P. Campbell, J. Chang-Claude, M. Hoffmeister, H. Brenner, M. Slattery, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204

cases and 154,587 controls of european and east asian ancestries. *Nature Genetics* 55.1 (2023), 89–99.

[115] P. J. Law, M. Timofeeva, C. Fernandez-Rozadilla, P. Broderick, J. Studd, J. Fernandez-Tajes, S. Farrington, V. Svinti, C. Palles, G. Orlando, A. Sud, A. Holroyd, S. Penegar, E. Theodoratou, P. Vaughan-Shaw, H. Campbell, L. Zgaga, C. Hayward, A. Campbell, S. Harris, I. J. Deary, J. Starr, L. Gatcombe, M. Pinna, S. Briggs, L. Martin, E. Jaeger, A. Sharma-Oates, J. East, S. Leedham, R. Arnold, E. Johnstone, H. Wang, D. Kerr, R. Kerr, T. Maughan, R. Kaplan, N. Al-Tassan, K. Palin, U. A. Hanninen, T. Cajuso, T. Tanskanen, J. Kondelin, E. Kaasinen, A. P. Sarin, J. G. Eriksson, H. Rissanen, P. Knekt, E. Pukkala, P. Jousilahti, V. Salomaa, S. Ripatti, A. Palotie, L. Renkonen-Sinisalo, A. Lepisto, J. Bohm, J. P. Mecklin, D. D. Buchanan, A. K. Win, J. Hopper, M. E. Jenkins, N. M. Lindor, P. A. Newcomb, S. Gallinger, D. Duggan, G. Casey, P. Hoffmann, M. M. Nothen, K. H. Jockel, D. F. Easton, P. D. P. Pharoah, J. Peto, F. Canzian, A. Swerdlow, R. A. Eeles, Z. Kote-Jarai, K. Muir, N. Pashayan, P. consortium, A. Harkin, K. Allan, J. McQueen, J. Paul, T. Iveson, M. Saunders, K. Butterbach, J. Chang-Claude, M. Hoffmeister, H. Brenner, I. Kirac, P. Matosevic, P. Hofer, S. Brezina, A. Gsur, J. P. Cheadle, L. A. Aaltonen, I. Tomlinson, R. S. Houlston, and M. G. Dunlop. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nature Communications* 10.1 (2019), 2154.

[116] F. Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLOS Genetics* 9.3 (2013), e1003348.

[117] J. McClellan and M. C. King. Genetic heterogeneity in human disease. *Cell* 141.2 (2010), 210–7.

[118] C. Fernández-Rozadilla, M. Álvarez-Barona, I. Quintana, A. López-Novo, J. Amigo, J. Cameselle-Teijeiro, E. Roman, D. González, X. Llor, L. Bujanda, X. Bessa, R. Jover, F. Balaguer, A. Castells, S. Castellvi-Bel, G. Capella, A. Carracedo, L. Valle, and C. Ruiz-Ponte. Exome sequencing of early-onset patients supports genetic heterogeneity in colorectal cancer. *Scientific Reports* 11.1 (2021), 1–9.

[119] F. Islami, A. Goding Sauer, K. D. Miller, R. L. Siegel, S. A. Fedewa, E. J. Jacobs, M. L. McCullough, A. V. Patel, J. Ma, I. Soerjomataram, W. D. Flanders, O. W. Brawley, S. M. Gapstur, and A. Jemal. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the united states. *CA: A Cancer Journal for Clinicians* 68.1 (2018), 31–54.

[120] J. DeCosse, S. Ngoi, J. Jacobson, and W. Cennerazzo. Gender and colorectal cancer. *European Journal of Cancer Prevention* (1993), 105–115.

[121] N. Papadimitriou, G. Markozannes, A. Kanellopoulou, E. Critselis, S. Alhardan, V. Karafousia, J. C. Kasimis, C. Katsaraki, A. Papadopoulou, M. Zografou, D. S. Lopez, D. S. M. Chan, M. Kyrgiou, E. Ntzani, A. J. Cross, M. T. Marrone, E. A. Platz, M. J. Gunter, and K. K. Tsilidis. An umbrella review of the evidence associating diet and cancer risk at 11 anatomical sites. *Nature Communications* 12.1 (2021), 4579.

[122] World Cancer Research Fund/American Institute for Cancer Research. Continuous update project expert report 2018. diet, nutrition, physical activity and colorectal cancer.
`https://www.wcrf.org/dietandcancer/colorectal-cancer`. 2018.

[123] B. C. Johnston, D. Zeraatkar, M. A. Han, R. W. Vernooij, C. Valli, R. El Dib, C. Marshall, P. J. Stover, S. Fairweather-Taitt, G. Wójcik, F. Bhatia, R. de Souza, C. Brotons, J. J. Meerpohl, C. J. Patel, B. Djulbegovic, P. Alonso-Coello, M. M. Bala, and G. H. Guyatt. Unprocessed Red Meat and Processed Meat Consumption: Dietary Guideline Recommendations From the Nutritional Recommendations (NutriRECS) Consortium. *Annals of Internal Medicine* 171.10 (2019), 756–764.

[124] V. Bouvard, D. Loomis, K. Z. Guyton, Y. Grosse, F. E. Ghissassi, L. Benbrahim-Tallaa, N. Guha, H. Mattock, and K. Straif. Carcinogenicity of consumption of red and processed meat. *The Lancet Oncology* 16.16 (2015), 1599–1600.

[125] A. R. Vieira, L. Abar, D. S. M. Chan, S. Vingeliene, E. Polemiti, C. Stevens, D. Greenwood, and T. Norat. Foods and beverages and colorectal cancer risk: a systematic review and meta-analysis of cohort studies, an update of the evidence of the wcrf-aicr continuous update project. *Annals of Oncology* 28.8 (2017), 1788–1802.

[126] L. Abar, A. R. Vieira, D. Aune, J. G. Sobiecki, S. Vingeliene, E. Polemiti, C. Stevens, D. C. Greenwood, D. S. M. Chan, S. Schlesinger, and T. Norat. Height and body fatness and colorectal cancer risk: an update of the wcrf-aicr systematic review of published prospective studies. *European Journal of Nutrition* 57.5 (2018), 1701–1720.

[127] N. K. Khankari, X.-O. Shu, W. Wen, P. Kraft, S. Lindström, U. Peters, J. Schildkraut, F. Schumacher, P. Bofetta, A. Risch, H. Bickeböller, C. I. Amos, D. Easton, R. A. Eeles, S. B. Gruber, C. A. Haiman, D. J. Hunter, S. J. Chanock, B. L. Pierce, W. Zheng, and on behalf of the Colorectal Transdisciplinary Study (CORECT) and Discovery Biology and Risk of Inherited Variants in Breast Cancer (DRIVE), Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE), and Transdisciplinary Research in Cancer of the Lung (TRICL). Association between adult height and risk of colorectal, lung, and prostate cancer: results from meta-analyses of prospective studies and mendelian randomization analyses. *PLOS Medicine* 13.9 (2016), e1002118.

[128] A. P. Thrift, J. Gong, U. Peters, J. Chang-Claude, A. Rudolph, M. L. Slattery, A. T. Chan, T. Esko, A. R. Wood, J. Yang, S. Vedantam, S. Gustafsson, T. H. Pers, G. Consortium, J. A. Baron, S. Bezieau, S. Küry, S. Ogino, S. I. Berndt, G. Casey, R. W. Haile, M. Du, T. A. Harrison, M. Thornquist, D. J. Duggan, L. Le Marchand, M. Lemire, N. M. Lindor, D. Seminara, M. Song, S. N. Thibodeau, M. Cotterchio, A. K. Win, M. A. Jenkins, J. L. Hopper, C. M. Ulrich, J. D. Potter, P. A. Newcomb, R. E. Schoen, M. Hoffmeister, H. Brenner, E. White, L. Hsu, and P. T. Campbell. Mendelian randomization study of height and risk of colorectal cancer. *International Journal of Epidemiology* 44.2 (2015), 662–672.

[129] C. Gao, C. J. Patel, K. Michailidou, U. Peters, J. Gong, J. Schildkraut, F. R. Schumacher, W. Zheng, P. Boffetta, I. Stucker, W. Willett, S. Gruber, D. F. Easton, D. J. Hunter, T. A. Sellers, C. Haiman, B. E. Henderson, R. J. Hung, C. Amos, B. L. Pierce, S. Lindström, P. Kraft, B. on behalf of: the Colorectal Transdisciplinary Study (CORECT); Discovery, R. of Inherited Variants in Breast Cancer (DRIVE); Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE); Follow-up of Ovarian Cancer Genetic Association, I. S. (FOCI); and T. R. in Cancer of the Lung (TRICL). Mendelian randomization study of adiposity-related traits and risk of breast, ovarian, prostate, lung and colorectal cancer. *International Journal of Epidemiology* 45.3 (2016), 896–908.

[130] C. J. Caspersen, K. E. Powell, and G. M. Christenson. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Reports* 100.2 (1985), 126–131.

[131] R. B. Gupta, N. Harpaz, S. Itzkowitz, S. Hossain, S. Matula, A. Kornbluth, C. Bodian, and T. Ullman. Histologic inflammation is a risk factor for progression to colorectal neoplasia in ulcerative colitis: a cohort study. *Gastroenterology* 133.4 (2007), 1099–1105.

[132] J. A. Eaden, K. R. Abrams, and J. F. Mayberry. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut* 48.4 (2001), 526–535.

[133] T. Jess, J. Simonsen, K. T. Jørgensen, B. V. Pedersen, N. M. Nielsen, and M. Frisch. Decreasing risk of colorectal cancer in patients with inflammatory bowel disease over 30 years. *Gastroenterology* 143.2 (2012), 375–381.e1.

[134] O. Olén, R. Erichsen, M. C. Sachs, L. Pedersen, J. Halfvarson, J. Askling, A. Ekbom, H. T. Sørensen, and J. F. Ludvigsson. Colorectal cancer in ulcerative colitis: a scandinavian population-based cohort study. *The Lancet* 395.10218 (2020), 123–131.

[135] P. L. Lakatos and L. Lakatos. Challenges in calculating the risk for colorectal cancer in patients with ulcerative colitis. *Clinical Gastroenterology and Hepatology* 10.10 (2012), 1179.

[136] S. Sebastian, H. V. Hernández, P. Myrelid, R. Kariv, E. Tsianos, M. Toruner, M. Marti-Gallostra, A. Spinelli, A. E. van der Meulen-de Jong, E. S. Yuksel, C. Gasche, S. Ardizzone, and S. Danese. Colorectal cancer in inflammatory bowel disease: Results of the 3rd ECCO pathogenesis scientific workshop (I). *Journal of Crohn's and Colitis* 8.1 (2014), 5–18.

[137] L. J. Herrinton. Regarding: a tale of two cohorts. *Gastroenterology* 144.3 (2013), e21–e22.

[138] D. Piovani, C. Hassan, A. Repici, L. Rimassa, C. Carlo-Stella, G. K. Nikolopoulos, E. Riboli, and S. Bonovas. Risk of cancer in inflammatory bowel diseases: umbrella review and reanalysis of meta-analyses. *Gastroenterology* (2022).

[139] M. G. Laukoetter, R. Mennigen, C. M. Hannig, N. Osada, E. Rijcken, T. Vowinkel, C. F. Krieglstein, N. Senninger, C. Anthoni, and M. Bruewer. Intestinal cancer risk in crohn's disease: a meta-analysis. *Journal of Gastrointestinal Surgery* 15.4 (2011), 576–583.

[140] C. Canavan, K. R. Abrams, and J. Mayberry. Meta-analysis: colorectal and small bowel cancer risk in patients with crohn's disease. *Alimentary Pharmacology & Therapeutics* 23.8 (2006), 1097–1104.

[141] A. C. von Roon, G. Reese, J. Teare, V. Constantinides, A. W. Darzi, and P. P. Tekkis. The risk of cancer in patients with crohn's disease. *Diseases of the Colon & Rectum* 50.6 (2007).

[142] P. M. Rothwell, M. Wilson, C.-E. Elwin, B. Norrving, A. Algra, C. P. Warlow, and T. W. Meade. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *The Lancet* 376.9754 (2010), 1741–1750.

[143] J. Chubak, E. P. Whitlock, S. B. Williams, A. Kamineni, B. U. Burda, D. S. Buist, and M. L. Anderson. Aspirin for the Prevention of Cancer Incidence and Mortality: Systematic Evidence Reviews for the U.S. Preventive Services Task ForceAspirin for the Prevention of Cancer Incidence and Mortality. *Annals of Internal Medicine* 164.12 (2016), 814–825.

[144] T. Tomić, S. Domínguez-López, and R. Barrios-Rodríguez. Non-aspirin non-steroidal anti-inflammatory drugs in prevention of colorectal cancer in people aged 40 or older: a systematic review and meta-analysis. *Cancer Epidemiology* 58 (2019), 52–62.

[145] C. W. Ng, A. A. Jiang, E. M. S. Toh, C. H. Ng, Z. H. Ong, S. Peng, H. Y. Tham, R. Sundar, C. S. Chong, and C. M. Khoo. Metformin and colorectal cancer: a systematic review, meta-analysis and meta-regression. *International Journal of Colorectal Disease* 35.8 (2020), 1501–1512.

[146] Y. S. Jung, C. H. Park, C. S. Eun, D. I. Park, and D. S. Han. Metformin use and the risk of colorectal adenoma: a systematic review and meta-analysis. *Journal of Gastroenterology and Hepatology* 32.5 (2017), 957–965.

[147] N. Thosani, S. N. Thosani, S. Kumar, Z. Nugent, C. Jimenez, H. Singh, and S. Guha. Reduced risk of colorectal cancer with use of oral bisphosphonates: a systematic review and meta-analysis. *Journal of Clinical Oncology* 31.5 (2013), 623–630. PMID: 23269990.

[148] Y. Y. Li, L. J. Gao, Y. X. Zhang, S. J. Liu, S. Cheng, Y. P. Liu, and C. X. Jia. Bisphosphonates and risk of cancers: a systematic review and meta-analysis. *British Journal of Cancer* 123.10 (2020), 1570–1581.

[149] Y. Liu, W. Tang, J. Wang, L. Xie, T. Li, Y. He, Y. Deng, Q. Peng, S. Li, and X. Qin. Association between statin use and colorectal cancer risk: a meta-analysis of 42 studies. *Cancer Causes & Control* 25.2 (2014), 237–249.

[150] Y. S. Jung, C. H. Park, C. S. Eun, D. I. Park, and D. S. Han. Statin use and the risk of colorectal adenoma: a meta-analysis. *Journal of Gastroenterology and Hepatology* 31.11 (2016), 1823–1830.

[151] H. Chu, J. Xin, Q. Yuan, Y. Wu, M. Du, R. Zheng, H. Liu, S. Wu, Z. Zhang, and M. Wang. A prospective study of the associations among fine particulate matter, genetic variants, and the risk of colorectal cancer. *Environment International* 147 (2021), 106309.

[152] H. B. Kim, J. Y. Shim, B. Park, and Y. J. Lee. Long-term exposure to air pollutants and cancer mortality: a meta-analysis of cohort studies. *International Journal of Environmental Research and Public Health* 15.11 (2018).

[153] D. Weller, D. Coleman, R. Robertson, P. Butler, J. Melia, C. Campbell, R. Parker, J. Patnick, and S. Moss. The uk colorectal cancer screening pilot: results of the second round of screening in england. *British Journal of Cancer* 97.12 (2007), 1601–5.

[154] M. Richards. Report of the independent review of adult screening programmes in england. 2019. `https://www.england.nhs.uk/wp-content/uploads/2019/02/report-of-the-independent-review-of-adult-screening-programme-in-england.pdf`, [Accessed 10-06-2022].

[155] R. C. Armitage and J. R. Morling. The impact of covid-19 on national screening programmes in england. *Public Health* 198 (2021), 174–176.

[156] A. I. Kooyker, E. Toes-Zoutendijk, A. W. J. Opstal-van Winden, M. C. W. Spaander, M. Buskermolen, H. J. van Vuuren, E. J. Kuipers, F. J. van Kemenade, C. Ramakers, M. G. J. Thomeer, E. Dekker, I. D. Nagtegaal, H. J. de Koning, M. E. van Leerdam, and I. Lansdorp-Vogelaar. The second round of the dutch colorectal cancer screening program: impact of an increased fecal immunochemical test cut-off level on yield of screening. *International Journal of Cancer* 147.4 (2020), 1098–1106.

[157] M. J. Frampton, P. Law, K. Litchfield, E. J. Morris, D. Kerr, C. Turnbull, I. P. Tomlinson, and R. S. Houlston. Implications of polygenic risk for personalised colorectal cancer screening. *Annals of Oncology* 27.3 (2016), 429–34.

[158] P. N, M. S, G. FJ, and P. PP. Cost-effectiveness and benefit-to-harm ratio of risk-stratified screening for breast cancer: a life-table model. *JAMA Oncology* 4.11 (2018), 1504–1510.

[159] UK National Screening Committee. Cmos' recommendations for the expanded remit of the uk nsc. `https://www.gov.uk/government/publications/uk-nsc-meeting-march-2022/cmos-recommendations-for-the-expanded-remit-of-the-uk-nsc`. 2022.

[160] T. Kortlever, M. van der Vlugt, and E. Dekker. Future of colorectal cancer screening: from one-size-fits-all to tailor-made. *Frontiers in Gastroenterology* 1 (2022).

[161] A. Umar, C. R. Boland, J. P. Terdiman, S. Syngal, A. d. l. Chapelle, J. R?schoff, R. Fishel, N. M. Lindor, L. J. Burgart, R. Hamelin, S. R. Hamilton, R. A. Hiatt, J. Jass, A. Lindblom, H. T. Lynch, P. Peltomaki, S. D. Ramsey, M. A. Rodriguez-Bigas, H. F. A. Vasen, E. T. Hawk, J. C. Barrett, A. N. Freedman, and S. Srivastava. Revised bethesda guidelines for hereditary nonpolyposis colorectal cancer (lynch syndrome) and microsatellite instability. *JNCI: Journal of the National Cancer Institute* 96.4 (2004), 261–268.

[162] L. Moreira, F. Balaguer, N. Lindor, A. de la Chapelle, H. Hampel, L. A. Aaltonen, J. L. Hopper, L. Le Marchand, S. Gallinger, P. A. Newcomb, R. Haile, S. N. Thibodeau, S. Gunawardena, M. A. Jenkins, D. D. Buchanan, J. D. Potter, J. A. Baron, D. J. Ahnen, V. Moreno, M. Andreu, M. Ponz de Leon, A. K. Rustgi, A. Castells, and the EPICOLON Consortium. Identification of lynch syndrome among patients with colorectal cancer. *JAMA* 308.15 (2012), 1555–1565.

[163] A. Latham, P. Srinivasan, Y. Kemel, J. Shia, C. Bandlamudi, D. Mandelker, S. Middha, J. Hechtman, A. Zehir, M. Dubard-Gault, C. Tran, C. Stewart, M. Sheehan, A. Penson, D. DeLair, R. Yaeger, J. Vijai, S. Mukherjee, J. Galle, M. A. Dickson, Y. Janjigian, E. M. O'Reilly, N. Segal, L. B. Saltz, D. Reidy-Lagunes, A. M. Varghese, D. Bajorin, M. I. Carlo, K. Cadoo, M. F. Walsh, M. Weiser, J. G. Aguilar, D. S. Klimstra, L. A. Diaz, J. Baselga, L. Zhang, M. Ladanyi, D. M. Hyman, D. B. Solit, M. E. Robson, B. S. Taylor, K. Offit, M. F. Berger, and Z. K. Stadler. Microsatellite instability is associated with the presence of lynch syndrome pan-cancer. *Journal of Clinical Oncology* 37.4 (2018), 286–295. PMID: 30376427.

[164] National Institute for Health and Care Excellence. Molecular testing strategies for lynch syndrome in people with colorectal cancer. *NICE guideline (DG27)* (2017).

[165] A. K. Win, M. A. Jenkins, J. G. Dowty, A. C. Antoniou, A. Lee, G. G. Giles, D. D. Buchanan, M. Clendenning, C. Rosty, D. J. Ahnen, S. N. Thibodeau, G. Casey, S. Gallinger, L. Le Marchand, R. W. Haile, J. D. Potter, Y. Zheng, N. M. Lindor, P. A. Newcomb, J. L. Hopper, and R. J. MacInnis. Prevalence and penetrance of major genes and polygenes for colorectal cancer. *Cancer Epidemiology, Biomarkers & Prevention* 26.3 (2017), 404–412.

[166] R. Manchanda, L. Sun, S. Patel, O. Evans, J. Wilschut, A. C. De Freitas Lopes, F. Gaba, A. Brentnall, S. Duffy, B. Cui, P. Coelho De Soarez, Z. Husain, J. Hopper, Z. Sadique, A. Mukhopadhyay, L. Yang, J. Berkhof, and R. Legood. Economic evaluation of population-based brca1/brca2 mutation testing across multiple countries and health systems. *Cancers (Basel)* 12.7 (2020).

[167] R. Manchanda, M. Burnell, F. Gaba, R. Desai, J. Wardle, S. Gessler, L. Side, S. Sanderson, K. Loggenberg, A. F. Brady, H. Dorkins, Y. Wallis, C. Chapman, C. Jacobs, R. Legood, U. Beller, I. Tomlinson, U. Menon, and I. Jacobs. Randomised trial of population-based brca testing in ashkenazi jews: long-term outcomes. *BJOG* 127.3 (2020), 364–375.

[168] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Journal of Clinical Epidemiology* 68.2 (2015), 134–43.

[169] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. Moons. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal* 338 (2009), b605.

[170] P. Royston, K. G. M. Moons, D. G. Altman, and Y. Vergouwe. Prognosis and prognostic research: developing a prognostic model. *British Medical Journal* 338 (2009).

[171] H. Wand, S. A. Lambert, C. Tamburro, M. A. Iacocca, J. W. O'Sullivan, C. Sillari, I. J. Kullo, R. Rowley, J. S. Dron, D. Brockman, E. Venner, M. I. McCarthy, A. C. Antoniou, D. F. Easton, R. A. Hegele, A. V. Khera, N. Chatterjee, C. Kooperberg, K. Edwards, K. Vlessis, K. Kinnear, J. N. Danesh, H. Parkinson, E. M. Ramos, M. C. Roberts, K. E. Ormond, M. J. Khoury, A. Janssens, K. A. B. Goddard, P. Kraft, J. A. L. MacArthur, M. Inouye, and G. L. Wojcik. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 591.7849 (2021), 211–219.

[172] J. A. Usher-Smith, A. Harshfield, C. L. Saunders, S. J. Sharp, J. Emery, F. M. Walter, K. Muir, and S. J. Griffin. External validation of risk prediction models for incident colorectal cancer using uk biobank. *British Journal Of Cancer* 118 (2018), 750 EP -.

[173] T. Smith, D. C. Muller, K. G. M. Moons, A. J. Cross, M. Johansson, P. Ferrari, G. Fagherazzi, P. H. M. Peeters, G. Severi, A. Hüsing, R. Kaaks, A. Tjonneland, A. Olsen, K. Overvad, C. Bonet, M. Rodriguez-Barranco, J. M. Huerta, A. Barricarte Gurrea, K. E. Bradbury, A. Trichopoulou, C. Bamia, P. Orfanos, D. Palli, V. Pala, P. Vineis, B. Bueno-de-Mesquita, B. Ohlsson, S. Harlid, B. Van Guelpen, G. Skeie, E. Weiderpass, M. Jenab, N. Murphy, E. Riboli, M. J. Gunter, K. J. Aleksandrova, and I. Tzoulaki. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the epic and uk biobank prospective cohort studies. *Gut* (2018).

[174] G. Colditz, K. Atwood, K. Emmons, R. Monson, W. Willett, D. Trichopoulos, and D. Hunter. Harvard report on cancer prevention volume 4: harvard cancer risk index. *Cancer Causes & Control* 11.6 (2000), 477–488.

[175] J. A. Driver, J. M. Gaziano, R. P. Gelber, I. M. Lee, J. E. Buring, and T. Kurth. Development of a risk score for colorectal cancer in men. *American Journal of Medicine* 120.3 (2007), 257–63.

[176] A. N. Freedman, M. L. Slattery, R. Ballard-Barbash, G. Willis, B. J. Cann, D. Pee, M. H. Gail, and R. M. Pfeiffer. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 27.5 (2009), 686–693.

[177] F. Guesmi, A. Zoghlami, D. Sghaiier, R. Nouira, and D. C. Alimentary factors promoting colorectal cancer risk: a prospective epidemiological study. *La Tunisie Médicale* 88.3 (2010), 184–189.

[178] C. M. Johnson, C. Wei, J. E. Ensor, D. J. Smolenski, C. I. Amos, B. Levin, and D. A. Berry. Meta-analyses of colorectal cancer risk factors. *Cancer Causes & Control* 24.6 (2013), 1207–1222.

[179] E. Ma, S. Sasazuki, M. Iwasaki, N. Sawada, and M. Inoue. 10-year risk of colorectal cancer: development and validation of a prediction model in middle-aged japanese men. *Cancer Epidemiology* 34.5 (2010), 534–541.

[180] S. Tao, M. Hoffmeister, and H. Brenner. Development and validation of a scoring system to identify individuals at high risk for advanced colorectal neoplasms who should undergo colonoscopy screening. *Clinical Gastroenterology and Hepatology* 12.3 (2014), 478–485.

[181] Y. S. Wei, J. C. Lu, L. Wang, P. Lan, H. J. Zhao, Z. Z. Pan, J. Huang, and J. P. Wang. Risk factors for sporadic colorectal cancer in southern chinese. *World Journal of Gastroenterology* 15.20 (2009), 2526–30.

[182] J. Hippisley-Cox and C. Coupland. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* 5.3 (2015).

[183] B. J. Wells, M. W. Kattan, G. S. Cooper, L. Jackson, and S. Koroukian. Colorectal cancer predicted risk online (crc-pro) calculator using data from the multi-ethnic cohort study. *The Journal of the American Board of Family Medicine* 27.1 (2014), 42–55.

[184] A. Steffen, T. I. A. Sørensen, S. Knüppel, N. Travier, M.-J. Sánchez, J. M. Huerta, J. R. Quirós, E. Ardanaz, M. Dorronsoro, B. Teucher, K. Li, H. B. Bueno-de-Mesquita, D. van der A, A. Mattiello, D. Palli, R. Tumino, V. Krogh, P. Vineis, A. Trichopoulou, P. Orfanos, D. Trichopoulos, B. Hedblad, P. Wallström, K. Overvad, J. Halkjær, A. Tjønneland, G. Fagherazzi, L. Dartois, F. Crowe, K.-T. Khaw, N. Wareham, L. Middleton, A. M. May, P. H. M. Peeters, and H. Boeing. Development and validation of a risk score predicting substantial weight gain over 5 years in middle-aged european men and women. *PLOS ONE* 8.7 (2013), 1–11.

[185] D. P. Taylor, G. J. Stoddard, R. W. Burt, M. S. Williams, J. A. Mitchell, P. J. Haug, and L. A. Cannon-Albright. How well does family history predict who will get colorectal cancer? implications for cancer screening and counseling. *Genetics in Medicine* 13.5 (2011), 385–91.

[186] L. Guo, H. Chen, G. Wang, Z. Lyu, X. Feng, L. Wei, X. Li, Y. Wen, M. Lu, Y. Chen, J. Shi, J. Ren, C. Lin, X. Yu, S. Chen, S. Wu, N. Li, M. Dai, and J. He. Development of a risk score for colorectal cancer in chinese males: a prospective cohort study. *Cancer Medicine* 9.2 (2020), 816–823.

[187] H. Hedlin, J. Weitlauf, C. J. Crandall, R. Nassir, J. A. Cauley, L. Garcia, R. Brunner, J. Robinson, M. L. Stefanick, and J. Robbins. Development of a comprehensive health-risk prediction tool for postmenopausal women. *Menopause* 26.12 (2019), 1385–1394.

[188] K. Aleksandrova, R. Reichmann, R. Kaaks, M. Jenab, H. B. Bueno-de-Mesquita, C. C. Dahm, A. K. Eriksen, A. Tjonneland, F. Artaud, M. C. Boutron-Ruault, G. Severi, A. Husing, A. Trichopoulou, A. Karakatsani, E. Peppa, S. Panico, G. Masala, S. Grioni, C. Sacerdote, R. Tumino, S. G. Elias, A. M. May, K. B. Borch, T. M. Sandanger, G. Skeie, M. J. Sanchez, J. M. Huerta, N. Sala, A. B. Gurrea, J. R. Quiros, P. Amiano, J. Berntsson, I. Drake, B. van Guelpen, S. Harlid, T. Key, E. Weiderpass, E. K. Aglago, A. J. Cross, K. K. Tsilidis, E. Riboli, and M. J. Gunter. Development and validation of a lifestyle-based model for colorectal cancer risk prediction: the lifecrc score. *BMC Medicine* 19.1 (2021), 1.

[189] J. A. Usher-Smith, S. J. Sharp, R. Luben, and S. J. Griffin. Development and validation of lifestyle-based models to predict incidence of the most common potentially preventable cancers. *Cancer Epidemiology, Biomarkers & Prevention* 28.1 (2019), 67–75.

[190] L. Peng, K. Weigl, D. Boakye, and H. Brenner. Risk scores for predicting advanced colorectal neoplasia in the average-risk population: a systematic review and meta-analysis. *American Journal of Gastroenterology* 113.12 (2018), 1788–1800.

[191] L. Peng, Y. Balavarca, K. Weigl, M. Hoffmeister, and H. Brenner. Head-to-head comparison of the performance of 17 risk models for predicting presence of advanced neoplasms in colorectal cancer screening. *American Journal of Gastroenterology* 114.9 (2019), 1520–1530.

[192] M. F. Kaminski, M. Polkowski, E. Kraszewska, M. Rupinski, E. Butruk, and J. Regula. A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. *Gut* 63.7 (2014), 1112–9.

[193] T. F. Imperiale, P. O. Monahan, T. E. Stump, and D. F. Ransohoff. Derivation and validation of a predictive model for advanced colorectal neoplasia in asymptomatic adults. *Gut* 70.6 (2021), 1155–1161.

[194] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21.1 (2010), 128–138.

[195] M. K. Thomsen, L. Pedersen, R. Erichsen, T. L. Lash, H. T. Sørensen, and E. M. Mikkelsen. Risk-stratified selection to colonoscopy in fit colorectal cancer screening: development and temporal validation of a prediction model. *British Journal of Cancer* 126.8 (2022), 1229–1235.

[196] V. H. Roos, F. G. J. Kallenberg, M. van der Vlugt, E. J. C. Bongers, C. M. Aalfs, P. M. M. Bossuyt, and E. Dekker. Addition of an online, validated family history questionnaire to the dutch fit-based screening programme did not improve its diagnostic yield. *British Journal of Cancer* 122.12 (2020), 1865–1871.

[197] I. Stegeman, T. R. de Wijkerslooth, E. M. Stoop, M. E. van Leerdam, E. Dekker, M. van Ballegooijen, E. J. Kuipers, P. Fockens, R. A. Kraaijenhagen, and P. M. Bossuyt. Combining risk factors with faecal immunochemical test outcome for selecting crc screenees for colonoscopy. *Gut* 63.3 (2014), 466–471.

[198] J. A. Cooper, N. Parsons, C. Stinton, C. Mathews, S. Smith, S. P. Halloran, S. Moss, and S. Taylor-Phillips. Risk-adjusted colorectal cancer screening using the fit and routine screening data: development of a risk prediction model. *British Journal of Cancer* 118.2 (2018), 285–293.

[199] J. A. Cooper, R. Ryan, N. Parsons, C. Stinton, T. Marshall, and S. Taylor-Phillips. The use of electronic healthcare records for colorectal cancer screening referral decisions and risk prediction model development. *BMC Gastroenterology* 20.1 (2020), 78.

[200] Y. M. Samarakoon, N. S. Gunawardena, A. Pathirana, M. N. Perera, and S. A. Hewage. Prediction of colorectal cancer risk among adults in a lower middle-income country. *Journal of Gastrointestinal Oncology* 10.3 (2019), 445–452.

[201] A. I. Sharara, A. El Mokahal, A. H. Harb, N. Khalaf, F. S. Sarkis, M. E.-H. M, N. M. Mansour, A. Malli, and R. Habib. Risk prediction rule for advanced neoplasia on screening colonoscopy for average-risk individuals. *World Journal of Gastroenterology* 26.37 (2020), 5705–5717.

[202] W. Guo, T. J. Key, and G. K. Reeves. Accelerometer compared with questionnaire measures of physical activity in relation to body size and composition: a large cross-sectional analysis of uk biobank. *BMJ Open* 9.1 (2019).

[203] K.-G. Yeoh, K.-Y. Ho, H.-M. Chiu, F. Zhu, J. Y. Ching, D.-C. Wu, T. Matsuda, J.-S. Byeon, S.-K. Lee, K.-L. Goh, et al. The asia-pacific colorectal screening score: a validated tool that stratifies risk for colorectal advanced neoplasia in asymptomatic asian subjects. *Gut* 60.9 (2011), 1236–1241.

[204] H. Chen, M. Lu, C. Liu, S. Zou, L. Du, X. Liao, D. Dong, D. Wei, Y. Gao, C. Zhu, et al. Comparative evaluation of participation and diagnostic yield of colonoscopy vs fecal immunochemical test vs risk-adapted screening in colorectal cancer screening: interim analysis of a multicenter randomized controlled trial (target-c). *Official journal of the American College of Gastroenterology| ACG* 115.8 (2020), 1264–1274.

[205] N. Chatterjee, J. Shi, and M. Garcia-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* 17.7 (2016), 392–406.

[206] A. Torkamani, N. E. Wineinger, and E. J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* 19.9 (2018), 581–590.

[207] N. R. Wray, S. H. Lee, D. Mehta, A. A. Vinkhuyzen, F. Dudbridge, and C. M. Middeldorp. Research review: polygenic methods and their application to psychiatric traits. *Journal of Child Psychology & Psychiatry* 55.10 (2014), 1068–87.

[208] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology* 37.2 (2013), 184–95.

[209] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* 41.6 (2017), 469–480.

[210] B. J. Vilhjálmsson et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics* 97.4 (2015), 576–592.

[211] F. Privé, J. Arbel, and B. J. Vilhjálmsson. Ldpred2: better, faster, stronger. *Bioinformatics* 36.22-23 (2020), 5424–5431.

[212] T. Ge, C. Y. Chen, Y. Ni, Y. A. Feng, and J. W. Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications* 10.1 (2019), 1776.

[213] A. C. J. W. Janssens, J. P. A. Ioannidis, C. M. van Duijn, J. Little, M. J. Khoury, and for the GRIPS Group. Strengthening the reporting of genetic risk prediction studies: the grips statement. *PLOS Medicine* 8.3 (2011), 1–4.

[214] L. McGeoch, C. L. Saunders, S. J. Griffin, J. D. Emery, F. M. Walter, D. J. Thompson, A. C. Antoniou, and J. A. Usher-Smith. Risk prediction models for colorectal cancer incorporating common genetic variants: a systematic review. *Cancer Epidemiology, Biomarkers & Prevention* 28.10 (2019), 1580–1593.

[215] C. L. Saunders, B. Kilian, D. J. Thompson, L. J. McGeoch, S. J. Griffin, A. C. Antoniou, J. D. Emery, F. M. Walter, J. Dennis, X. Yang, and J. A. Usher-Smith. External validation of risk prediction models incorporating common genetic variants for incident colorectal cancer using uk biobank. *Cancer Prevention Research* (2020), canprevres.0521.2019.

[216] X. Li, M. Timofeeva, A. Spiliopoulou, P. McKeigue, Y. He, X. Zhang, V. Svinti, H. Campbell, R. S. Houlston, I. P. M. Tomlinson, S. M. Farrington, M. G. Dunlop, and E. Theodoratou. Prediction of colorectal cancer risk based on profiling with common genetic variants. *International Journal of Cancer* 147.12 (2020), 3431–3437.

[217] M. Abe, H. Ito, I. Oze, M. Nomura, Y. Ogawa, and K. Matsuo. The more from east-asian, the better: risk prediction of colorectal cancer risk by gwas-identified snps among japanese. *Journal of Cancer Research and Clinical Oncology* 143.12 (2017), 2481–2492.

[218] M. Dunlop, A. Tenesa, S. Farrington, S. Ballereau, D. Brewster, T. Koessler, P. Pharoah, C. Schafmayer, J. Hampe, H. Volzke, J. Chang-Claude, M. Hoffmeister, H. Brenner, S. Von Holst, S. Picelli, A. Lindblom, M. Jenkins, J. Hopper, G. Casey, D. Duggan, P. Newcomb, A. Abuli, X. Bessa, C. Ruiz-Ponte, S. Castellvi-Bel, I. Niittymaki, S. Tuupanen, A. Karhu, L. Aaltonen, B. Zanke, T. Hudson, S. Gallinger, E. Barclay, L. Martin, M. Gorman, L. Carvajal-Carmona, A. Walther, D. Kerr, S. Lubbe, P. Broderick, I. Chandler, A. Pittman, S. Penegar, H. Campbell, I. Tomlinson, and R. Houlston. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42 103 individuals. *Gut* 62.6 (2013), 871–881.

[219] L. G. Fritsche, S. Patil, L. J. Beesley, P. VandeHaar, M. Salvatore, Y. Ma, R. B. Peng, D. Taliun, X. Zhou, and B. Mukherjee. Cancer prsweb: an online repository with polygenic risk scores for major cancer traits and their evaluation in two independent biobanks. *The American Journal of Human Genetics* 107.5 (2020), 815–836.

[220] S. Hosono, H. Ito, I. Oze, M. Watanabe, K. Komori, Y. Yatabe, Y. Shimizu, H. Tanaka, and K. Matsuo. A risk prediction model for colorectal cancer using genome-wide association study-identified polymorphisms and established risk factors among japanese: results from two independent case-control studies. *European Journal of Cancer Prevention* 25.6 (2016), 500–7.

[221] L. Hsu, J. Jeon, H. Brenner, S. B. Gruber, R. E. Schoen, S. I. Berndt, A. T. Chan, J. Chang-Claude, M. Du, J. Gong, T. A. Harrison, R. B. Hayes, M. Hoffmeister, C. M. Hutter, Y. Lin, R. Nishihara, S. Ogino, R. L. Prentice, F. R. Schumacher, D. Seminara, M. L. Slattery, D. C. Thomas, M. Thornquist, P. A. Newcomb, J. D. Potter, Y. Zheng, E. White, U. Peters, S. Colorectal Transdisciplinary, Genetics, and C. Epidemiology of Colorectal Cancer. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology* 148.7 (2015), 1330–9 e14.

[222] J. R. Huyghe, S. A. Bien, T. A. Harrison, H. M. Kang, S. Chen, S. L. Schmit, D. V. Conti, C. Qu, J. Jeon, C. K. Edlund, P. Greenside, M. Wainberg, F. R. Schumacher, J. D. Smith, D. M. Levine, S. C. Nelson, N. A. Sinnott-Armstrong, D. Albanes, M. H. Alonso, K. Anderson,

C. Arnau-Collell, V. Arndt, C. Bamia, B. L. Banbury, J. A. Baron, S. I. Berndt, S. Bézieau, D. T. Bishop, J. Boehm, H. Boeing, H. Brenner, S. Brezina, S. Buch, D. D. Buchanan, A. Burnett-Hartman, K. Butterbach, B. J. Caan, P. T. Campbell, C. S. Carlson, S. Castellvıé-Bel, A. T. Chan, J. Chang-Claude, S. J. Chanock, M.-D. Chirlaque, S. H. Cho, C. M. Connolly, A. J. Cross, K. Cuk, K. R. Curtis, A. de la Chapelle, K. F. Doheny, D. Duggan, D. F. Easton, S. G. Elias, F. Elliott, D. R. English, E. J. M. Feskens, J. C. Figueiredo, R. Fischer, L. M. FitzGerald, D. Forman, M. Gala, S. Gallinger, W. J. Gauderman, G. G. Giles, E. Gillanders, J. Gong, P. J. Goodman, W. M. Grady, J. S. Grove, A. Gsur, M. J. Gunter, R. W. Haile, J. Hampe, H. Hampel, S. Harlid, R. B. Hayes, P. Hofer, M. Hoffmeister, J. L. Hopper, W.-L. Hsu, W.-Y. Huang, T. J. Hudson, D. J. Hunter, G. Ibañez-Sanz, G. E. Idos, R. Ingersoll, R. D. Jackson, E. J. Jacobs, M. A. Jenkins, A. D. Joshi, C. E. Joshu, T. O. Keku, T. J. Key, H. R. Kim, E. Kobayashi, L. N. Kolonel, C. Kooperberg, T. Kühn, S. Küry, S.-S. Kweon, S. C. Larsson, C. A. Laurie, L. Le Marchand, S. M. Leal, S. C. Lee, F. Lejbkowicz, M. Lemire, C. I. Li, L. Li, W. Lieb, Y. Lin, A. Lindblom, N. M. Lindor, H. Ling, T. L. Louie, S. Männistö, S. D. Markowitz, V. Martıén, G. Masala, C. E. McNeil, M. Melas, R. L. Milne, L. Moreno, N. Murphy, R. Myte, A. Naccarati, P. A. Newcomb, K. Offit, S. Ogino, N. C. Onland-Moret, B. Pardini, P. S. Parfrey, R. Pearlman, V. Perduca, P. D. P. Pharoah, M. Pinchev, E. A. Platz, R. L. Prentice, E. Pugh, L. Raskin, G. Rennert, H. S. Rennert, E. Riboli, M. Rodrıéguez-Barranco, J. Romm, L. C. Sakoda, C. Schafmayer, R. E. Schoen, D. Seminara, M. Shah, T. Shelford, M.-H. Shin, K. Shulman, S. Sieri, M. L. Slattery, M. C. Southey, Z. K. Stadler, C. Stegmaier, Y.-R. Su, C. M. Tangen, S. N. Thibodeau, D. C. Thomas, S. S. Thomas, A. E. Toland, A. Trichopoulou, C. M. Ulrich, D. J. Van Den Berg, F. J. B. van Duijnhoven, B. Van Guelpen, H. van Kranen, J. Vijai, K. Visvanathan, P. Vodicka, L. Vodickova, V. Vymetalkova, K. Weigl, S. J. Weinstein, E. White, A. K. Win, C. R. Wolf, A. Wolk, M. O. Woods, A. H. Wu, S. H. Zaidi, B. W. Zanke, Q. Zhang, W. Zheng, P. C. Scacheri, J. D. Potter, M. C. Bassik, A. Kundaje, G. Casey, V. Moreno, G. R. Abecasis, D. A. Nickerson, S. B. Gruber, L. Hsu, and U. Peters. Discovery of common and rare genetic risk variants for colorectal cancer. *Nature Genetics* 51.1 (2019), 76–87.

[223] G. Ibañez-Sanz, A. Diez-Villanueva, M. H. Alonso, F. Rodriguez-Moranta, B. Perez-Gomez, M. Bustamante, V. Martin, J. Llorca, P. Amiano, E. Ardanaz, A. Tardon, J. J. Jimenez-Moleon, R. Peiro, J. Alguacil, C. Navarro, E. Guino, G. Binefa, P. Fernandez-Navarro, A. Espinosa, V. Davila-Batista, A. J. Molina, C. Palazuelos, G. Castano-Vinyals, N. Aragones, M. Kogevinas, M. Pollan, and V. Moreno. Risk model for colorectal cancer in spanish population using environmental and genetic factors: results from the mcc-spain study. *Scientific Reports* 7 (2017), 43263.

[224] M. Iwasaki, S. Tanaka-Mizuno, A. Kuchiba, T. Yamaji, N. Sawada, A. Goto, T. Shimazu, S. Sasazuki, H. Wang, L. L. Marchand, and S. Tsugane. Inclusion of a genetic risk score into a validated risk prediction model for colorectal cancer in japanese men improves performance. *Cancer Prevention Research* 10.9 (2017), 535.

[225] M. A. Jenkins, E. Makalic, J. G. Dowty, D. F. Schmidt, G. S. Dite, R. J. MacInnis, D. Ait Ouakrim, M. Clendenning, L. B. Flander, O. K. Stanesby, J. L. Hopper, A. K. Win, and D. D. Buchanan. Quantifying the utility of single nucleotide polymorphisms to guide colorectal cancer screening. *Future Oncology* 12.4 (2016), 503–513.

[226] J. Jeon, M. Du, R. E. Schoen, M. Hoffmeister, P. A. Newcomb, S. I. Berndt, B. Caan, P. T. Campbell, A. T. Chan, J. Chang-Claude, G. G. Giles, J. Gong, T. A. Harrison, J. R. Huyghe, E. J. Jacobs, L. Li, Y. Lin, L. L. Marchand, J. D. Potter, C. Qu, S. A. Bien, N. Zubair, R. J. Macinnis, D. D. Buchanan, J. L. Hopper, Y. Cao, R. Nishihara, G. Rennert, M. L. Slattery, D. C. Thomas, M. O. Woods, R. L. Prentice, S. B. Gruber, Y. Zheng, H. Brenner, R. B. Hayes, E. White, U. Peters, and L. Hsu. Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors. *Gastroenterology* 154.8 (2018), 2152–2164.e19.

[227] T. Smith, M. J. Gunter, I. Tzoulaki, and D. C. Muller. The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the uk biobank prospective cohort study. *British Journal of Cancer* 119.8 (2018), 1036–1039.

[228] M. Thomas, L. C. Sakoda, M. Hoffmeister, E. A. Rosenthal, J. K. Lee, F. J. van Duijnhoven, E. A. Platz, A. H. Wu, C. H. Dampier, A. de la Chapelle, A. Wolk, A. D. Joshi, A. Burnett-Hartman, A. Gsur, A. Lindblom, A. Castells, A. K. Win, B. Namjou, B. Van Guelpen, C. M. Tangen, Q. He, C. I. Li, C. Schafmayer, C. E. Joshu, C. M. Ulrich, D. T. Bishop, D. D. Buchanan, D. Schaid, D. A. Drew, D. C. Muller, D. Duggan, D. R. Crosslin, D. Albanes, E. L. Giovannucci, E. Larson, F. Qu, F. Mentch, G. G. Giles, H. Hakonarson, H. Hampel, I. B. Stanaway, J. C. Figueiredo, J. R. Huyghe, J. Minnier, J. Chang-Claude, J. Hampe, J. B. Harley, K. Visvanathan, K. R. Curtis, K. Offit, L. Li, L. Le Marchand, L. Vodickova, M. J. Gunter, M. A. Jenkins, M. L. Slattery, M. Lemire, M. O. Woods, M. Song, N. Murphy, N. M. Lindor, O. Dikilitas, P. D. Pharoah, P. T. Campbell, P. A. Newcomb, R. L. Milne, R. J. MacInnis, S. Castellví-Bel, S. Ogino, S. I. Berndt, S. Bézieau, S. N. Thibodeau, S. J. Gallinger, S. H. Zaidi, T. A. Harrison, T. O. Keku, T. J. Hudson, V. Vymetalkova, V. Moreno, V. Martín, V. Arndt, W.-Q. Wei, W. Chung, Y.-R. Su, R. B. Hayes, E. White, P. Vodicka, G. Casey, S. B. Gruber, R. E. Schoen, A. T. Chan, J. D. Potter, H. Brenner, G. P. Jarvik, D. A. Corley, U. Peters, and L. Hsu. Genome-wide modeling of polygenic risk score in colorectal cancer risk. *The American Journal of Human Genetics* (2020).

[229] H. M. Wang, T. H. Chang, F. M. Lin, T. H. Chao, W. C. Huang, C. Liang, C. F. Chu, C. M. Chiu, W. Y. Wu, M. C. Chen, C. T. Weng, S. L. Weng, F. F. Chiang, and H. D. Huang. A new method for post genome-wide association study (gwas) analysis of colorectal cancer in taiwan. *Gene* 518.1 (2013), 107–13.

[230] J. Xin, H. Chu, S. Ben, Y. Ge, W. Shao, Y. Zhao, Y. Wei, G. Ma, S. Li, D. Gu, Z. Zhang, M. Du, and M. Wang. Evaluating the effect of multiple genetic risk score models on colorectal cancer risk prediction. *Gene* 673 (2018), 174–180.

[231] J. M. Yarnall, D. J. Crouch, and C. M. Lewis. Incorporating non-genetic risk factors and behavioural modifications into risk prediction models for colorectal cancer. *Cancer Epidemiology* 37.3 (2013), 324–9.

[232] N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics* 14 (2013), 507–515.

[233] N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher. The genetic interpretation of area under the roc curve in genomic profiling. *PLOS Genetics* 6.2 (2010), e1000864.

[234] J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner, D. Liu, A. E. Locke, S. Balasubramanian, A. Yadav, N. Banerjee, C. E. Gillies, A. Damask, S. Liu, X. Bai, A. Hawes, E. Maxwell, L. Gurski, K. Watanabe, J. A. Kosmicki, V. Rajagopal, J. Mighty, C. Regeneron Genetics, DiscovEhr, M. Jones, L. Mitnaul, E. Stahl, G. Coppola, E. Jorgenson, L. Habegger, W. J. Salerno, A. R. Shuldiner, L. A. Lotta, J. D. Overton, M. N. Cantor, J. G. Reid, G. Yancopoulos, H. M. Kang, J. Marchini, A. Baras, G. R. Abecasis, and M. A. R. Ferreira. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature* 599.7886 (2021), 628–634.

[235] R. Mihaescu, M. J. Pencina, A. Alonso, K. L. Lunetta, S. R. Heckbert, E. J. Benjamin, and A. C. Janssens. Incremental value of rare genetic variants for the prediction of multifactorial diseases. *Genome Medicine* 5.8 (2013), 76.

[236] G. Abraham, R. Malik, E. Yonova-Doing, A. Salim, T. Wang, J. Danesh, A. S. Butterworth, J. M. M. Howson, M. Inouye, and M. Dichgans. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nature Communications* 10.1 (2019), 5819.

[237] P. R. Carr, K. Weigl, D. Edelmann, L. Jansen, J. Chang-Claude, H. Brenner, and M. Hoffmeister. Estimation of absolute risk of colorectal cancer based on healthy lifestyle, genetic risk, and colonoscopy status in a population-based study. *Gastroenterology* 159.1 (2020), 129–138 e9.

[238] L. Kachuri, R. E. Graff, K. Smith-Byrne, T. J. Meyers, S. R. Rashkin, E. Ziv, J. S. Witte, and M. Johansson. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nature Communications* 11.1 (2020), 6084.

[239] T. J. Iveson, R. S. Kerr, M. P. Saunders, J. Cassidy, N. H. Hollander, J. Tabernero, A. Haydon, B. Glimelius, A. Harkin, K. Allan, J. McQueen, C. Scudder, K. A. Boyd, A. Briggs, A. Waterston, L. Medley, C. Wilson, R. Ellis, S. Essapen, A. S. Dhadda, M. Harrison, S. Falk, S. Raouf, C. Rees, R. K. Olesen, D. Propper, J. Bridgewater, A. Azzabi, D. Farrugia, A. Webb, D. Cunningham, T. Hickish, A. Weaver, S. Gollins, H. S. Wasan, and J. Paul. 3 versus 6 months of adjuvant oxaliplatin-fluoropyrimidine combination therapy for colorectal cancer (scot): an international, randomised, phase 3, non-inferiority trial. *The Lancet Oncology* 19.4 (2018), 562–578.

[240] A. Schmermund, S. Möhlenkamp, A. Stang, D. Grönemeyer, R. Seibel, H. Hirche, K. Mann, W. Siffert, K. Lauterbach, J. Siegrist, K.-H. Jöckel, and R. Erbel. Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the heinz nixdorf recall study. *American Heart Journal* 144.2 (2002), 212–218.

[241] B. Winney, A. Boumertit, T. Day, D. Davison, C. Echeta, I. Evseeva, K. Hutnik, S. Leslie, K. Nicodemus, E. C. Royrvik, S. Tonks, X. Yang, J. Cheshire, P. Longley, P. Mateos, A. Groom, C. Relton, D. T. Bishop, K. Black, E. Northwood, L. Parkinson, T. M. Frayling, A. Steele, J. R. Sampson, T. King, R. Dixon, D. Middleton, B. Jennings, R. Bowden, P. Donnelly, and W. Bodmer. People of the british isles: preliminary analysis of genotypes and surnames in a uk-control population. *European Journal Of Human Genetics* 20 (2011), 203–210.

[242] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E. C. Royrvik, B. Cunliffe, W. T. C. C. C. 2, I. M. S. G. Consortium, D. J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, and W. Bodmer. The fine-scale genetic structure of the british population. *Nature* 519 (2015), 309–104.

[243] S. Penegar, W. Wood, S. Lubbe, I. Chandler, P. Broderick, E. Papaemmanuil, G. Sellick, R. Gray, J. Peto, and R. Houlston. National study of colorectal cancer genetics. *British Journal of Cancer* 97.9 (2007), 1305–1309.

[244] I. P. Tomlinson, E. Webb, L. Carvajal-Carmona, P. Broderick, K. Howarth, A. M. Pittman, S. Spain, S. Lubbe, A. Walther, K. Sullivan, E. Jaeger, S. Fielding, A. Rowan, J. Vijayakrishnan, E. Domingo, I. Chandler, Z. Kemp, M. Qureshi, S. M. Farrington, A. Tenesa, J. G. Prendergast, R. A. Barnetson, S. Penegar, E. Barclay, W. Wood, L. Martin, M. Gorman, H. Thomas, J. Peto, D. T. Bishop, R. Gray, E. R. Maher, A. Lucassen, D. Kerr, D. G. R. Evans, T. C. Consortium, C. Schafmayer, S. Buch, H. Völzke, J. Hampe, S. Schreiber, U. John, T. Koessler, P. Pharoah, T. van Wezel, H. Morreau, J. T. Wijnen, J. L. Hopper, M. C. Southey, G. G. Giles, G. Severi, S. Castellvıé-Bel, C. Ruiz-Ponte, A. Carracedo, A. Castells, T. E. Consortium, A. Försti, K. Hemminki, P. Vodicka, A. Naccarati, L. Lipton, J. W. Ho, K. K. Cheng, P. C. Sham, J. Luk, J. A. Agúndez, J. M. Ladero, M. de la Hoya, T. Caldés, I. Niittymäki, S. Tuupanen, A. Karhu, L. Aaltonen, J.-B. Cazier, H. Campbell, M. G. Dunlop, and R. S. Houlston. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature Genetics* 40 (2008), 623–630.

[245] The Practical Consortium. Practical. 2022. `http://practical.icr.ac.uk`, [Accessed 03-08-2022].

[246] The Breast Cancer Association Consortium. Bcac – breast cancer association consortium. 2022. `https://bcac.ccge.medschl.cam.ac.uk/`, [Accessed 03-08-2022].

[247] R. S. Houlston, E. Webb, P. Broderick, A. M. Pittman, M. C. Di Bernardo, S. Lubbe, I. Chandler, J. Vijayakrishnan, K. Sullivan, S. Penegar, L. Carvajal-Carmona, K. Howarth, E. Jaeger, S. L. Spain, A. Walther, E. Barclay, L. Martin, M. Gorman, E. Domingo, A. S. Teixeira, D. Kerr, J.-B. Cazier, I. Niittymäki, S. Tuupanen, A. Karhu, L. A. Aaltonen, I. P. M. Tomlinson, S. M. Farrington, A. Tenesa, J. G. D. Prendergast, R. A. Barnetson, R. Cetnarskyj, M. E. Porteous, P. D. P. Pharoah, T. Koessler, J. Hampe, S. Buch, C. Schafmayer, J. Tepel, S. Schreiber, H. Völzke, J. Chang-Claude, M. Hoffmeister, H. Brenner, B. W. Zanke, A. Montpetit, T. J. Hudson, S. Gallinger, H. Campbell, M. G. Dunlop, C. Study, Group, C. Colorectal Cancer Association Study, Group, R. G. I. C. Co, Group, and

C. International Colorectal Cancer Genetic Association. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetics* 40.12 (2008), 1426–1435.

[248] R. S. Houlston, J. Cheadle, S. E. Dobbins, A. Tenesa, A. M. Jones, K. Howarth, S. L. Spain, P. Broderick, E. Domingo, S. Farrington, J. G. D. Prendergast, A. M. Pittman, E. Theodoratou, C. G. Smith, B. Olver, A. Walther, R. A. Barnetson, M. Churchman, E. E. M. Jaeger, S. Penegar, E. Barclay, L. Martin, M. Gorman, R. Mager, E. Johnstone, R. Midgley, I. Niittymäki, S. Tuupanen, J. Colley, S. Idziaszczyk, T. C. Consortium, H. J. W. Thomas, A. M. Lucassen, D. G. R. Evans, E. R. Maher, T. C. Consortium, T. C. C. Group, T. C. C. Group, T. Maughan, A. Dimas, E. Dermitzakis, J.-B. Cazier, L. A. Aaltonen, P. Pharoah, D. J. Kerr, L. G. Carvajal-Carmona, H. Campbell, M. G. Dunlop, and I. P. M. Tomlinson. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nature Genetics* 42 (2010), 973–977.

[249] B. H. Smith, A. Campbell, P. Linksted, B. Fitzpatrick, C. Jackson, S. M. Kerr, I. J. Deary, D. J. Macintyre, H. Campbell, M. McGilchrist, L. J. Hocking, L. Wisely, I. Ford, R. S. Lindsay, R. Morton, C. N. Palmer, A. F. Dominiczak, D. J. Porteous, and A. D. Morris. Cohort profile: generation scotland: scottish family health study (gs:sfhs). the study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology* 42.3 (2013), 689–700.

[250] I. J. Deary, A. J. Gow, A. Pattie, and J. M. Starr. Cohort profile: the lothian birth cohorts of 1921 and 1936. *International Journal of Epidemiology* 41.6 (2012), 1576–84.

[251] U. Peters, C. M. Hutter, L. Hsu, F. R. Schumacher, D. V. Conti, C. S. Carlson, C. K. Edlund, R. W. Haile, S. Gallinger, B. W. Zanke, M. Lemire, J. Rangrej, R. Vijayaraghavan, A. T. Chan, A. Hazra, D. J. Hunter, J. Ma, C. S. Fuchs, E. L. Giovannucci, P. Kraft, Y. Liu, L. Chen, S. Jiao, K. W. Makar, D. Taverna, S. B. Gruber, G. Rennert, V. Moreno, C. M. Ulrich, M. O. Woods, R. C. Green, P. S. Parfrey, R. L. Prentice, C. Kooperberg, R. D. Jackson, A. Z. LaCroix, B. J. Caan, R. B. Hayes, S. I. Berndt, S. J. Chanock, R. E. Schoen, J. Chang-Claude, M. Hoffmeister, H. Brenner, B. Frank, S. Bézieau, S. Küry, M. L. Slattery, J. L. Hopper, M. A. Jenkins, L. Le Marchand, N. M. Lindor, P. A. Newcomb, D. Seminara, T. J. Hudson, D. J. Duggan, J. D. Potter, and G. Casey. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Human Genetics* 131.2 (2012), 217–234.

[252] D. J. Hunter, P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, J. Wang, K. Yu, N. Chatterjee, N. Orr, W. C. Willett, G. A. Colditz, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, R. B. Hayes, M. Tucker, D. S. Gerhard, J. F. Fraumeni Jr, R. N. Hoover, G. Thomas, and S. J. Chanock. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* 39 (2007), 870 EP -.

[253] R. A. Adams, A. M. Meade, M. T. Seymour, R. H. Wilson, A. Madi, D. Fisher, S. L. Kenny, E. Kay, E. Hodgkinson, M. Pope, P. Rogers, H. Wasan, S. Falk, S. Gollins, T. Hickish, E. M. Bessell, D. Propper, M. J. Kennedy, R. Kaplan, T. S. Maughan, and M. C. T. Investigators. Intermittent versus continuous oxaliplatin and fluoropyrimidine combination chemotherapy for first-line treatment of advanced colorectal cancer: results of the randomised phase 3 mrc coin trial. *Lancet Oncol* 12.7 (2011), 642–53.

[254] H. Wasan, A. M. Meade, R. Adams, R. Wilson, C. Pugh, D. Fisher, B. Sydes, A. Madi, B. Sizer, C. Lowdell, G. Middleton, R. Butler, R. Kaplan, T. Maughan, and C.-. B. investigators. Intermittent chemotherapy plus either intermittent or continuous cetuximab for first-line treatment of patients with kras wild-type advanced colorectal cancer (coin-b): a randomised phase 2 trial. *Lancet Oncol* 15.6 (2014), 631–9.

[255] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447.7145 (2007), 661–78.

[256] P. Hofer, M. Hagmann, S. Brezina, E. Dolejsi, K. Mach, G. Leeb, A. Baierl, S. Buch, H. Sutterlüty-Fall, J. Karner-Hanusch, M. M. Bergmann, T. Bachleitner-Hofmann, A. Stift, A. Gerger, K. Rötzer, J. Karner, S. Stättner, M. Waldenberger, T. Meitinger, K. Strauch, J. Linseisen, C. Gieger, F. Frommlet, and A. Gsur. Bayesian and frequentist analysis of an austrian genome-wide association study of colorectal cancer and advanced adenomas. *Oncotarget* 8.58 (2017), 98623–98634.

[257] Y. He, M. Timofeeva, S. M. Farrington, P. Vaughan-Shaw, V. Svinti, M. Walker, L. Zgaga, X. Meng, X. Li, A. Spiliopoulou, X. Jiang, E. Hyppönen, P. Kraft, D. P. Kiel, S. consortium, C. Hayward, A. Campbell, D. Porteous, K. Vucic, I. Kirac, M. Filipovic, S. E. Harris, I. J. Deary, R. Houlston, I. P. Tomlinson, H. Campbell, E. Theodoratou, and M. G. Dunlop. Exploring causality in the association between circulating 25-hydroxyvitamin d and colorectal cancer risk: a large mendelian randomisation study. *BMC Medicine* 16.1 (2018), 142, 142–142.

[258] H. Brenner, J. Chang-Claude, C. M. Seiler, A. Rickert, and M. Hoffmeister. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Annals of Internal Medicine* 154.1 (2011), 22–30.

[259] G. Orlando, P. J. Law, R. S. Houlston, H. Jarvinen, A. Lepisto, L. Renkonen-Sinisalo, J. Bohm, J.-P. Mecklin, B. F. Meyer, N. A. Al-Tassan, S. M. Wakil, C. Palles, E. Barclay, I. P. Tomlinson, L. Martin, M. G. Dunlop, M. N. Timofeeva, S. Farrington, A. Tenesa, J. P. Cheadle, A. Gylfe, E. Kaasinen, J. Kondelin, K. Palin, L. A. Aaltonen, S. Tuupanen, T. Cajuso, T. Tanskanen, U. A. H?nninen, J. Taipale, H. Campbell, C. G. Smith, S. Idziaszczyk, T. S. Maughan, R. Kaplan, R. Kerr, D. Kerr, D. D. Buchanan, A. K. Win, J. Hopper, M. Jenkins, N. M. Lindor, P. A. Newcomb, S. Gallinger, D. Conti, F. Schumacher, G. Casey, A.-P. Sarin, S. Ripatti, A. Palotie, J. Kaprio, H. Rissanen, P. Knekt, P. Jousilahti, V. Salomaa, J. G. Eriksson, and E. Pukkala. Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Human Molecular Genetics* 25.11 (2016), 2349–2359.

[260] R. S. Midgley, C. C. McConkey, E. C. Johnstone, J. A. Dunn, J. L. Smith, S. A. Grumett, P. Julier, C. Iveson, Y. Yanagisawa, B. Warren, M. J. Langman, and D. J. Kerr. Phase iii randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the victor trial. *Journal of Clinical Oncology* 28.30 (2010), 4575–4580. PMID: 20837956.

[261] R. S. Kerr, S. Love, E. Segelov, E. Johnstone, B. Falcon, P. Hewett, A. Weaver, D. Church, C. Scudder, S. Pearson, P. Julier, F. Pezzella, I. Tomlinson, E. Domingo, and D. J. Kerr. Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (quasar 2): an open-label, randomised phase 3 trial. *The Lancet Oncology* 17.11 (2016), 1543–1557.

[262] C. Power and J. Elliott. Cohort profile: 1958 british birth cohort (national child development study). *International Journal of Epidemiology* 35.1 (2006), 34–41.

[263] J. C. Taylor, H. C. Martin, S. Lise, J. Broxholme, J.-B. Cazier, A. Rimmer, A. Kanapin, G. Lunter, S. Fiddy, C. Allan, A. R. Aricescu, M. Attar, C. Babbs, J. Becq, D. Beeson, C. Bento, P. Bignell, E. Blair, V. J. Buckle, K. Bull, O. Cais, H. Cario, H. Chapel, R. R. Copley, R. Cornall, J. Craft, K. Dahan, E. E. Davenport, C. Dendrou, O. Devuyst, A. L. Fenwick, J. Flint, L. Fugger, R. D. Gilbert, A. Goriely, A. Green, I. H. Greger, R. Grocock, A. V. Gruszczyk, R. Hastings, E. Hatton, D. Higgs, A. Hill, C. Holmes, M. Howard, L. Hughes, P. Humburg, D. Johnson, F. Karpe, Z. Kingsbury, U. Kini, J. C. Knight, J. Krohn, S. Lamble, C. Langman, L. Lonie, J. Luck, D. McCarthy, S. J. McGowan, M. F. McMullin, K. A. Miller, L. Murray, A. H. Németh, M. A. Nesbit, D. Nutt, E. Ormondroyd, A. B. Oturai, A. Pagnamenta, S. Y. Patel, M. Percy, N. Petousi, P. Piazza, S. E. Piret, G. Polanco-Echeverry, N. Popitsch, F. Powrie, C. Pugh, L. Quek, P. A. Robbins, K. Robson, A. Russo, N. Sahgal, P. A. van Schouwenburg, A. Schuh, E. Silverman, A. Simmons, P. S. Sørensen, E. Sweeney, J. Taylor, R. V. Thakker, I. Tomlinson, A. Trebes, S. R. Twigg, H. H. Uhlig, P. Vyas, T. Vyse, S. A. Wall, H. Watkins, M. P. Whyte, L. Witty, B. Wright, C. Yau, D. Buck, S. Humphray, P. J. Ratcliffe, J. I. Bell, A. O. Wilkie, D. Bentley, P. Donnelly, and G. McVean. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics* 47.7 (2015), 717–726.

[264] N. Whiffin, S. E. Dobbins, F. J. Hosking, C. Palles, A. Tenesa, Y. Wang, S. M. Farrington, A. M. Jones, P. Broderick, H. Campbell, P. A. Newcomb, G. Casey, D. V. Conti, F. Schumacher, S. Gallinger, N. M. Lindor, J. Hopper, M. Jenkins, M. G. Dunlop, I. P. Tomlinson, and R. S. Houlston. Deciphering the genetic architecture of low-penetrance susceptibility to colorectal cancer. *Human Molecular Genetics* 22.24 (2013), 5075–82.

[265] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun,

J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church, and C. A. Reid. Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. *Science* 327.5961 (2010), 78–81.

[266] A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, and N. E. Allen. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* 186.9 (2017), 1026–1034.

[267] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature* 562.7726 (2018), 203–209.

[268] QResearch. Qresearch. `https://www.qresearch.org`. 2019.

[269] P. Townsend, P. Phillimor, and B. A. *Health and Deprivation: Inequality and the North*. Croom Helm, 1987.

[270] ClinRisk Ltd. Qcancer®(15yr,colorectal). Computer Program. 2015.

[271] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols* 5 (2010), 1564.

[272] S. Purcell and C. Chang. Plink 1.9. `www.cog-genomics.org/plink/1.9/`. 2018.

[273] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* 39.7 (2007), 906–913.

[274] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics* 5.6 (2009), 1–15.

[275] Rayner, Will. *Genotyping chips strand and build files*. 2018. `https://www.well.ox.ac.uk/~wrayner/strand/`, [Accessed: February 2018].

[276] O. Delaneau, J. Marchini, and T. 1. G. P. Consortium. Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel. *Nature Communications* 5 (2014), 3934.

[277] The UK10K Consortium. The uk10k project identifies rare variants in health and disease. *Nature* 526 (2015), 82–90.

[278] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526 (2015), 68–74.

[279] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30 (2001), 97–101.

[280] A. L. Price, M. E. Weale, N. Patterson, S. R. Myers, A. C. Need, K. V. Shianna, D. Ge, J. I. Rotter, E. Torres, K. D. Taylor, D. B. Goldstein, and D. Reich. Long-range ld can confound genome scans in admixed populations. *American Journal of Human Genetics* 83.1 (2008), 132–5, 132–5.

[281] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38 (2006), 904–909.

[282] S. W. Choi, T. S. Mak, and P. F. O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* 15.9 (2020), 2759–2772.

[283] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), 187–220.

[284] E. Steyerberg. Selection of main effects. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer New York, 2009.

[285] F. E. Harrell. *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York, NY: Springer, 2001.

[286] J. Z. Liu, F. Tozzi, D. M. Waterworth, S. G. Pillai, P. Muglia, L. Middleton, W. Berrettini, C. W. Knouff, X. Yuan, G. Waeber, P. Vollenweider, M. Preisig, N. J. Wareham, J. H. Zhao, R. J. F. Loos, I. Barroso, K.-T. Khaw, S. Grundy, P. Barter, R. Mahley, A. Kesaniemi, R. McPherson, J. B. Vincent, J. Strauss, J. L. Kennedy, A. Farmer, P. McGuffin, R. Day, K. Matthews, P. Bakke, A. Gulsvik, S. Lucae, M. Ising, T. Brueckl, S. Horstmann, H.-E. Wichmann, R. Rawal, N. Dahmen, C. Lamina, O. Polasek, L. Zgaga, J. Huffman, S. Campbell, J. Kooner, J. C. Chambers, M. S. Burnett, J. M. Devaney, A. D. Pichard, K. M. Kent, L. Satler, J. M. Lindsay, R. Waksman, S. Epstein, J. F. Wilson, S. H. Wild, H. Campbell, V. Vitart, M. P. Reilly, M. Li, L. Qu, R. Wilensky, W. Matthai, H. H. Hakonarson, D. J. Rader, A. Franke, M. Wittig, A. Schäfer, M. Uda, A. Terracciano, X. Xiao, F. Busonero, P. Scheet, D. Schlessinger, D. S. Clair, D. Rujescu, G. R. Abecasis, H. J. Grabe, A. Teumer, H. Völzke, A. Petersmann, U. John, I. Rudan, C. Hayward, A. F. Wright, I. Kolcic, B. J. Wright, J. R. Thompson, A. J. Balmforth, A. S. Hall, N. J. Samani, C. A. Anderson, T. Ahmad, C. G. Mathew, M. Parkes, J. Satsangi, M. Caulfield, P. B. Munroe, M. Farrall, A. Dominiczak, J. Worthington, W. Thomson, S. Eyre, A. Barton, T. W. T. C. C. Consortium, V. Mooser, C. Francks, and J. Marchini. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature Genetics* 42 (2010), 436–440.

[287] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143.1 (1982), 29–36. PMID: 7063747.

[288] R. H. Somers. A new asymmetric measure of association for ordinal variables. *American Sociological Review* 27 (1962), 799–811.

[289] J. Harrell F. E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247.18 (1982), 2543–6.

[290] P. Royston and W. Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in Medicine* 23.5 (2004), 723–48.

[291] P. Royston and D. G. Altman. External validation of a cox prognostic model: principles and methods. *BMC Medical Research Methodology* 13.1 (2013), 33.

[292] D. R. Cox. Note on grouping. *Journal of the American Statistical Association* 52.280 (1957), 543–547.

[293] P. C. Austin and E. W. Steyerberg. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine* 33.3 (2014), 517–535.

[294] C. Freeman and J. Marchini. *GTOOL*. 2012. GTOOL version 0.7.5.

[295] G. Band and J. Marchini. *qctool v2*.

[296] J. E. Wigginton and G. R. Abecasis. Pedstats: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 21.16 (), 3445–7.

[297] H. Li, B. Handsaker, P. Danecek, S. McCarthy, and J. Marshall. *BCFtools*. 2021. BCFtools.

[298] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019.

[299] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software* 4.43 (2019), 1686.

[300] F. Privé, H. Aschard, A. Ziyatdinov, and M. G. Blum. Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. *Bioinformatics* (2018).

[301] T. M. Therneau. *A Package for Survival Analysis in R*. 2015. version 2.38.

[302] F. E. Harrell Jr. *rms: Regression Modeling Strategies*. 2019. R package version 5.1-4.

[303] original by Gareth Ambler and modified by Axel Benner. *mfp: Multivariable Fractional Polynomials*. 2015. R package version 1.5.2.

[304] F. E. Harrell Jr, with contributions from Charles Dupont, and many others. *Hmisc: Harrell Miscellaneous*. 2021. R package version 4.5-0.

[305] T. J. Aragon. *epitools: Epidemiology Tools*. 2021. R package version 0.5-10.1.

[306] E. Waring, M. Quinn, A. McNamara, E. Arino de la Rubia, H. Zhu, and S. Ellis. *skimr: Compact and Flexible Summaries of Data*. 2020. R package version 2.1.2.

[307] G. Grolemund and H. Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software* 40.3 (2011), 1–25.

[308] F. Solt and Y. Hu. *interplot: Plot the Effects of Variables in Interaction Terms*. 2021. R package version 0.2.3.

[309] B. Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. 2017. R package version 2.3.

[310] E. Harrison, T. Drake, and R. Ots. *finalfit*. 2021. R package version 1.0.3.

[311] C. Gandrud. *simPH: Simulate and Plot Estimates from Cox Proportional Hazards Models*. 2019. R package version 1.3.13.

[312] S. L. Schmit, C. K. Edlund, F. R. Schumacher, J. Gong, T. A. Harrison, J. R. Huyghe, C. Qu, M. Melas, D. J. Van Den Berg, H. Wang, S. Tring, S. J. Plummer, D. Albanes, M. H. Alonso, C. I. Amos, K. Anton, A. K. Aragaki, V. Arndt, E. L. Barry, S. I. Berndt, S. Bezieau, S. Bien, A. Bloomer, J. Boehm, M. C. Boutron-Ruault, H. Brenner, S. Brezina, D. D. Buchanan, K. Butterbach, B. J. Caan, P. T. Campbell, C. S. Carlson, J. E. Castelao, A. T. Chan, J. Chang-Claude, S. J. Chanock, I. Cheng, Y. W. Cheng, L. S. Chin, J. M. Church, T. Church, G. A. Coetzee, M. Cotterchio, M. Cruz Correa, K. R. Curtis, D. Duggan, D. F. Easton, D. English, E. J. M. Feskens, R. Fischer, L. M. FitzGerald, B. K. Fortini, L. G. Fritsche, C. S. Fuchs, M. Gago-Dominguez, M. Gala, S. J. Gallinger, W. J. Gauderman, G. G. Giles, E. L. Giovannucci, S. M. Gogarten, C. Gonzalez-Villalpando, E. M. Gonzalez-Villalpando, W. M. Grady, J. K. Greenson, A. Gsur, M. Gunter, C. A. Haiman, J. Hampe, S. Harlid, J. F. Harju, R. B. Hayes, P. Hofer, M. Hoffmeister, J. L. Hopper, S. C. Huang, J. M. Huerta, T. J. Hudson, D. J. Hunter, G. E. Idos, M. Iwasaki, R. D. Jackson, E. J. Jacobs, S. H. Jee, M. A. Jenkins, W. H. Jia, S. Jiao, A. D. Joshi, L. N. Kolonel, S. Kono, C. Kooperberg, V. Krogh, T. Kuehn, S. Kury, A. LaCroix, C. A. Laurie, F. Lejbkowicz, M. Lemire, H. J. Lenz, D. Levine, et al. Novel common genetic susceptibility loci for colorectal cancer. *Journal of the National Cancer Institute* 111.2 (2019), 146–157.

[313] S. Briggs and I. Tomlinson. Germline and somatic polymerase epsilon and delta mutations define a new class of hypermutated colorectal and endometrial cancers. *Journal of Pathology* 230.2 (2013), 148–53.

[314] European Genome-Phenome Archive. *WTCCC2 People of the British Isles (POBI) samples using Illumina 1.2M array*. 2018. `https://ega-archive.org/datasets/EGAD00010000632`, [Accessed: January 2018].

[315] L. R. Cardon and L. J. Palmer. Population stratification and spurious allelic association. *Lancet* 361.9357 (2003), 598–604.

[316] Y. S. Aulchenko, M. V. Struchalin, and C. M. van Duijn. Probabel package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 11 (2010), 134.

[317] J. Wakefield. A bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics* 81.2 (2007), 208–27.

[318] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. Gcta: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* 88.1 (2011), 76–82.

[319] I. P. Tomlinson, M. Dunlop, H. Campbell, B. Zanke, S. Gallinger, T. Hudson, T. Koessler, P. D. Pharoah, I. Niittymaki, S. Tuupanen, L. A. Aaltonen, K. Hemminki, A. Lindblom, A. Forsti, O. Sieber, L. Lipton, T. van Wezel, H. Morreau, J. T. Wijnen, P. Devilee, K. Matsuda, Y. Nakamura, S. Castellvi-Bel, C. Ruiz-Ponte, A. Castells, A. Carracedo, J. W. Ho, P. Sham, R. M. Hofstra, P. Vodicka, H. Brenner, J. Hampe, C. Schafmayer, J. Tepel, S. Schreiber, H. Volzke, M. M. Lerch, C. A. Schmidt, S. Buch, V. Moreno, C. M. Villanueva, P. Peterlongo, P. Radice, M. M. Echeverry, A. Velez, L. Carvajal-Carmona, R. Scott, S. Penegar, P. Broderick, A. Tenesa, and R. S. Houlston. Cogent

(colorectal cancer genetics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer. *British Journal of Cancer* 102.2 (2010), 447–54.

[320] J. Mao, Z. Sun, Y. Cui, N. Du, H. Guo, J. Wei, Z. Hao, and L. Zheng. Pcbp2 promotes the development of glioma by regulating fhl3/tgf-beta/smad signaling pathway. *Journal of Cell Physiology* 235.4 (2020), 3280–3291.

[321] Z. Huang, C. Yu, L. Yu, H. Shu, and X. Zhu. The roles of fhl3 in cancer. *Frontiers in Oncology* 12 (2022), 887828.

[322] E. Siebring-van Olst, M. Blijlevens, R. X. de Menezes, I. H. van der Meulen-Muileman, E. F. Smit, and V. W. van Beusechem. A genome-wide sirna screen for regulators of tumor suppressor p53 activity in human non-small cell lung cancer cells identifies components of the rna splicing machinery as targets for anticancer treatment. *Molecular Oncology* 11.5 (2017), 534–551.

[323] L. Del Bosque-Plata, E. P. Hernandez-Cortes, and C. Gragnoli. The broad pathogenetic role of tcf7l2 in human diseases beyond type 2 diabetes. *Journal of Cell Physiology* 237.1 (2022), 301–312.

[324] L. J. Barber, J. L. Youds, J. D. Ward, M. J. McIlwraith, N. J. O'Neil, M. I. Petalcorin, J. S. Martin, S. J. Collis, S. B. Cantor, M. Auclair, H. Tissenbaum, S. C. West, A. M. Rose, and S. J. Boulton. Rtel1 maintains genomic stability by suppressing homologous recombination. *Cell* 135.2 (2008), 261–71.

[325] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y. J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M. S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature* 463.7283 (2010), 899–905.

[326] G. Han, G. Yang, D. Hao, Y. Lu, K. Thein, B. S. Simpson, J. Chen, R. Sun, O. Alhalabi, R. Wang, M. Dang, E. Dai, S. Zhang, F. Nie, S. Zhao, C. Guo, A. Hamza, B. Czerniak, C. Cheng, A. Siefker-Radtke, K. Bhat, A. Futreal, G. Peng, J. Wargo, W. Peng, H. Kadara, J. Ajani, C. Swanton, K. Litchfield, J. R. Ahnert, J. Gao, and L. Wang. 9p21 loss confers a cold tumor immune microenvironment and primary resistance to immune checkpoint therapy. *Nature Communications* 12.1 (2021), 5606.

[327] T. Kojima, T. Shimazui, S. Hinotsu, A. Joraku, T. Oikawa, K. Kawai, R. Horie, H. Suzuki, R. Nagashima, K. Yoshikawa, T. Michiue, M. Asashima, H. Akaza, and K. Uchida. Decreased expression of cxxc4 promotes a malignant phenotype in renal cell carcinoma by activating wnt signaling. *Oncogene* 28.2 (2009), 297–305.

[328] J. Lu, S. Lu, J. Li, Q. Yu, L. Liu, and Q. Li. Mir-629-5p promotes colorectal cancer progression through targetting cxxc finger protein 4. *Bioscience Reports* 38.4 (2018).

[329] A. N. Sigafoos, B. D. Paradise, and M. E. Fernandez-Zapico. Hedgehog/gli signaling pathway: transduction, regulation, and implications for disease. *Cancers (Basel)* 13.14 (2021).

[330] L. Rosano, F. Spinella, and A. Bagnato. Endothelin 1 in cancer: biological implications and therapeutic opportunities. *Nature Reviews Cancer* 13.9 (2013), 637–51.

[331] C. Liu, C. E. Banister, C. C. Weige, D. Altomare, J. H. Richardson, C. M. Contreras, and P. J. Buckhaults. Prdm1 silences stem cell-related genes and inhibits proliferation of human colon tumor organoids. *Proceedings of the National Academy of Sciences USA* 115.22 (2018), E5066–E5075.

[332] A. P. Klein, B. M. Wolpin, H. A. Risch, R. Z. Stolzenberg-Solomon, E. Mocci, M. Zhang, F. Canzian, E. J. Childs, J. W. Hoskins, A. Jermusyk, J. Zhong, F. Chen, D. Albanes, G. Andreotti, A. A. Arslan, A. Babic, W. R. Bamlet, L. Beane-Freeman, S. I. Berndt, A. Blackford, M. Borges, A. Borgida, P. M. Bracci, L. Brais, P. Brennan, H. Brenner, B. Bueno-de-Mesquita, J. Buring, D. Campa, G. Capurso, G. M. Cavestro, K. G. Chaffee, C. C. Chung, S. Cleary, M. Cotterchio, F. Dijk, E. J. Duell, L. Foretova, C. Fuchs, N. Funel, S. Gallinger, M. G. JM, M. Gazouli, G. G. Giles, E. Giovannucci, M. Goggins, G. E. Goodman, P. J. Goodman, T. Hackert, C. Haiman, P. Hartge, M. Hasan, P. Hegyi, K. J. Helzlsouer, J. Herman, I. Holcatova, E. A. Holly, R. Hoover, R. J. Hung, E. J. Jacobs, K. Jamroziak, V. Janout, R. Kaaks, K. T. Khaw, E. A. Klein, M. Kogevinas, C. Kooperberg, M. H. Kulke, J. Kupcinskas, R. J. Kurtz, D. Laheru, S. Landi, R. T. Lawlor, I. M. Lee, L. LeMarchand, L. Lu, N. Malats, A. Mambrini, S. Mannisto, R. L. Milne, B. Mohelnikova-Duchonova, R. E. Neale, J. P. Neoptolemos, A. L. Oberg, S. H. Olson, I. Orlow, C. Pasquali, A. V. Patel, U. Peters, R. Pezzilli, M. Porta, F. X. Real, N. Rothman, G. Scelo, H. D. Sesso, G. Severi, X. O. Shu, D. Silverman, J. P. Smith, P. Soucek, et al. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nature Communications* 9.1 (2018), 556.

[333] L. Qin, X. Cao, T. Kaneko, C. Voss, X. Liu, G. Wang, and S. S. Li. Dynamic interplay of two molecular switches enabled by the mek1/2-erk1/2 and il-6-stat3 signaling axes controls epithelial cell migration in response to growth factors. *Journal of Biological Chemistry* 297.4 (2021), 101161.

[334] A. Kong and N. J. Cox. Allele-sharing models: lod scores and accurate linkage tests. *American Journal of Human Genetics* 61.5 (1997), 1179–88.

[335] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna,

B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and C. Exome Aggregation. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536.7616 (2016), 285–91.

[336] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The ensembl variant effect predictor. *Genome Biology* 17.1 (2016), 122.

[337] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Research* 11.5 (2001), 863–74.

[338] V. Ramensky, P. Bork, and S. Sunyaev. Human non-synonymous snps: server and survey. *Nucleic Acids Research* 30.17 (2002), 3894–900.

[339] L. D. Ward and M. Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* 40.Database issue (2012), D930–4.

[340] P. A. Zweidler-Mckay, H. L. Grimes, M. M. Flubacher, and P. N. Tsichlis. Gfi-1 encodes a nuclear zinc finger protein that binds dna and functions as a transcriptional repressor. *Journal of Molecular Cell Biology* 16.8 (1996), 4024–34.

[341] L. McGhee, J. Bryan, L. Elliott, H. L. Grimes, A. Kazanjian, J. N. Davis, and S. Meyers. Gfi-1 attaches to the nuclear matrix, associates with eto (mtg8) and histone deacetylase proteins, and represses transcription using a tsa-sensitive mechanism. *Journal of Cellular Biochemistry* 89.5 (2003), 1005–18.

[342] M. S. Chen, Y. H. Lo, X. Chen, C. S. Williams, J. M. Donnelly, 2. Criss Z. K., S. Patel, J. M. Butkus, J. Dubrulle, M. J. Finegold, and N. F. Shroyer. Growth factor-independent 1 is a tumor suppressor gene in colorectal cancer. *Molecular Cancer Research* 17.3 (2019), 697–708.

[343] D. Martin-Zanca, S. H. Hughes, and M. Barbacid. A human oncogene formed by the fusion of truncated tropomyosin and protein tyrosine kinase sequences. *Nature* 319.6056 (1986), 743–8.

[344] A. P. G. Consortium. Aacr project genie: powering precision medicine through an international consortium. *Cancer Discovery* 7.8 (2017), 818–831.

[345] T. Tsunoda and S. Shirasawa. Roles of zfat in haematopoiesis, angiogenesis and cancer development. *Anticancer Research* 33.7 (2013), 2833–7.

[346] T. Fujimoto, K. Doi, M. Koyanagi, T. Tsunoda, Y. Takashima, Y. Yoshida, T. Sasazuki, and S. Shirasawa. Zfat is an antiapoptotic molecule and critical for cell survival in molt-4 cells. *FEBS Letters* 583.3 (2009), 568–72.

[347] M. Ramakrishna, L. H. Williams, S. E. Boyle, J. L. Bearfoot, A. Sridhar, T. P. Speed, K. L. Gorringe, and I. G. Campbell. Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis. *PLoS One* 5.4 (2010), e9983.

[348] C. Rosa-Ferreira and S. Munro. Arl8 and skip act together to link lysosomes to kinesin-1. *Developmental Cell* 21.6 (2011), 1171–8.

[349] J. Pu, C. M. Guardia, T. Keren-Kaplan, and J. S. Bonifacino. Mechanisms and functions of lysosome positioning. *Journal of Cell Science* 129.23 (2016), 4329–4339.

[350] T. Li, F. Zhang, Z. Wu, L. Cui, X. Zhao, J. Wang, and Y. Hu. Plekhm2-alk: a novel fusion in small-cell lung cancer and durable response to alk inhibitors. *Lung Cancer* 139 (2020), 146–150.

[351] J. Heuser. Changes in lysosome shape and distribution correlated with changes in cytoplasmic ph. *Journal of Cell Biology* 108.3 (1989), 855–64.

[352] J. J. Steffan, B. C. Williams, T. Welbourne, and J. A. Cardelli. Hgf-induced invasion by prostate tumor cells requires anterograde lysosome trafficking and activity of na+-h+ exchangers. *Journal of Cell Science* 123.Pt 7 (2010), 1151–9.

[353] D. Xie, N. F. Ma, Z. Z. Pan, H. X. Wu, Y. D. Liu, G. Q. Wu, H. F. Kung, and X. Y. Guan. Overexpression of eif-5a2 is associated with metastasis of human colorectal carcinoma. *Human Pathology* 39.1 (2008), 80–6.

[354] W. Zhu, M. Y. Cai, Z. T. Tong, S. S. Dong, S. J. Mai, Y. J. Liao, X. W. Bian, M. C. Lin, H. F. Kung, Y. X. Zeng, X. Y. Guan, and D. Xie. Overexpression of eif5a2 promotes colorectal carcinoma cell aggressiveness by upregulating mta1 through c-myc to induce epithelial-mesenchymaltransition. *Gut* 61.4 (2012), 562–75.

[355] N. Shivapurkar, A. Maitra, S. Milchgrub, and A. F. Gazdar. Deletions of chromosome 4 occur early during the pathogenesis of colorectal carcinoma. *Human Pathology* 32.2 (2001), 169–77.

[356] A. D. Beggs, A. Jones, N. Shepherd, A. Arnaout, C. Finlayson, A. M. Abulafi, D. G. Morton, G. M. Matthews, S. V. Hodgson, and I. P. Tomlinson. Loss of expression and promoter methylation of slit2 are associated with sessile serrated adenoma formation. *PLOS Genetics* 9.5 (2013), e1003488.

[357] N. Brusselaers, K. Ekwall, and M. Durand-Dubief. Copy number of 8q24.3 drives hsf1 expression and patient outcome in cancer: an individual patient data meta-analysis. *Human Genomics* 13.1 (2019), 54.

[358] E. J. Douglas, H. Fiegler, A. Rowan, S. Halford, D. C. Bicknell, W. Bodmer, I. P. Tomlinson, and N. P. Carter. Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Research* 64.14 (2004), 4817–25.

[359] Y. Bosse, Z. Li, J. Xia, V. Manem, R. Carreras-Torres, A. Gabriel, N. Gaudreault, D. Albanes, M. C. Aldrich, A. Andrew, S. Arnold, H. Bickeboller, S. E. Bojesen, P. Brennan, H. Brunnstrom, N. Caporaso, C. Chen, D. C. Christiani, J. K. Field, G. Goodman, K. Grankvist, R. Houlston, M. Johansson, M. Johansson, L. A. Kiemeney, S. Lam, M. T. Landi, P. Lazarus, L. Le Marchand, G. Liu, O. Melander, G. Rennert, A. Risch, S. M. Rosenberg, M. B. Schabath, S. Shete, Z. Song, V. L. Stevens, A. Tardon, H. E. Wichmann, P. Woll, S. Zienolddiny, M. Obeidat, W. Timens, R. J. Hung, P. Joubert, C. I. Amos, and J. D. McKay. Transcriptome-wide association study reveals candidate causal genes for lung cancer. *International Journal of Cancer* 146.7 (2020), 1862–1878.

[360] L. Latonen, K. A. Leinonen, T. Gronlund, R. L. Vessella, T. L. Tammela, O. R. Saramaki, and T. Visakorpi. Amplification of the 9p13.3 chromosomal region in prostate cancer. *Genes Chromosomes & Cancer* 55.8 (2016), 617–25.

[361] D. Q. Calcagno, S. S. Takeno, C. O. Gigek, M. F. Leal, F. Wisnieski, E. S. Chen, T. M. Araujo, E. M. Lima, M. I. Melaragno, S. Demachki, P. P. Assumpcao, R. R. Burbano, and M. C. Smith. Identification of il11ra and melk amplification in gastric cancer by comprehensive genomic profiling of gastric cancer cell lines. *World Journal of Gastroenterology* 22.43 (2016), 9506–9514.

[362] M. Guled, S. Myllykangas, J. Frierson H. F., S. E. Mills, S. Knuutila, and E. B. Stelow. Array comparative genomic hybridization analysis of olfactory neuroblastoma. *Mod Pathol* 21.6 (2008), 770–8.

[363] J. H. Rubenstein, A. Tavakkoli, E. Koeppe, P. Ulintz, J. M. Inadomi, H. Morgenstern, H. Appelman, J. M. Scheiman, P. Schoenfeld, V. Metko, and E. M. Stoffel. Family history of colorectal or esophageal cancer in barrett's esophagus and potentially explanatory genetic variants. *Clinical and Translational Gastroenterology* 11.4 (2020), e00151.

[364] X. Guo, W. Liu, Y. Pan, P. Ni, J. Ji, L. Guo, J. Zhang, J. Wu, J. Jiang, X. Chen, Q. Cai, J. Li, J. Zhang, Q. Gu, B. Liu, Z. Zhu, and Y. Yu. Homeobox gene irx1 is a tumor suppressor gene in gastric carcinoma. *Oncogene* 29.27 (2010), 3908–20.

[365] M. M. Kuster, M. A. Schneider, A. M. Richter, S. Richtmann, H. Winter, M. Kriegsmann, S. S. Pullamsetti, T. Stiewe, R. Savai, T. Muley, and R. H. Dammann. Epigenetic inactivation of the tumor suppressor irx1 occurs frequently in lung adenocarcinoma and its silencing is associated with impaired prognosis. *Cancers (Basel)* 12.12 (2020).

[366] G. Kulasekaran, M. Chaineau, V. E. C. Piscopo, F. Verginelli, M. Fotouhi, M. Girard, Y. Tang, R. Dali, R. Lo, S. Stifani, and P. S. McPherson. An arf/rab cascade controls the growth and invasiveness of glioblastoma. *Journal of Cell Biology* 220.2 (2021).

[367] H. Tang, X. Liu, Z. Wang, X. She, X. Zeng, M. Deng, Q. Liao, X. Guo, R. Wang, X. Li, F. Zeng, M. Wu, and G. Li. Interaction of hsa-mir-381 and glioma suppressor lrrc4 is involved in glioma growth. *Brain Research* 1390 (2011), 21–32.

[368] S. Majumder, W. R. Taylor, P. H. Foote, C. K. Berger, C. W. Wu, D. W. Mahoney, W. R. Bamlet, K. N. Burger, N. Postier, J. de la Fuente, K. A. Doering, G. P. Lidgard, H. T. Allawi, G. M. Petersen, S. T. Chari, D. A. Ahlquist, and J. B. Kisiel. High detection rates of pancreatic cancer across stages by plasma assay of novel methylated dna markers and ca19-9. *Clinical Cancer Research* 27.9 (2021), 2523–2532.

[369] X. Lin, L. Zhou, J. Zhong, L. Zhong, R. Zhang, T. Kang, and Y. Wu. Rna-binding protein rbm28 can translocate from the nucleolus to the nucleoplasm to inhibit the transcriptional activity of p53. *Journal of Biological Chemistry* 298.2 (2021), 101524.

[370] I. P. Tomlinson and W. F. Bodmer. Chromosome 11q in sporadic colorectal carcinoma: patterns of allele loss and their significance for tumorigenesis. *J Clin Pathol* 49.5 (1996), 386–90.

[371] J. Koreth, C. J. Bakkenist, and J. O. McGee. Allelic deletions at chromosome 11q22-q23.1 and 11q25-qterm are frequent in sporadic breast but not colorectal cancers. *Oncogene* 14.4 (1997), 431–7.

[372] Y. Wang, W. Wu, M. Zhu, C. Wang, W. Shen, Y. Cheng, L. Geng, Z. Li, J. Zhang, J. Dai, H. Ma, L. Chen, Z. Hu, G. Jin, and H. Shen. Integrating expression-related snps into genome-wide gene- and pathway-based analyses identified novel lung cancer susceptibility genes. *International Journal of Cancer* 142.8 (2018), 1602–1610.

[373] S. C. Kim, Y. K. Shin, Y. A. Kim, S. G. Jang, and J. L. Ku. Identification of genes inducing resistance to ionizing radiation in human rectal cancer cell lines: re-sensitization of radio-resistant rectal cancer cells through down regulating ndrg1. *BMC Cancer* 18.1 (2018), 594.

[374] J. A. Lees, M. Messa, E. W. Sun, H. Wheeler, F. Torta, M. R. Wenk, P. De Camilli, and K. M. Reinisch. Lipid transport by tmem24 at er-plasma membrane contacts regulates pulsatile insulin secretion. *Science* 355.6326 (2017).

[375] R. A. Saxton and D. M. Sabatini. Mtor signaling in growth, metabolism, and disease. *Cell* 169.2 (2017), 361–371.

[376] M. J. Sanaei, A. Baghery Saghchy Khorasani, A. Pourbagheri-Sigaroodi, S. Shahrokh, M. R. Zali, and D. Bashash. The pi3k/akt/mtor axis in colorectal cancer: oncogenic alterations, non-coding rnas, therapeutic opportunities, and the emerging role of nanoparticles. *Journal of Cell Physiology* (2021).

[377] H. Kobayashi, K. A. Gieniec, T. R. M. Lannagan, T. Wang, N. Asai, Y. Mizutani, T. Iida, R. Ando, E. M. Thomas, A. Sakai, N. Suzuki, M. Ichinose, J. A. Wright, L. Vrbanac, J. Q. Ng, J. Goyne, G. Radford, M. J. Lawrence, T. Sammour, Y. Hayakawa, S. Klebe, A. E. Shin, S. Asfaha, M. L. Bettington, F. Rieder, N. Arpaia, T. Danino, L. M. Butler, A. D. Burt, S. J. Leedham, A. K. Rustgi, S. Mukherjee, M. Takahashi, T. C. Wang, A. Enomoto, S. L. Woods, and D. L. Worthley. The origin and contribution of cancer-associated fibroblasts in colorectal carcinogenesis. *Gastroenterology* 162.3 (2022), 890–906.

[378] Y. Choi, J. Nam, D. J. Whitcomb, Y. S. Song, D. Kim, S. Jeon, J. W. Um, S. G. Lee, J. Woo, S. K. Kwon, Y. Li, W. Mah, H. M. Kim, J. Ko, K. Cho, and E. Kim. Salm5 trans-synaptically interacts with lar-rptps in a splicing-dependent manner to regulate synapse development. *Scientific Reports* 6 (2016), 26676.

[379] D. R. de Bruijn, A. H. van Dijk, R. Pfundt, A. Hoischen, G. F. Merkx, G. A. Gradek, H. Lybaek, A. Stray-Pedersen, H. G. Brunner, and G. Houge. Severe progressive autism associated with two de novo changes: a 2.6-mb 2q31.1 deletion and a balanced t(14;21)(q21.1;p11.2) translocation with long-range epigenetic silencing of lrfn5 expression. *Molecular Syndromology* 1.1 (2010), 46–57.

[380] L. Ying, A. Sharma, A. Chhoda, N. Ruzgar, N. Hasan, R. Kwak, C. L. Wolfgang, T. H. Wang, J. W. Kunstman, R. R. Salem, L. D. Wood, C. Iacobuzio-Donahue, E. B. Schneider, J. J. Farrell, and N. Ahuja. Methylation-based cell-free dna signature for early detection of pancreatic cancer. *Pancreas* 50.9 (2021), 1267–1273.

[381] Y. Wu, L. Yu, G. Bi, K. Luo, G. Zhou, and S. Zhao. Identification and characterization of two novel human scan domain-containing zinc finger genes znf396 and znf397. *Gene* 310 (2003), 193–201.

[382] J. A. Sinnott and P. Kraft. Artifact due to differential error when cases and controls are imputed from different platforms. *Hum Genet* 131.1 (2012), 111–9.

[383] K. T. Zondervan and L. R. Cardon. Designing candidate gene and genome-wide case-control association studies. *Nature Protocols* 2.10 (2007), 2492–501.

[384] D. E. Reich and D. B. Goldstein. Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology* 20.1 (2001), 4–16.

[385] C. B. Gilks, S. E. Bear, H. L. Grimes, and P. N. Tsichlis. Progression of interleukin-2 (il-2)-dependent rat t cell lymphoma lines to il-2-independent growth following activation of a gene (gfi-1) encoding a novel zinc finger protein. *Journal of Molecular Cell Biology* 13.3 (1993), 1759–68.

[386] M. Zornig, T. Schmidt, H. Karsunky, A. Grzeschiczek, and T. Moroy. Zinc finger protein gfi-1 cooperates with myc and pim-1 in t-cell lymphomagenesis by reducing the requirements for il-2. *Oncogene* 12.8 (1996), 1789–801.

[387] B. Scheijen, J. Jonkers, D. Acton, and A. Berns. Characterization of pal-1, a common proviral insertion site in murine leukemia virus-induced lymphomas of c-myc and pim-1 transgenic mice. *Journal of Virology* 71.1 (1997), 9–16.

[388] T. Schmidt, M. Zornig, R. Beneke, and T. Moroy. Momulv proviral integrations identified by sup-f selection in tumors from infected myc/pim bitransgenic mice correlate with activation of the gfi-1 gene. *Nucleic Acids Research* 24.13 (1996), 2528–34.

[389] S. Basu, Q. Liu, Y. Qiu, and F. Dong. Gfi-1 represses cdkn2b encoding p15ink4b through interaction with miz-1. *Proceedings of the National Academy of Sciences USA* 106.5 (2009), 1433–8.

[390] C. Esteban-Jurado, M. Vila-Casadesus, P. Garre, J. J. Lozano, A. Pristoupilova, S. Beltran, J. Munoz, T. Ocana, F. Balaguer, M. Lopez-Ceron, M. Cuatrecasas, S. Franch-Exposito, J. M. Pique, A. Castells, A. Carracedo, C. Ruiz-Ponte, A. Abuli, X. Bessa, M. Andreu, L. Bujanda, T. Caldes, and S. Castellvi-Bel. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genetics in Medicine* 17.2 (2015), 131–42.

[391] E. Sharif-Askari, L. Vassen, C. Kosan, C. Khandanpour, M. C. Gaudreau, F. Heyd, T. Okayama, J. Jin, M. E. Rojas, H. L. Grimes, H. Zeng, and T. Moroy. Zinc finger protein gfi1 controls the endotoxin-mediated toll-like receptor inflammatory response by antagonizing nf-kappab p65. *Journal of Molecular Cell Biology* 30.16 (2010), 3929–42.

[392] W. Xing, Y. Xiao, X. Lu, H. Zhu, X. He, W. Huang, E. S. Lopez, J. Wong, H. Ju, L. Tian, F. Zhang, H. Xu, S. D. Wang, X. Li, M. Karin, and H. Ren. Gfi1 downregulation promotes inflammation-linked metastasis of colorectal cancer. *Cell Death & Differentiation* 24.5 (2017), 929–943.

[393] N. Ashour, J. C. Angulo, A. Gonzalez-Corpas, M. J. Orea, M. V. T. Lobo, R. Colomer, B. Colas, M. Esteller, and S. Ropero. Epigenetic regulation of gfi1 in endocrine-related cancers: a role regulating tumor growth. *International Journal of Molecular Sciences* 21.13 (2020).

[394] P. A. Northcott, C. Lee, T. Zichner, A. M. Stutz, S. Erkek, D. Kawauchi, D. J. Shih, V. Hovestadt, M. Zapatka, D. Sturm, D. T. Jones, M. Kool, M. Remke, F. M. Cavalli, S. Zuyderduyn, G. D. Bader, S. VandenBerg, L. A. Esparza, M. Ryzhova, W. Wang, A. Wittmann, S. Stark, L. Sieber, H. Seker-Cin, L. Linke, F. Kratochwil, N. Jager, I. Buchhalter, C. D. Imbusch, G. Zipprich, B. Raeder, S. Schmidt, N. Diessl, S. Wolf, S. Wiemann, B. Brors, C. Lawerenz, J. Eils, H. J. Warnatz, T. Risch, M. L. Yaspo, U. D. Weber, C. C. Bartholomae, C. von Kalle, E. Turanyi, P. Hauser, E. Sanden, A. Darabi, P. Siesjo, J. Sterba, K. Zitterbart, D. Sumerauer, P. van Sluis, R. Versteeg, R. Volckmann, J. Koster, M. U. Schuhmann, M. Ebinger, H. L. Grimes, G. W. Robinson, A. Gajjar, M. Mynarek, K. von Hoff, S. Rutkowski, T. Pietsch, W. Scheurlen, J. Felsberg, G. Reifenberger, A. E. Kulozik, A. von Deimling, O. Witt, R. Eils, R. J. Gilbertson, A. Korshunov, M. D. Taylor, P. Lichter, J. O. Korbel, R. J. Wechsler-Reya, and S. M. Pfister. Enhancer hijacking activates gfi1 family oncogenes in medulloblastoma. *Nature* 511.7510 (2014), 428–34.

[395] X. Kuai, L. Li, R. Chen, K. Wang, M. Chen, B. Cui, Y. Zhang, J. Li, H. Zhu, H. Zhou, J. Huang, J. Qin, Z. Wang, W. Wei, and D. Gao. Scf(fbxw7)/gsk3beta-mediated gfi1 degradation suppresses proliferation of gastric cancer cells. *Cancer Research* 79.17 (2019), 4387–4398.

[396] Y. Huang, R. Ruan, Y. Fang, K. Wu, L. Yao, R. Zhang, and W. He. Gfi1 promotes the proliferation and migration of esophageal squamous cell carcinoma cells through the inhibition of socs1 expression. *International Journal of Molecular Medicine* 48.4 (2021).

[397] R. E. Person, F. Q. Li, Z. Duan, K. F. Benson, J. Wechsler, H. A. Papadaki, G. Eliopoulos, C. Kaufman, S. J. Bertolone, B. Nakamoto, T. Papayannopoulou, H. L. Grimes, and M. Horwitz. Mutations in proto-oncogene gfi1 cause human neutropenia and target ela2. *Nature Genetics* 34.3 (2003), 308–12.

[398] B. Dave, T. Hsu, W. Hong, and S. Pathak. Nonrandom distribution of mutagen-induced chromosome breaks in lymphocytes of patients with different malignancies. *International Journal of Oncology* 5.4 (1994), 733–40.

[399] B. J. Dave, V. L. Hopwood, T. M. King, H. Jiang, M. R. Spitz, and S. Pathak. Genetic susceptibility to lung cancer as determined by lymphocytic chromosome analysis. *Cancer Epidemiology, Biomarkers & Prevention* 4.7 (1995), 743–9.

[400] N. A. Heerema, H. N. Sather, M. G. Sensel, P. Kraft, J. B. Nachman, P. G. Steinherz, B. J. Lange, R. S. Hutchinson, G. H. Reaman, M. E. Trigg, D. C. Arthur, P. S. Gaynon, and F. M. Uckun. Frequency and clinical significance of cytogenetic abnormalities in pediatric t-lineage acute lymphoblastic leukemia: a report from the children's cancer group. *J Clin Oncol* 16.4 (1998), 1270–8.

[401] Z. Kan, B. S. Jaiswal, J. Stinson, V. Janakiraman, D. Bhatt, H. M. Stern, P. Yue, P. M. Haverty, R. Bourgon, J. Zheng, M. Moorhead, S. Chaudhuri, L. P. Tomsho, B. A. Peters, K. Pujara, S. Cordes, D. P. Davis, V. E. Carlton, W. Yuan, L. Li, W. Wang, C. Eigenbrot, J. S. Kaminker, D. A. Eberhard, P. Waring, S. C. Schuster, Z. Modrusan, Z. Zhang, D. Stokoe, F. J. de Sauvage, M. Faham, and S. Seshagiri. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* 466.7308 (2010), 869–73.

[402] H. Liang, L. W. Cheung, J. Li, Z. Ju, S. Yu, K. Stemke-Hale, T. Dogruluk, Y. Lu, X. Liu, C. Gu, W. Guo, S. E. Scherer, H. Carter, S. N. Westin, M. D. Dyer, R. G. Verhaak, F. Zhang, R. Karchin, C. G. Liu, K. H. Lu, R. R. Broaddus, K. L. Scott, B. T. Hennessy, and G. B. Mills. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. *Genome Research* 22.11 (2012), 2120–9.

[403] J. Zhang, X. Qin, Q. Sun, H. Guo, X. Wu, F. Xie, Q. Xu, M. Yan, J. Liu, Z. Han, and W. Chen. Transcriptional control of pax4-regulated mir-144/451 modulates metastasis by suppressing adams expression. *Oncogene* 34.25 (2015), 3283–95.

[404] N. Tsuchiya, M. Ochiai, K. Nakashima, T. Ubagai, T. Sugimura, and H. Nakagama. Snd1, a component of rna-induced silencing complex, is up-regulated in human colon cancers and implicated in early stage colon carcinogenesis. *Cancer Research* 67.19 (2007), 9568–76.

[405] N. Tsuchiya and H. Nakagama. Microrna, snd1, and alterations in translational regulation in colon carcinogenesis. *Mutation Research* 693.1-2 (2010), 94–100.

[406] O. Levi-Galibov, H. Lavon, R. Wassermann-Dozorets, M. Pevsner-Fischer, S. Mayer, E. Wershof, Y. Stein, L. E. Brown, W. Zhang, G. Friedman, R. Nevo, O. Golani, L. H. Katz, R. Yaeger, I. Laish, J. A. Porco, E. Sahai, D. S. Shouval, D. Kelsen, and R. Scherz-Shouval. Heat shock factor 1-dependent extracellular matrix remodeling mediates the transition from chronic intestinal inflammation to colon cancer. *Nature Communications* 11.1 (2020), 6245.

[407] S. Saha, A. Bardelli, P. Buckhaults, V. E. Velculescu, C. Rago, B. St Croix, K. E. Romans, M. A. Choti, C. Lengauer, K. W. Kinzler, and B. Vogelstein. A phosphatase associated with metastasis of colorectal cancer. *Science* 294.5545 (2001), 1343–6.

[408] L. Xu, R. B. Corcoran, J. W. Welsh, D. Pennica, and A. J. Levine. Wisp-1 is a wnt-1- and beta-catenin-responsive oncogene. *Genes & Development* 14.5 (2000), 585–95.

[409] D. Pennica, T. A. Swanson, J. W. Welsh, M. A. Roy, D. A. Lawrence, J. Lee, J. Brush, L. A. Taneyhill, B. Deuel, M. Lew, C. Watanabe, R. L. Cohen, M. F. Melhem, G. G. Finley, P. Quirke, A. D. Goddard, K. J. Hillan, A. L. Gurney, D. Botstein, and A. J. Levine. Wisp genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. *Proceedings of the National Academy of Sciences USA* 95.25 (1998), 14717–22.

[410] J. Jen, H. Kim, S. Piantadosi, Z. F. Liu, R. C. Levitt, P. Sistonen, K. W. Kinzler, B. Vogelstein, and S. R. Hamilton. Allelic loss of chromosome 18q and prognosis in colorectal cancer. *New England Journal of Medicine* 331.4 (1994), 213–21.

[411] D. Jia, S. M. Hasso, J. Chan, D. Filingeri, P. A. D'Amore, L. Rice, C. Pampo, D. W. Siemann, D. Zurakowski, S. J. Rodig, and M. A. Moses. Transcriptional repression of vegf by znf24: mechanistic studies and vascular consequences in vivo. *Blood* 121.4 (2013), 707–15.

[412] X. Liu, X. Ge, Z. Zhang, X. Zhang, J. Chang, Z. Wu, W. Tang, L. Gan, M. Sun, and J. Li. Microrna-940 promotes tumor cell invasion and metastasis by downregulating znf24 in gastric cancer. *Oncotarget* 6.28 (2015), 25418–28.

[413] B. Pang, Y. Wang, and X. Chang. A novel tumor suppressor gene, znf24, inhibits the development of nsclc by inhibiting the wnt signaling pathway to induce cell senescence. *Frontiers in Oncology* 11 (2021), 664369.

[414] J. Xiong, P. Jiang, L. Zhong, and Y. Wang. The novel tumor suppressor gene znf24 induces thca cells senescence by regulating wnt signaling pathway, resulting in inhibition of thca tumorigenesis and invasion. *Frontiers in Oncology* 11 (2021), 646511.

[415] F. Liu, L. Yan, Z. Wang, Y. Lu, Y. Chu, X. Li, Y. Liu, D. Rui, S. Nie, and X. H. Metformin therapy and risk of colorectal adenomas and colorectal cancer in type 2 diabetes mellitus patients: a systematic review and meta-analysis. *Oncotarget* 8.9 (2017), 16017–16026.

[416] M. Katoh. Functional and cancer genomics of asxl family members. *British Journal of Cancer* 109.2 (2013), 299–306.

[417] L. K. Su and Y. Qi. Characterization of human mapre genes and their proteins. *Genomics* 71.2 (2001), 142–9.

[418] D. A. Goldspink, J. R. Gadsby, G. Bellett, J. Keynton, B. J. Tyrrell, E. K. Lund, P. P. Powell, P. Thomas, and M. M. Mogensen. The microtubule end-binding protein eb2 is a central regulator of microtubule reorganisation in apico-basal epithelial differentiation. *Journal of Cell Science* 126.Pt 17 (2013), 4000–14.

[419] H. Liu, J. Yue, H. Huang, X. Gou, S. Y. Chen, Y. Zhao, and X. Wu. Regulation of focal adhesion dynamics and cell motility by the eb2 and hax1 protein complex. *Journal of Biological Chemistry* 290.52 (2015), 30771–82.

[420] F. Stenner, H. Liewen, S. Gottig, R. Henschler, N. Markuly, S. Kleber, M. Faust, A. Mischo, S. Bauer, M. Zweifel, A. Knuth, C. Renner, and A. Wadle. Rp1 is a phosphorylation target of ck2 and is involved in cell adhesion. *PLoS One* 8.7 (2013), e67595.

[421] M. Iimori, S. Watanabe, S. Kiyonari, K. Matsuoka, R. Sakasai, H. Saeki, E. Oki, H. Kitao, and Y. Maehara. Phosphorylation of eb2 by aurora b and cdk1 ensures mitotic progression and genome stability. *Nature Communications* 7 (2016), 11117.

[422] I. Abiatari, S. Gillen, T. DeOliveira, T. Klose, K. Bo, N. A. Giese, H. Friess, and J. Kleeff. The microtubule-associated protein mapre2 is involved in perineural invasion of pancreatic cancer cells. *International Journal of Oncology* 35.5 (2009), 1111–6.

[423] F. J. Zhong, Y. M. Li, C. Xu, B. Sun, J. L. Wang, and L. Y. Yang. Eb2 promotes hepatocellular carcinoma proliferation and metastasis via mapk/erk pathway by modulating microtubule dynamics. *Clinical Science (Lond)* 135.7 (2021), 847–864.

[424] K. G. M. Moons, R. F. Wolff, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett. Probast: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Annals of Internal Medicine* 170.1 (2019), W1–W33.

[425] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* 12.3 (2015), 1–10.

[426] J. A. Hutcheon, A. Chiolero, and J. A. Hanley. Random measurement error and regression dilution bias. *British Medical Journal* 340 (2010), c2289.

[427] UK Biobank. *UK Biobank Malignant Cancer Summary Report*. 2021. `https://biobank.ndph.ox.ac.uk/~bbdatan/CancerSummaryReport.html`, [Accessed: 18/08/2021].

[428] UK Biobank. *UK Biobank Death Summary Report*. 2021. `https://biobank.ctsu.ox.ac.uk/~bbdatan/DeathSummaryReport.html`, [Accessed: 18/08/2021].

[429] NHS. *Alcohol units*. 2018. `https://www.nhs.uk/live-well/alcohol-support/calculating-alcohol-units/`, [Accessed January 2019].

[430] D. Hoaglin, F. Mosteller, and J. Tukey. *Understanding Robust and Exploratory Data Analysis*. New York: Wiley, 2000.

[431] E. O. Ogundimu, D. G. Altman, and G. S. Collins. Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of clinical epidemiology* 76 (2016), 175–182.

[432] R. D. Riley, K. I. Snell, J. Ensor, D. L. Burke, J. Harrell Frank E, K. G. Moons, and G. S. Collins. Minimum sample size for developing a multivariable prediction model: part ii - binary and time-to-event outcomes. *Statistics in Medicine* 38.7 (2019), 1276–1296.

[433] J. Ensor, E. C. Martin, and R. D. Riley. *pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model*. 2021. [R package version 1.1.0].

[434] R. C. Jinks, P. Royston, and M. K. B. Parmar. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Medical Research Methodology* 15 (2015), 82–82.

[435] P. Royston. Explained variation for survival models. *The Stata Journal* 6.1 (2006), 83–96.

[436] G. S. Collins, E. O. Ogundimu, and D. G. Altman. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine* 35.2 (2016), 214–226.

[437] M. Delgado-Rodriguez and J. Llorca. Bias. *Journal of Epidemiology and Community Health* 58.8 (2004), 635–41.

[438] G. Holden, G. Rosenberg, K. Barker, S. Tuhrim, and B. Brenner. The recruitment of research participants: *Social Work in Health Care* 19.2 (1993), 1–44.

[439] T. A. Manolio, B. K. Weis, C. C. Cowie, R. N. Hoover, K. Hudson, B. S. Kramer, C. Berg, R. Collins, W. Ewart, J. M. Gaziano, S. Hirschfeld, P. M. Marcus, D. Masys, C. A. McCarty, J. McLaughlin, A. V. Patel, T. Peakman, N. L. Pedersen, C. Schaefer, J. A. Scott, T. Sprosen, M. Walport, and F. S. Collins. New Models for Large Prospective Studies: Is There a Better Way? *American Journal of Epidemiology* 175.9 (2012), 859–866.

[440] J. A. Usher-Smith, S. J. Sharp, and S. J. Griffin. The spectrum effect in tests for risk prediction, screening, and diagnosis. *British Medical Journal* 353 (2016), i3139.

[441] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, and S. Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* 50.9 (2018), 1219–1224.

[442] G. Hill, J. Connelly, R. Hebert, J. Lindsay, and W. Millar. Neyman's bias re-visited. *Journal of Clinical Epidemiology* 56.4 (2003), 293–6.

[443] R. D. Riley, J. Ensor, K. I. E. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. M. Moons, G. Collins, and M. van Smeden. Calculating the sample size required for developing a clinical prediction model. *British Medical Journal* 368 (2020).

[444] J. W. Bartlett, O. Harel, and J. R. Carpenter. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American Journal of Epidemiology* 182.8 (2015), 730–6.

[445] Y. Hirst, S. Stoffel, G. Baio, L. McGregor, and C. von Wagner. Uptake of the english bowel (colorectal) cancer screening programme: an update 5 years after the full roll-out. *European Journal of Cancer* (2018).

[446] P. D. P. Pharoah, A. Antoniou, M. Bobrow, R. L. Zimmern, D. F. Easton, and B. A. J. Ponder. Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics* 31 (2002), 33–36.

[447] International Schizophrenia Consortium, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460.7256 (2009), 748–52.

[448] N. Mavaddat, K. Michailidou, J. Dennis, M. Lush, L. Fachal, A. Lee, J. P. Tyrer, T. H. Chen, Q. Wang, M. K. Bolla, X. Yang, M. A. Adank, T. Ahearn, K. Aittomaki, J. Allen, I. L. Andrulis, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, P. L. Auer, P. Auvinen, M. Barrdahl, L. E. Beane Freeman, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A. L. Borresen-Dale, H. Brauch, M. Bremer, H. Brenner, A. Brentnall, I. W. Brock, A. Brooks-Wilson, S. Y. Brucker, T. Bruning, B. Burwinkel, D. Campa, B. D. Carter, J. E. Castelao, S. J. Chanock, R. Chlebowski, H. Christiansen, C. L. Clarke, J. M. Collee, E. Cordina-Duverger, S. Cornelissen, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, T. Dork, I. Dos-Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. H. Eliassen, C. Ellberg, C. Engel, M. Eriksson, D. G. Evans, P. A. Fasching,

J. Figueroa, O. Fletcher, H. Flyger, A. Forsti, L. Fritschi, M. Gabrielson, M. Gago-Dominguez, S. M. Gapstur, J. A. Garcia-Saenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, I. R. Gilyazova, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. Gonzalez-Neira, G. I. Grenaker Alnaes, M. Grip, J. Gronwald, A. Grundy, P. Guenel, L. Haeberle, E. Hahnen, C. A. Haiman, N. Hakansson, U. Hamann, S. E. Hankinson, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *American Journal of Human Genetics* 104.1 (2019), 21–34.

[449] N. Pashayan, S. W. Duffy, S. Chowdhury, T. Dent, H. Burton, D. E. Neal, D. F. Easton, R. Eeles, and P. Pharoah. Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. *British Journal Of Cancer* 104 (2011), 1656–1663.

[450] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* 33.2 (2003), 177–82.

[451] T. B. Bigdeli, D. Lee, B. T. Webb, B. P. Riley, V. I. Vladimirov, A. H. Fanous, K. S. Kendler, and S.-A. Bacanu. A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. *Bioinformatics* 32.17 (2016), 2598–2603.

[452] H. Zhong and R. L. Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* 9.4 (2008), 621–634.

[453] J. Shi, J.-H. Park, J. Duan, S. T. Berndt, W. Moy, K. Yu, L. Song, W. Wheeler, X. Hua, D. Silverman, M. Garcia-Closas, C. A. Hsiung, J. D. Figueroa, V. K. Cortessis, N. Malats, M. R. Karagas, P. Vineis, I.-S. Chang, D. Lin, B. Zhou, A. Seow, K. Matsuo, Y.-C. Hong, N. E. Caporaso, B. Wolpin, E. Jacobs, G. M. Petersen, A. P. Klein, D. Li, H. Risch, A. R. Sanders, L. Hsu, R. E. Schoen, H. Brenner, MGS (Molecular Genetics of Schizophrenia) GWAS Consortium, GECCO (The Genetics and Epidemiology of Colorectal Cancer Consortium), The GAME-ON/TRICL (Transdisciplinary Research in Cancer of the Lung) GWAS Consortium, PRACTICAL (PRostate cancer AssoCiation group To Investigate Cancer Associated aLterations) Consortium, PanScan Consortium, The GAME-ON/ELLIPSE Consortium, R. Stolzenberg-Solomon, P. Gejman, Q. Lan, N. Rothman, L. T. Amundadottir, M. T. Landi, D. F. Levinson, S. J. Chanock, and N. Chatterjee. Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLOS Genetics* 12.12 (2016), e1006493.

[454] Office for National Statistics. *Number of people registered with colorectal cancer and number of deaths caused by colorectal cancer in England*. 2019.

[455] Office for National Statistics. *Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland*. 2019.

[456] National Cancer Institute and Information Management Services. *DevCan: Probability of Developing or Dying of Cancer Software*. 2009. Version 6.7.6.

[457] S. W. Choi and P. F. O'Reilly. Prsice-2: polygenic risk score software for biobank-scale data. *Gigascience* 8.7 (2019).

[458] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G. B. Blum. Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics* 105.6 (2019), 1213–1221.

[459] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* 69.1 (2001), 1–14.

[460] F. Privé, H. Aschard, and M. G. B. Blum. Efficient implementation of penalized regression for genetic risk prediction. *Genetics* 212.1 (2019), 65–74.

[461] L. R. Lloyd-Jones, J. Zeng, J. Sidorenko, L. Yengo, G. Moser, K. E. Kemper, H. W. Wang, Z. L. Zheng, R. Magi, T. Esko, A. Metspalu, N. R. Wray, M. E. Goddard, J. Yang, and P. M. Visscher. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications* 10 (2019).

[462] F. Privé. Ldpred2 code. 2020. `https://github.com/privefl/paper-ldpred2/tree/master/code`, [Accessed Oct-2020].

[463] F. E. Harrell, K. L. Lee, and D. B. Mark. Tutorial in biostatistics: multivariable prognostic models. *Statistics in Medicine* 15 (1996), 361–387.

[464] J. Elliott, B. Bodinier, T. A. Bond, M. Chadeau-Hyam, E. Evangelou, K. G. M. Moons, A. Dehghan, D. C. Muller, P. Elliott, and I. Tzoulaki. Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA* 323.7 (2020), 636–645.

[465] A. C. Justice, K. E. Covinsky, and J. A. Berlin. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 130.6 (1999), 515–24.

[466] D. G. Altman and P. Royston. What do we mean by validating a prognostic model? *Statistics in Medicine* 19.4 (), 453–473.

[467] A. N. Archambault, Y.-R. Su, J. Jeon, M. Thomas, Y. Lin, D. V. Conti, A. K. Win, L. C. Sakoda, I. Lansdorp-Vogelaar, E. F. Peterse, A. G. Zauber, D. Duggan, A. N. Holowatyj, J. R. Huyghe, H. Brenner, M. Cotterchio, S. Bézieau, S. L. Schmit, C. K. Edlund, M. C. Southey, R. J. MacInnis, P. T. Campbell, J. Chang-Claude, M. L. Slattery, A. T. Chan, A. D. Joshi, M. Song, Y. Cao, M. O. Woods, E. White, S. J. Weinstein, C. M. Ulrich, M. Hoffmeister, S. A. Bien, T. A. Harrison, J. Hampe, C. I. Li, C. Schafmayer, K. Offit, P. D. Pharoah, V. Moreno, A. Lindblom, A. Wolk, A. H. Wu, L. Li, M. J. Gunter, A. Gsur, T. O. Keku, R. Pearlman, D. T. Bishop, S. Castellvıé-Bel, L. Moreira, P. Vodicka, E. Kampman, G. G. Giles, D. Albanes, J. A. Baron, S. I. Berndt, S. Brezina, S. Buch, D. D. Buchanan, A. Trichopoulou, G. Severi, M.-D. Chirlaque, M.-J. Sánchez, D. Palli, T. Kühn, N. Murphy, A. J. Cross, A. N. Burnett-Hartman, S. J. Chanock, A. d. l. Chapelle, D. F. Easton, F. Elliott, D. R. English, E. J. Feskens, L. M. FitzGerald, P. J. Goodman, J. L. Hopper, T. J. Hudson, D. J. Hunter, E. J. Jacobs, C. E. Joshu, S. Küry, S. D. Markowitz, R. L. Milne, E. A. Platz, G. Rennert, H. S. Rennert, F. R. Schumacher, R. S. Sandler, D. Seminara, C. M. Tangen, S. N. Thibodeau, A. E. Toland, F. J. van Duijnhoven, K. Visvanathan, L. Vodickova, J. D. Potter, S. Männistö, K. Weigl, J. Figueiredo, V. Martıén, S. C. Larsson, P. S. Parfrey, W.-Y. Huang, H.-J. Lenz, J. E. Castelao, M. Gago-Dominguez, V. Muñoz-Garzón, C. Mancao,

C. A. Haiman, L. R. Wilkens, E. Siegel, E. Barry, B. Younghusband, B. Van Guelpen, S. Harlid, A. Zeleniuch-Jacquotte, P. S. Liang, M. Du, G. Casey, N. M. Lindor, L. Le Marchand, S. J. Gallinger, M. A. Jenkins, P. A. Newcomb, S. B. Gruber, R. E. Schoen, H. Hampel, D. A. Corley, L. Hsu, U. Peters, and R. B. Hayes. Cumulative burden of colorectal cancer-associated genetic variants is more strongly associated with early-onset vs late-onset cancer. *Gastroenterology* (2019), S0016-5085(19)41937-9.

[468] S. Li. Negative age-dependence of the polygenic risk score gradient for colorectal cancer. *Gastroenterology* 160.6 (2021), 2214–2215.

[469] M. Thomas, L. C. Sakoda, M. Hoffmeister, E. A. Rosenthal, J. K. Lee, F. J. B. van Duijnhoven, E. A. Platz, A. H. Wu, C. H. Dampier, A. de la Chapelle, A. Wolk, A. D. Joshi, A. Burnett-Hartman, A. Gsur, A. Lindblom, A. Castells, A. K. Win, B. Namjou, B. Van Guelpen, C. M. Tangen, Q. He, C. I. Li, C. Schafmayer, C. E. Joshu, C. M. Ulrich, D. T. Bishop, D. D. Buchanan, D. Schaid, D. A. Drew, D. C. Muller, D. Duggan, D. R. Crosslin, D. Albanes, E. L. Giovannucci, E. Larson, F. Qu, F. Mentch, G. G. Giles, H. Hakonarson, H. Hampel, I. B. Stanaway, J. C. Figueiredo, J. R. Huyghe, J. Minnier, J. Chang-Claude, J. Hampe, J. B. Harley, K. Visvanathan, K. R. Curtis, K. Offit, L. Li, L. Le Marchand, L. Vodickova, M. J. Gunter, M. A. Jenkins, M. L. Slattery, M. Lemire, M. O. Woods, M. Song, N. Murphy, N. M. Lindor, O. Dikilitas, P. D. P. Pharoah, P. T. Campbell, P. A. Newcomb, R. L. Milne, R. J. MacInnis, S. Castellvi-Bel, S. Ogino, S. I. Berndt, S. Bezieau, S. N. Thibodeau, S. J. Gallinger, S. H. Zaidi, T. A. Harrison, T. O. Keku, T. J. Hudson, V. Vymetalkova, V. Moreno, V. Martin, V. Arndt, W. Q. Wei, W. Chung, Y. R. Su, R. B. Hayes, E. White, P. Vodicka, G. Casey, S. B. Gruber, R. E. Schoen, A. T. Chan, J. D. Potter, H. Brenner, G. P. Jarvik, D. A. Corley, U. Peters, and L. Hsu. Response to li and hopper. *American Journal of Human Genetics* 108.3 (2021), 527–529.

[470] M. Isgut, J. Sun, A. A. Quyyumi, and G. Gibson. Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later. *Genome Medicine* 13.1 (2021), 13.

[471] A. P. Patel, M. Wang, A. C. Fahed, H. Mason-Suares, D. Brockman, R. Pelletier, S. Amr, K. Machini, M. Hawley, L. Witkowski, C. Koch, A. Philippakis, C. A. Cassa, P. T. Ellinor, S. Kathiresan, K. Ng, M. Lebo, and A. V. Khera. Association of rare pathogenic dna variants for familial hypercholesterolemia, hereditary breast and ovarian cancer syndrome, and lynch syndrome with disease risk in adults according to family history. *JAMA Network Open* 3.4 (2020), e203959.

[472] A. C. Fahed, M. Wang, J. R. Homburger, A. P. Patel, A. G. Bick, C. L. Neben, C. Lai, D. Brockman, A. Philippakis, P. T. Ellinor, C. A. Cassa, M. Lebo, K. Ng, E. S. Lander, A. Y. Zhou, S. Kathiresan, and A. V. Khera. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications* 11.1 (2020), 3635.

[473] M. A. Jenkins, D. D. Buchanan, J. Lai, E. Makalic, G. S. Dite, A. K. Win, M. Clendenning, I. M. Winship, R. B. Hayes, J. R. Huyghe, U. Peters, S. Gallinger, L. L. Marchand, J. C. Figueiredo, R. K. Pai, P. A. Newcomb,

J. M. Church, G. Casey, and J. L. Hopper. Assessment of a polygenic risk score for colorectal cancer to predict risk of lynch syndrome colorectal cancer. *JNCI Cancer Spectrum* 5.2 (2021).

[474] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics* 100.4 (2017), 635–649.

[475] L. M. Helsingen, P. O. Vandvik, H. C. Jodal, T. Agoritsas, L. Lytvyn, J. C. Anderson, R. Auer, S. B. Murphy, M. A. Almadi, D. A. Corley, C. Quinlan, J. M. Fuchs, A. McKinnon, A. Qaseem, A. F. Heen, R. A. C. Siemieniuk, M. Kalager, J. A. Usher-Smith, I. Lansdorp-Vogelaar, M. Bretthauer, and G. Guyatt. Colorectal cancer screening with faecal immunochemical testing, sigmoidoscopy or colonoscopy: a clinical practice guideline. *British Medical Journal* 367 (2019), l5515.

[476] P. Royston and W. Sauerbrei. Improving the robustness of fractional polynomial models by preliminary covariate transformation: a pragmatic approach. *Computational Statistics & Data Analysis* 51.9 (2007), 4240–4253.

[477] P. Royston, G. Ambler, and W. Sauerbrei. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 28.5 (1999), 964–74.

[478] A. J. Vickers and E. B. Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26.6 (2006), 565–74.

[479] B. Van Calster, L. Wynants, J. F. M. Verbeek, J. Y. Verbakel, E. Christodoulou, A. J. Vickers, M. J. Roobol, and E. W. Steyerberg. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol* 74.6 (2018), 796–804.

[480] S. G. Pauker and J. P. Kassirer. Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine* 293.5 (1975), 229–34.

[481] A. J. Vickers, A. M. Cronin, E. B. Elkin, and M. Gonen. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making* 8 (2008), 53.

[482] A. J. Vickers, B. Van Calster, and E. W. Steyerberg. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *British Medical Journal* 352 (2016), i6.

[483] A. J. Vickers, B. van Calster, and E. W. Steyerberg. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research* 3 (2019), 18.

[484] D. H. Kim, P. J. Pickhardt, A. J. Taylor, W. K. Leung, T. C. Winter, J. L. Hinshaw, D. V. Gopal, M. Reichelderfer, R. H. Hsu, and P. R. Pfau. Ct colonography versus colonoscopy for the detection of advanced neoplasia. *New England Journal of Medicine* 357.14 (2007), 1403–12.

[485] E. M. Stoop, M. C. de Haan, T. R. de Wijkerslooth, P. M. Bossuyt, M. van Ballegooijen, C. Y. Nio, M. J. van de Vijver, K. Biermann, M. Thomeer, M. E. van Leerdam, P. Fockens, J. Stoker, E. J. Kuipers, and E. Dekker. Participation and yield of colonoscopy versus non-cathartic ct colonography in population-based screening for colorectal cancer: a randomised controlled trial. *Lancet Oncol* 13.1 (2012), 55–64.

[486] L. E. Johns and R. S. Houlston. A systematic review and meta-analysis of familial colorectal cancer risk. *American Journal of Gastroenterology* 96.10 (2001), 2992–3003.

[487] S. J. Li, L. D. Sharples, S. C. Benton, O. Blyuss, C. Mathews, P. Sasieni, and S. W. Duffy. Faecal immunochemical testing in bowel cancer screening: estimating outcomes for different diagnostic policies. *Journal of Medical Screening* 28.3 (2021), 277–285.

[488] S. G. Baker, E. Schuit, E. W. Steyerberg, M. J. Pencina, A. Vickers, K. G. Moons, B. W. Mol, and K. S. Lindeman. How to interpret a small increase in auc with an additional risk prediction marker: decision analysis comes through. *Statistics in Medicine* 33.22 (2014), 3946–59.

[489] A. Sud, C. Turnbull, and R. Houlston. Will polygenic risk scores for cancer ever be clinically useful? *npj Precision Oncology* 5.1 (2021), 40.

[490] Our Future Health. Web Page. 2022.

[491] J. A. Usher-Smith, F. M. Walter, J. D. Emery, A. K. Win, and S. J. Griffin. Risk prediction models for colorectal cancer: a systematic review. *Cancer Prevention Research* 9.1 (2016), 13–26.

[492] J. Hippisley-Cox and C. Coupland. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* 62.594 (2012), e29–37.

[493] G. S. Collins and D. G. Altman. Identifying patients with undetected colorectal cancer: an independent validation of qcancer (colorectal). *British Journal of Cancer* 107.2 (2012), 260–5.

[494] T. Yang, X. Li, Z. Montazeri, J. Little, S. M. Farrington, J. P. Ioannidis, M. G. Dunlop, H. Campbell, M. Timofeeva, and E. Theodoratou. Gene–environment interactions and colorectal cancer risk: an umbrella review of systematic reviews and meta-analyses of observational studies. *International Journal of Cancer* 0.0 ().

[495] T. Yang, X. Li, S. M. Farrington, M. G. Dunlop, H. Campbell, M. Timofeeva, and E. Theodoratou. A systematic analysis of interactions between environmental risk factors and genetic variation in susceptibility to colorectal cancer. *Cancer Epidemiology, Biomarkers & Prevention* 29.6 (2020), 1145–1153.

[496] Y. Tian, A. E. Kim, S. A. Bien, Y. Lin, C. Qu, T. Harrison, R. Carreras-Torres, V. Diez-Obrero, N. Dimou, D. A. Drew, A. Hidaka, J. R. Huyghe, K. M. Jordahl, J. Morrison, N. Murphy, M. Obon-Santacana, C. M. Ulrich, J. Ose, A. R. Peoples, E. A. Ruiz-Narvaez, A. Shcherbina, M. Stern, Y. R. Su, F. J. B. van Duijnhoven, V. Arndt, J. Baurley, S. I. Berndt, D. T. Bishop, H. Brenner, D. D. Buchanan, A. T. Chan, J. C. Figueiredo, S. Gallinger, S. B. Gruber, S. Harlid, M. Hoffmeister, M. A. Jenkins, A. D. Joshi, T. O. Keku, S. C. Larsson,

L. Le Marchand, L. Li, G. G. Giles, R. L. Milne, H. Nan, R. Nassir, S. Ogino, A. Budiarto, E. A. Platz, J. D. Potter, R. L. Prentice, G. Rennert, L. C. Sakoda, R. E. Schoen, M. L. Slattery, S. N. Thibodeau, B. Van Guelpen, K. Visvanathan, E. White, A. Wolk, M. O. Woods, A. H. Wu, P. T. Campbell, G. Casey, D. V. Conti, M. J. Gunter, A. Kundaje, J. P. Lewinger, V. Moreno, P. A. Newcomb, B. Pardamean, D. C. Thomas, K. K. Tsilidis, U. Peters, W. J. Gauderman, L. Hsu, and J. Chang-Claude. Genome-wide interaction analysis of genetic variants with menopausal hormone therapy for colorectal cancer risk. *Journal of the National Cancer Institute* (2022).

[497] X. Chen, L. Jansen, F. Guo, M. Hoffmeister, J. Chang-Claude, and H. Brenner. Smoking, genetic predisposition, and colorectal cancer risk. *Clinical and Translational Gastroenterology* 12.3 (2021), e00317.

[498] J. Choi, G. Jia, W. Wen, X. O. Shu, and W. Zheng. Healthy lifestyles, genetic modifiers, and colorectal cancer risk: a prospective cohort study in the uk biobank. *American Journal of Clinical Nutrition* 113.4 (2021), 810–820.

[499] P. R. Carr, K. Weigl, L. Jansen, V. Walter, V. Erben, J. Chang-Claude, H. Brenner, and M. Hoffmeister. Healthy lifestyle factors associated with lower risk of colorectal cancer irrespective of genetic risk. *Gastroenterology* 155.6 (2018), 1805–1815 e5.

[500] V. Erben, P. R. Carr, F. Guo, K. Weigl, M. Hoffmeister, and H. Brenner. Individual and joint associations of genetic risk and healthy lifestyle score with colorectal neoplasms among participants of screening colonoscopy. *Cancer Prev Res (Phila)* 14.6 (2021), 649–658.

[501] K. Weigl, J. Chang-Claude, P. Knebel, L. Hsu, M. Hoffmeister, and H. Brenner. Strongly enhanced colorectal cancer risk stratification by combining family history and genetic risk score. *Clinical Epidemiology* 10 (2018), 143–152.

[502] M. A. Jenkins, A. K. Win, J. G. Dowty, R. J. MacInnis, E. Makalic, D. F. Schmidt, G. S. Dite, M. Kapuscinski, M. Clendenning, C. Rosty, I. M. Winship, J. D. Emery, S. Saya, F. A. Macrae, D. J. Ahnen, D. Duggan, J. C. Figueiredo, N. M. Lindor, R. W. Haile, J. D. Potter, M. Cotterchio, S. Gallinger, P. A. Newcomb, D. D. Buchanan, G. Casey, and J. L. Hopper. Ability of known susceptibility snps to predict colorectal cancer risk for persons with and without a family history. *Familial Cancer* (2019).

[503] Y. Zheng, X. Hua, A. K. Win, R. J. MacInnis, S. Gallinger, L. L. Marchand, N. M. Lindor, J. A. Baron, J. L. Hopper, J. G. Dowty, A. C. Antoniou, J. Zheng, M. A. Jenkins, and P. A. Newcomb. A new comprehensive colorectal cancer risk prediction model incorporating family history, personal characteristics, and environmental factors. *Cancer Epidemiology, Biomarkers & Prevention* 29.3 (2020), 549–557.

[504] M. C. S. Wong, C. H. Chan, J. Lin, J. L. W. Huang, J. Huang, Y. Fang, W. W. L. Cheung, C. P. Yu, J. C. T. Wong, G. Tse, J. C. Y. Wu, and F. K. L. Chan. Lower relative contribution of positive family history to colorectal cancer risk with increasing age: a systematic review and meta-analysis of 9.28 million individuals. *American Journal of Gastroenterology* 113.12 (2018), 1819–1827.

[505] K. Weigl, H. Thomsen, Y. Balavarca, J. N. Hellwege, M. J. Shrubsole, and H. Brenner. Genetic risk score is associated with prevalence of advanced neoplasms in a colorectal cancer screening population. *Gastroenterology* 155.1 (2018), 88–98 e10.

[506] M. J. Northcutt, Z. Shi, M. Zijlstra, A. Shah, S. Zheng, E. F. Yen, O. Khan, M. I. Beig, P. Imas, A. Vanderloo, O. Ansari, J. Xu, and J. L. Goldstein. Polygenic risk score is a predictor of adenomatous polyps at screening colonoscopy. *BMC Gastroenterology* 21.1 (2021), 65.

[507] F. Guo, X. Chen, J. Chang-Claude, M. Hoffmeister, and H. Brenner. Colorectal cancer risk by genetic variants in populations with and without colonoscopy history. *JNCI Cancer Spectrum* 5.1 (2021).

[508] M. Wolbers, M. T. Koller, J. C. Witteman, and E. W. Steyerberg. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* 20.4 (2009), 555–61.

[509] N. Bradshaw, S. Holloway, I. Penman, M. G. Dunlop, and M. E. Porteous. Colonoscopy surveillance of individuals at risk of familial colorectal cancer. *Gut* 52.12 (2003), 1748–51.

[510] F. C. Tsai and W. B. Strum. Impact of a family history of colorectal cancer on the prevalence of advanced neoplasia at colonoscopy in 4,967 asymptomatic patients. *Digestive Diseases and Sciences* 57.12 (2012), 3234–9.

[511] UK Government. Genome uk: the future of healthcare. `https://www.gov.uk/government/publications/genome-uk-the-future-of-healthcare.`. 2020.

[512] G. J. Annas and S. Elias. 23andme and the fda. *New England Journal of Medicine* 370.11 (2014), 985–8.

[513] Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nature Medicine* 27.11 (2021), 1876–1884.

[514] F. Guo, K. Weigl, P. R. Carr, T. Heisser, L. Jansen, P. Knebel, J. Chang-Claude, M. Hoffmeister, and H. Brenner. Use of polygenic risk scores to select screening intervals after negative findings from colonoscopy. *Clinical Gastroenterology and Hepatology* 18.12 (2020), 2742–2751 e7.

[515] L. Henneman, D. R. Timmermans, C. M. Bouwman, M. C. Cornel, and H. Meijers-Heijboer. 'a low risk is still a risk': exploring women's attitudes towards genetic testing for breast cancer susceptibility in order to target disease prevention. *Public Health Genomics* 14.4-5 (2011), 238–247.

[516] M. Koitsalu, M. A. Sprangers, M. Eklund, K. Czene, P. Hall, H. Gronberg, and Y. Brandberg. Public interest in and acceptability of the prospect of risk-stratified screening for breast and prostate cancer. *Acta Oncologica* 55.1 (2016), 45–51.

[517] Cancer Research UK. Bowel cancer incidence statistics. 2022. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence`, [Accessed 07-07-2022].

[518] A. B. Knudsen, C. M. Rutter, E. F. P. Peterse, A. P. Lietz, C. L. Seguin, R. G. S. Meester, L. A. Perdue, J. S. Lin, R. L. Siegel, V. P. Doria-Rose, E. J. Feuer, A. G. Zauber, K. M. Kuntz, and I. Lansdorp-Vogelaar. Colorectal cancer screening: an updated modeling study for the us preventive services task force. *Journal of the American Medical Association* 325.19 (2021), 1998–2011.

[519] S. Briggs and I. Slade. Evaluating the integration of genomics into cancer screening programmes: challenges and opportunities. *Curr Genetics in Medicine Rep* 7.2 (2019), 63–74.

[520] A. E. Hall, S. Chowdhury, N. Pashayan, N. Hallowell, P. Pharoah, and H. Burton. What ethical and legal principles should guide the genotyping of children as part of a personalised screening programme for common cancer? *Journal of Medical Ethics* 40.3 (2014), 163–167.

[521] PHG Foundation. Implementing polygenic scores for cardiovascular disease into nhs health checks. `https://www.phgfoundation.org/media/491/download/PRS%20Implementation%20Report%2017%20Sept%202021.pdf?v=3&inline=1`. 2021.

[522] E. Dekker. Combining risk factors and faecal immunochemical testing in colorectal cancer screening: a randomized controlled trial. `https://clinicaltrials.gov/ct2/show/NCT04490551`. 2022. Identifier: NCT04490551.

[523] S. Benafif, H. Ni Raghallaigh, E. McGrowder, E. J. Saunders, M. N. Brook, S. Saya, R. Rageevakumar, S. Wakerell, D. James, A. Chamberlain, N. Taylor, M. Hogben, B. Benton, L. D'Mello, K. Myhill, C. Mikropoulos, H. Bowen-Perkins, I. Rafi, M. Ferris, A. Beattie, S. Kuganolipava, T. Sevenoaks, J. Bower, P. Kumar, S. Hazell, N. M. deSouza, A. Antoniou, E. Bancroft, Z. Kote-Jarai, and R. Eeles. The barcode1 pilot: a feasibility study of using germline single nucleotide polymorphisms to target prostate cancer screening. *BJU International* 129.3 (2022), 325–336.

[524] myPeBS. International randomized study comparing personalized, risk-stratified to standard breast cancer screening in women aged 40-70. `https://clinicaltrials.gov/ct2/show/NCT03672331?term=mypebs&rank=1`. 2018.

[525] Y. Shieh, M. Eklund, L. Madlensky, S. D. Sawyer, C. K. Thompson, A. Stover Fiscalini, E. Ziv, L. J. Van't Veer, L. J. Esserman, J. A. Tice, and I. Athena Breast Health Network. Breast cancer screening in the precision medicine era: risk-based screening in a population-based trial. *Journal of the National Cancer Institute* 109.5 (2017).

[526] H. Devlin. Are genetic tests useful to predict cancer? `https://www.theguardian.com/society/2019/mar/23/are-predictive-genetic-test-useful-to-predict-cancer-matt-hancock`. 2019.

[527] Y. Ding, K. Hou, K. S. Burch, S. Lapinska, F. Prive, B. Vilhjalmsson, S. Sankararaman, and B. Pasaniuc. Large uncertainty in individual polygenic risk score estimation impacts prs-based risk stratification. *Nature Genetics* 54.1 (2022), 30–39.

[528] A. C. F. Lewis and R. C. Green. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Medicine* 13.1 (2021), 14.

[529] CanGene-CanVar. Cangene-canvar. `https://www.cangene-canvaruk.org/about-cangene-canvar`. 2022.

[530] G. J. Hollands, D. P. French, S. J. Griffin, A. T. Prevost, S. Sutton, S. King, and T. M. Marteau. The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. *British Medical Journal* 352 (2016), i1102.

[531] K. F. J. Stewart, A. Wesselius, M. A. C. Schreurs, A. Schols, and M. P. Zeegers. Behavioural changes, sharing behaviour and psychological responses after receiving direct-to-consumer genetic test results: a systematic review and meta-analysis. *Journal of Community Genetics* 9.1 (2018), 1–18.

[532] S. Oliveri, F. Ferrari, A. Manfrinati, and G. Pravettoni. A systematic review of the psychological implications of genetic testing: a comparative analysis among cardiovascular, neurodegenerative and cancer diseases. *Frontiers in Genetics* 9 (2018), 624.

[533] A. J. Vickers, A. Sud, J. Bernstein, and R. Houlston. Polygenic risk scores to stratify cancer screening should predict mortality not incidence. *npj Precision Oncology* 6.1 (2022), 1–6.

[534] R. J. Klein, E. Vertosick, D. Sjoberg, D. Ulmert, A. C. Ronn, C. Haggstrom, E. Thysell, G. Hallmans, A. Dahlin, P. Stattin, O. Melander, A. Vickers, and H. Lilja. Prostate cancer polygenic risk score and prediction of lethal prostate cancer. *npj Precision Oncology* 6.1 (2022), 25.

[535] UK Government Department of Health and Social Care. Code on genetic testing and insurance. `https://www.gov.uk/government/publications/code-on-genetic-testing-and-insurance`. 2018.

[536] C. Campbell, A. Douglas, L. Williams, G. Cezard, D. H. Brewster, D. Buchanan, K. Robb, G. Stanners, D. Weller, R. J. Steele, M. Steiner, and R. Bhopal. Are there ethnic and religious variations in uptake of bowel cancer screening? a retrospective cohort study among 1.7 million people in scotland. *BMJ Open* 10.10 (2020), e037011.

[537] J. M. Carethers and C. A. Doubeni. Causes of socioeconomic disparities in colorectal cancer and intervention framework and strategies. *Gastroenterology* 158.2 (2020), 354–367.

[538] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* 51.4 (2019), 584–591.

[539] C. G. Victora, G. Joseph, I. C. M. Silva, F. S. Maia, J. P. Vaughan, F. C. Barros, and A. J. D. Barros. The inverse equity hypothesis: analyses of institutional deliveries in 286 national surveys. *American Journal of Public Health* 108.4 (2018), 464–471.

[540] J. Madhusoodanan. Health-care inequality could deepen with precision oncology. *Nature* 585 (2020), S13–S15.

[541] T. Wang, L. Antonacci-Fulton, K. Howe, H. A. Lawson, J. K. Lucas, A. M. Phillippy, A. B. Popejoy, M. Asri, C. Carson, M. J. P. Chaisson, X. Chang, R. Cook-Deegan, A. L. Felsenfeld, R. S. Fulton, E. P. Garrison, N. A. Garrison, T. A. Graves-Lindsay, H. Ji, E. E. Kenny, B. A. Koenig, D. Li, T. Marschall, J. F. McMichael, A. M. Novak, D. Purushotham, V. A. Schneider, B. I. Schultz, M. W. Smith, H. J. Sofia, T. Weissman, P. Flicek, H. Li, K. H. Miga, B. Paten, E. D. Jarvis, I. M. Hall, E. E. Eichler, D. Haussler, and C. Human Pangenome Reference. The human pangenome project: a global resource to map genomic diversity. *Nature* 604.7906 (2022), 437–446.

[542] E. A. Khramtsova, L. K. Davis, and B. E. Stranger. The role of sex in the genomics of human complex traits. *Nature Reviews Genetics* 20.3 (2019), 173–190.

[543] T. Ge, C.-Y. Chen, B. M. Neale, M. R. Sabuncu, and J. W. Smoller. Phenome-wide heritability analysis of the uk biobank. *PLoS Genetics* 13.4 (2017), e1006711.

[544] V. Boraska, A. Jerončić, V. Colonna, L. Southam, D. R. Nyholt, N. William Rayner, J. R. Perry, D. Toniolo, E. Albrecht, W. Ang, S. Bandinelli, M. Barbalic, I. Barroso, J. S. Beckmann, R. Biffar, D. Boomsma, H. Campbell, T. Corre, J. Erdmann, T. Esko, K. Fischer, N. Franceschini, T. M. Frayling, G. Girotto, J. R. Gonzalez, T. B. Harris, A. C. Heath, I. M. Heid, W. Hoffmann, A. Hofman, M. Horikoshi, J. H. Zhao, A. U. Jackson, J. J. Hottenga, A. Jula, M. Kahonen, K. T. Khaw, L. A. Kiemeney, N. Klopp, Z. Kutalik, V. Lagou, L. J. Launer, T. Lehtimaki, M. Lemire, M. L. Lokki, C. Loley, J. Luan, M. Mangino, I. Mateo Leach, S. E. Medland, E. Mihailov, G. W. Montgomery, G. Navis, J. Newnham, M. S. Nieminen, A. Palotie, K. Panoutsopoulou, A. Peters, N. Pirastu, O. Polasek, K. Rehnstrom, S. Ripatti, G. R. Ritchie, F. Rivadeneira, A. Robino, N. J. Samani, S. Y. Shin, J. Sinisalo, J. H. Smit, N. Soranzo, L. Stolk, D. W. Swinkels, T. Tanaka, A. Teumer, A. Tonjes, M. Traglia, J. Tuomilehto, A. Valsesia, W. H. van Gilst, J. B. van Meurs, A. V. Smith, J. Viikari, J. M. Vink, G. Waeber, N. M. Warrington, E. Widen, G. Willemsen, A. F. Wright, B. W. Zanke, L. Zgaga, C. Wellcome Trust Case Control, M. Boehnke, A. P. d'Adamo, E. de Geus, E. W. Demerath, M. den Heijer, J. G. Eriksson, L. Ferrucci, C. Gieger, V. Gudnason, et al. Genome-wide meta-analysis of common variant differences between men and women. *Human Molecular Genetics* 21.21 (2012), 4805–4815.

[545] S. Stringer, T. J. Polderman, and D. Posthuma. Majority of human traits do not show evidence for sex-specific genetic and environmental effects. *Scientific Reports* 7.1 (2017), 1–7.

[546] J. Sherman, A. MacNeill, and C. Thiel. Reducing pollution from the health care industry. *JAMA* 322.11 (2019), 1043–1044.

[547] National Human Genome Research Institute. Web Page. 2022.

[548] H3Africa Consortium, C. Rotimi, A. Abayomi, A. Abimiku, V. M. Adabayeri, C. Adebamowo, E. Adebiyi, A. D. Ademola, A. Adeyemo, D. Adu, D. Affolabi, G. Agongo, S. Ajayi, S. Akarolo-Anthony, R. Akinyemi, A. Akpalu, M. Alberts, O. Alonso Betancourt, A. M. Alzohairy, G. Ameni, O. Amodu, G. Anabwani, K. Andersen, F. Arogundade, O. Arulogun, D. Asogun, R. Bakare, N. Balde, M. L. Baniecki, C. Beiswanger, A. Benkahla, L. Bethke, M. Boehnke, V. Boima,

J. Brandful, A. I. Brooks, F. C. Brosius, C. Brown, B. Bucheton, D. T. Burke, B. G. Burnett, S. Carrington-Lawrence, N. Carstens, J. Chisi, A. Christoffels, R. Cooper, H. Cordell, N. Crowther, T. Croxton, J. de Vries, L. Derr, P. Donkor, S. Doumbia, A. Duncanson, I. Ekem, A. El Sayed, M. E. Engel, J. C. Enyaru, D. Everett, F. M. Fadlelmola, E. Fakunle, K. H. Fischbeck, A. Fischer, O. Folarin, J. Gamieldien, R. F. Garry, S. Gaseitsiwe, R. Gbadegesin, A. Ghansah, M. Giovanni, P. Goesbeck, F. X. Gomez-Olive, D. S. Grant, R. Grewal, M. Guyer, N. A. Hanchard, C. T. Happi, S. Hazelhurst, B. J. Hennig, C. Hertz, Fowler, W. Hide, F. Hilderbrandt, C. Hugo-Hamman, M. E. Ibrahim, R. James, Y. Jaufeerally-Fakim, C. Jenkins, U. Jentsch, P. P. Jiang, M. Joloba, V. Jongeneel, F. Joubert, M. Kader, K. Kahn, P. Kaleebu, S. H. Kapiga, S. K. Kassim, I. Kasvosve, J. Kayondo, et al. Research capacity. enabling the genomic revolution in africa. *Science* 344.6190 (2014), 1346–8.

[549] Y. Ruan, Y. F. Lin, Y. A. Feng, C. Y. Chen, M. Lam, Z. Guo, I. Stanley Global Asia, L. He, A. Sawa, A. R. Martin, S. Qin, H. Huang, and T. Ge. Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics* 54.5 (2022), 573–580.

[550] O. Weissbrod, M. Kanai, H. Shi, S. Gazal, W. J. Peyrot, A. V. Khera, Y. Okada, P. Biobank Japan, A. R. Martin, H. K. Finucane, and A. L. Price. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nature Genetics* 54.4 (2022), 450–458.

[551] D. Marnetto, K. Pärna, K. Läll, L. Molinaro, F. Montinaro, T. Haller, M. Metspalu, R. Mägi, K. Fischer, and L. Pagani. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nature Communications* 11.1 (2020), 1–9.

[552] J. M. Cairns, S. Greenley, O. Bamidele, and D. Weller. A scoping review of risk-stratified bowel screening: current evidence, future directions. *Cancer Causes & Control* 33.5 (2022), 653–685.

[553] A. M. Thomas, P. Manghi, F. Asnicar, E. Pasolli, F. Armanini, M. Zolfo, F. Beghini, S. Manara, N. Karcher, C. Pozzi, S. Gandini, D. Serrano, S. Tarallo, A. Francavilla, G. Gallo, M. Trompetto, G. Ferrero, S. Mizutani, H. Shiroma, S. Shiba, T. Shibata, S. Yachida, T. Yamada, J. Wirbel, P. Schrotz-King, C. M. Ulrich, H. Brenner, M. Arumugam, P. Bork, G. Zeller, F. Cordero, E. Dias-Neto, J. C. Setubal, A. Tett, B. Pardini, M. Rescigno, L. Waldron, A. Naccarati, and N. Segata. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine* 25.4 (2019), 667–678.

[554] C. L. Andaur Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, and L. Hooft. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *British Medical Journal* 375 (2021), n2281.

[555] J. K. Min, H. J. Yang, M. S. Kwak, C. W. Cho, S. Kim, K. S. Ahn, S. K. Park, J. M. Cha, and D. I. Park. Deep neural network-based prediction of the risk of advanced colorectal neoplasia. *Gut and Liver* 15.1 (2021), 85–91.

[556] P. Dixon, E. Keeney, J. C. Taylor, S. Wordsworth, and R. M. Martin. Can polygenic risk scores contribute to cost-effective cancer screening? a systematic review. *Genetics in Medicine* (2022).

[557] C. Thomas, O. Mandrik, C. L. Saunders, D. Thompson, S. Whyte, S. Griffin, and J. A. Usher-Smith. The costs and benefits of risk stratification for colorectal cancer screening based on phenotypic and genetic risk: a health economic analysis. *Cancer Prev Res (Phila)* 14.8 (2021), 811–822.

[558] S. K. Naber, S. Kundu, K. M. Kuntz, W. D. Dotson, M. S. Williams, A. G. Zauber, N. Calonge, D. T. Zallen, T. G. Ganiats, E. M. Webber, K. A. B. Goddard, N. B. Henrikson, M. van Ballegooijen, A. Cecile, J. W. Janssens, and I. Lansdorp-Vogelaar. Cost-effectiveness of risk-stratified colorectal cancer screening based on polygenic risk – current status and future potential. *JNCI Cancer Spectrum* (2019). pkz086.

[559] D. R. Cenin, S. K. Naber, A. C. de Weerdt, M. A. Jenkins, D. B. Preen, H. C. Ee, P. C. O'Leary, and I. Lansdorp-Vogelaar. Cost-effectiveness of personalized screening for colorectal cancer based on polygenic risk and family history. *Cancer epidemiology, biomarkers & prevention* 29.1 (2020), 10–21.