Prediction of Colorectal Cancer Risk Based on Profiling with Common Genetic Variants

TITLE Prediction of Colorectal Cancer Risk Based on Profiling with Common Genetic Variants

AUTHORS Xue Li[1,2], Maria Timofeeva[3,4,5], Athina Spiliopoulou[6], Paul McKeigue[6], Yazhou He[1,4], Xiaomeng Zhang[1], Victoria Svinti[3,4], Harry Campbell[1], Richard S Houlston[7], Ian PM Tomlinson[4], Susan M Farrington[3,4], Malcolm G Dunlop[3,4§], Evropi Theodoratou[1, 4§]

1 Centre for Global Health, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
2 School of Public Health, Zhejiang University, Hangzhou, China
3 Colon Cancer Genetics Group, Cancer Research UK Edinburgh Centre and Medical Research Council Human Genetics Unit, Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom
4 Cancer Research UK Edinburgh Centre, Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom
5 DIAS, Danish Institute for Advanced Study, Department of Public Health, University of Southern Denmark, Odense, Denmark
6 Centre for Population Health Sciences, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
7 Division of Genetics and Epidemiology, the Institute of Cancer Research, London, SW7 3RP, UK

§ Corresponding authors
Evropi Theodoratou, Centre for Global Health, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom, e.theodoratou@ed.ac.uk, (+44) 0131 650 3210
Malcolm G Dunlop, Institute of Genetics and Molecular Medicine, University of Edinburgh, United Kingdom, Malcolm.Dunlop@igmm.ed.ac.uk, (+44) 0131 537 1547

Keywords: colorectal cancer, genome-wide association study, polygenic risk score, genetic prediction

Abbreviations: BMI, body mass index; CCRR1, Colon Cancer Family Registry 1; CCFR2, Colon Cancer Family Registry 2; CI: confidence interval; COIN, Continuous or Intermittent trial; CORSA, Colorectal cancer Study of Austria; CRC, colorectal cancer; CRP, C-reactive protein; DACHS, Darmkrebs Chancen der Verhütung durch Screening; FIN, Finnish Colorectal Cancer Predisposition Study; GWAS, genome-wide association study; IBD, inflammatory bowel disease; LD, linkage disequilibrium; NSCCG, National Study of Colorectal Cancer Genetics; OR, odds ratios; PC, principal components; PRS, polygenic risk scores; QC, quality control; ROC, receiver-operating characteristic curve; SCOT, Short Course Oncology Treatment trial; SD, standard deviation; SNP, single nucleotide polymorphism; SOCCS, Study of Colorectal Cancer in Scotland; UKBB, UK Biobank; VD, vitamin D; WHR, waist hip rate.

Novelty and Impact

This study shows that a weighted genomic risk score including 116 CRC susceptibility SNPs is the score with the best prediction performance, while deconstructing genetic risk into multiple regional scores or

inclusion of additional SNPs above the genome-wide significance threshold showed no further improvement on prediction performance. Modelling the levels of PRS with age and sex in the general UK population shows that employing genetic risk profiling can achieve a moderate degree of risk discrimination that could be helpful to identify a sub-population with higher CRC risk due to genetic susceptibility.

ABSTRACT

Increasing numbers of common genetic variants associated with colorectal cancer (CRC) have been identified. This study aimed to determine whether risk prediction based on common genetic variants might enable stratification for CRC risk. Meta-analysis of eleven genome-wide association studies (GWAS) comprising 16,871 cases and 26,328 controls was performed to capture CRC susceptibility variants. Genetic prediction models with several candidate polygenic risk scores (PRSs) were generated from Scottish CRC case–control studies (6478 cases and 11,043 controls), and the score with the best performance was then tested in UK Biobank (4800 cases and 20,287 controls). A weighted PRS of 116 CRC SNPs ($wPRS_{116}$) was found with the best predictive performance, reporting a c-statistics of 0.60 and an odds ratio (OR) of 1.46 (95%CI: 1.41-1.50, per SD increase) in Scottish dataset. The predictive performance of this $wPRS_{116}$ was consistently validated in UK Biobank dataset with c-statistics of 0.61 and an OR of 1.49 (95%CI: 1.44-1.54, per SD increase). Modelling the levels of PRS with age and sex in the general UK population shows that employing genetic risk profiling can achieve a moderate degree of risk discrimination that could be helpful to identify a subpopulation with higher CRC risk due to genetic susceptibility.

INTRODUCTION

Colorectal cancer (CRC) is one of the most common cancers, with 1.8 million new cases and almost 0.8 million deaths globally in 2018.[1] Substantial evidence showed that screening can reduce CRC mortality by allowing early detection and removal of precancerous lesions. Policy makers and clinicians rely on risk classification to determine which individuals to screen. To date, these classification schemes are predominantly based on age and/or a simple classification of family history. Stratifying the average risk population into risk categories offers the potential of tailoring surveillance intensity.

Comprehensive information on genetic susceptibility could contribute importantly to CRC risk stratification, given that the heritability of CRC has been estimated to be around 16%-35%[2] and the sibling recurrence risk ratio is about 2.0.[3] We previously assessed the utility of CRC genetic risk profiling with a panel of 10 common genetic variants associated with CRC susceptibility.[4] Although discrimination ability was low (c-statistic of 0.56), we showed that genotype data provides additional information to that from family history alone.[4] Others have also showed that personalised screening using polygenic risk scores (PRSs) have the potential to identify high-risk subgroups most likely to benefit from targeted CRC prevention strategies.[5] Incorporating more complete genetic information is expected to improve risk stratification and the combined effect of multiple risk loci has the potential to achieve a degree of risk discrimination that is useful for CRC risk stratification.

In this study, we aimed to derive, optimize and test polygenic risk scores (PRSs) for prediction of CRC and to apply the PRSs with the best predictive performance in population settings for risk stratification. We developed models by incorporating genetic information of CRC and several markers that comprise potential CRC risk factors or complex traits co-occurring with CRC. To gauge the broader future potential of genetic risk modelling, we assessed the utility of genetic risk scores in categorizing risk subgroups within the general population by projecting the risk models to the UK population.

MATERIALS AND METHODS

Studies

We made use of 11 previously published GWASs (i.e., CCRR1,[6] CCFR2,[7] COIN,[8] CORSA,[9] Croatia,[10] DACHS,[11] FIN,[12] NSCCG-OncoArray,[13] SCOT,[14] UK1,[15] and VQ58 [16]) to generate a list of genetic variants associated with CRC risk. A series of Scottish CRC case–control studies were used to test the predictive performance of polygenic risk scores (PRSs). The developed PRSs were further evaluated in an independent test dataset from UK Biobank. Schematic representation of the study design is shown in Supporting Information Fig. S1. Standard quality control (QC) measures were applied to each of the datasets. After QC process, a total of 16,871 cases and 26,328 controls were finally included for the derivation of genetic susceptibility SNPs, 6,478 cases and 11,043 controls from Scottish dataset were

included for the development of PRSs, and 4,800 cases and 20,287 controls from UK Biobank were included to validate the predictive performance of the PRSs developed. Details are described in Supporting Information Methods and Table S1.

Polygenic risk scores

Genome-wide polygenic score: We performed a meta-GWAS of 11 studies to obtain a list of genome-wide significant SNPs ($p<5\times10^{-8}$) and their per-allele odds ratios (ORs) and standard errors for CRC risk. The meta-analysis SNPs were pruned to only those with an $r^2 < 0.1$ and a distance greater than 500 kb. For completeness, we also included the genetic risk variants reported in early published CRC GWASs (Table S2). A weighted genome-wide PRS (*w*PRS) was computed using both previously known susceptibility variants and independent variants identified by the meta-GWAS.

Regional genetic scores: We additionally constructed regional genetic scores by including SNPs associated with CRC and its risk factors (i.e., vitamin D [VD], C-reactive protein [CRP], body mass index [BMI], waist hip rate [WHR], and inflammatory bowel disease [IBD]) by using the GENOSCORES library (https://pm2.phs.ed.ac.uk/genoscores/). This is similar to the approach used for LDpred,[17] in which the correction for linkage disequilibrium (LD) between SNPs was based on pre-multiplying the vector of weights by the generalized inverse of the correlation matrix estimated from 1000G reference panel of European ancestry.

Model development and evaluation

We constructed prediction models in the Scottish dataset by incorporating genetic CRC risk in forms of either PRSs or regional genetic scores with adjustment for the first 10 genetic principal components (PCs). A sequence of logistic models was fitted for: (i) a weighted PRS of identified CRC GWAS SNPs; (ii) regional genetic scores for CRC; and (iii) regional genetic scores for CRC and other relevant traits. A series of stepwise backward logistic regressions was conducted on regional genetic scores to obtain an optimized set of scores determined by the Akaike information criterion. The discriminatory accuracy of the models was evaluated by the area under the receiver-operating characteristic curve (ROC, known as c-statistic) with 10-fold cross-validation. These models were further assessed by the stratification of anatomic tumor sites (i.e., proximal colon, distal colon and rectum). The PRS model with the best performance was then evaluated in UK Biobank. Odds ratios (OR) were then derived for per SD increase in PRS for overall, and site-specific, CRC risk. To simplify the interpretation of PRS, we categorized it into percentiles based on its distribution in controls.

Combined effect of PRS and family history

To evaluate the incremental contribution of combining PRS and family history for prediction, we

additionally calculated the expected information for discrimination (expected weight of evidence, denoted as Λ).[18] Briefly, the expected information for discrimination is the expected log-likelihood ratio in favor of correct assignment as case or control, taken as the average of the values in cases and controls. One advantage of using Λ is that the contributions of independent variables to predictive performance are additive on the scale of Λ. For a logistic regression model, the sampling distribution of Λ is asymptotically Gaussian. In this situation, the c-statistic can be viewed as a mapping of Λ, which takes values from 0 to infinity to the interval from 0.5 to 1.[18] The rationale and theoretical explanations are presented in Supplementary Methods. Family history of colorectal cancer (CRC) was considered as a categorical variable, dependent on the presence or absence of at least one first-degree relative affected by CRC at any age at the time of recruitment.

Estimation of absolute risk for developing CRC

The absolute risk of CRC for individuals in each risk category was calculated after accounting for competing risks of dying from causes other than CRC by using the formula described previously.[19] Specifically, we obtained sex- and age-dependent UK CRC incidence and mortality rates for 2016 mid-year from the Office for National Statistics (http://www.ons.gov.uk/). The mortality rates for non-CRC causes were estimated by subtracting the age- and sex-specific CRC mortality rates from the overall mortality rates. Full details of these calculations are provided in Supplementary Methods.

RESULTS

The meta-analysis of 11 GWASs resulted in the identification of 1,593 genetic variants associated with CRC at $p<5\times10^{-8}$. After adding SNPs reported in other GWAS and excluding SNPs in LD, a list of 116 SNPs (Table S2) were retained for the creation of a weighted polygenic risk score ($wPRS_{116}$). We additionally created 35 regional genetic scores that included 1,593 SNPs with $p<5\times10^{-8}$ (Table 1). We also used more liberal p-value thresholds and created 40 genetic scores comprising of 1,837 SNPs at $p<10^{-7}$ and 41 genetic scores comprising of 2,712 SNPs at $p<10^{-6}$. The genes harbored in these genomic regions were annotated and are presented in Table S3. We additionally created 17 regional scores for CRP, 5 for VD, 85 for IBD, 69 for BMI and 48 for WHR with p-value threshold setting as 5×10^{-8}. More liberal p-value thresholds ($p<10^{-7}$ and $p<10^{-6}$) were also applied for these traits, and the number of regional genetic scores created and SNPs included are present Table S4.

We set out to optimize these derived scores by examining their discriminative ability in the Scottish dataset (Table S5). More specifically, the combined effect of 116 CRC SNPs in the form of $wPRS_{116}$ was significantly associated with CRC risk (OR=1.46, 95% CI: 1.41 to 1.50, $p=1.71\times10^{-116}$, 1 SD

increase of $wPRS_{116}$) and showed moderate discriminative ability (c-statistic=0.60) (Table 1). When stratifying CRC status by tumor sites, the predictive ability of $wPRS_{116}$ had less accuracy than that for overall CRC risk (Table 1). With inclusion of multiple regional scores, the best model of regional genetic risk scores (including 31 CRC scores, 7 CRP scores, 2 VD scores, 25 IBD scores, 18 BMI scores and 7 WHR scores) yielded a c-statistics of 0.60 (Table S4). When comparing to the $wPRS_{116}$, the regional scores showed no further improvement on overall CRC prediction (the p-value of likelihood-ratio test for assessing predictive accuracy between the models of $wPRS_{116}$ and genetic risk scores was close to 1).

We next tested the predictive performance of $wPRS_{116}$ in the UK Biobank dataset. Similarly, the $wPRS_{116}$ showed moderate discrimination ability with a c-statistic of 0.610 and an OR of 1.49 (95% CI: 1.44 to 1.54, $p=6.67\times10^{-128}$, per SD increase) (Table 2). For individuals in the lowest 1% of $wPRS_{116}$, the OR compared with the middle quintile (40%-60%) was 0.32 (95%CI: 0.19 to 0.54, $p=8.51\times10^{-6}$); by contrast, for individuals in the highest 1% of the PRS distribution, the corresponding estimated OR was 3.25 (95%CI: 2.50 to 4.22, $p=1.52\times10^{-17}$) (Fig.1, Table S6). When considering CRC risk separately for proximal colon, distal colon and rectum, it showed no improvement in predictive performance. We then explored the modification effect of the $wPRS_{116}$ by sex, age, or family history, but found no evidence of an interaction effect (Table S7, $P_{interaction} = 0.426$ for multiplicative interaction with sex, $P_{interaction} = 0.688$ with age, $P_{interaction} = 0.388$ with family history), therefore we did not fit additional interaction terms in the model.

We then assessed the incremental contribution of adding $wPRS_{116}$ and family history to a baseline model of age, sex and the first 10 genetic PCs as predictors in UK Biobank. Baseline model on age, sex and the 10 PCs yielded a c-statistic of 0.53, and the corresponding estimate of Λ was 0.01 bits (Table S8). When adding family history alone, the c-statistic increased to 0.55 and the corresponding Λ was 0.02 bits. Adding both family history and $wPRS_{116}$ yielded c-statistic of 0.610 and an incremental value of 0.10 bits, which showed significantly improvement over family history alone. We recalibrated the posterior probabilities by fitting a logistic regression model with the response variable as outcome and the logit of the posterior probability as the predictive variable. It showed that recalibration of the posterior probabilities increases the test log-likelihood only by 1 natural log units for baseline + family history and showed no increases for the baseline + FH + $wPRS_{116}$, indicating that both these models were well-calibrated (Fig. S2).

To gauge the potential public health impact of applying such risk prediction model in the general population, we estimated the 10-year absolute risk of the general UK population (Fig. S3, Table S9). We observed that the estimated absolute CRC risk for individuals at the highest 1% of PRS began to increase sharply after 45 years old, and reached a risk of 22.1% in men and 14.4% in women by 75 years old. As 50 years old is the recommended starting age of screening in Scotland, we used the

average risk at this age as reference threshold (0.48% for man and 0.33% women). Individuals in the top 10% of $wPRS_{116}$ would reach or exceed this level of risk at 45 years old, which is 5 years earlier than the average risk population; in contrast, individuals in the bottom 10% of PRS would stay below this average risk until 60 years old. If we considered individuals with 10-year absolute risk ≥5% as high risk group, with risk strata by $wPRS_{116}$ in population settings, we will able to identify 10% men and 5% women meriting intensive screening at 65 years old.

DISCUSSION

In this study we describe a systematic approach to derive, validate and test a number of candidate genetic risk scores with incorporating information from hundreds to thousands of common genetic variants to predict polygenic susceptibility of CRC. We evaluated the predictive performance of both a genomic risk score and a series of regional genetic scores that were built based on the summary statistics from multiple GWASs. Our study shows that a weighted genomic risk score including 116 CRC susceptibility SNPs is the score with the best performance, while deconstructing genetic risk into multiple regional scores or inclusion of additional SNPs above the genome-wide significance threshold showed no further improvement on prediction performance. By implementing the PRSs developed, we show that the inclusion of genetic factors into a baseline model of age, sex and family history results in a significant improvement of CRC risk stratification. It should be noted that family history data in this study were collected based on self-reported bowel cancer history of parents and siblings. Therefore any potential recall bias on family history may lead to the prediction improvement being less accurate.

ROC analysis of the genetic model that included $wPRS_{116}$ showed an improved but still modest discriminative performance (c-statistic: 0.60 in Scottish dataset, 0.61 in UK Biobank dataset). To our knowledge, the best predictive performance achieved by PRS along with age and family history was 0.69 and 0.60 for Korean men and women,[20] but the SNPs were chosen from the same dataset used to generate the model, and therefore the reported c-statistics are likely inflated. Other genetic models showed consistently low to modest discriminatory abilities.[21] Hsu *et al* developed sex-specific models by using family history and 27 common genetic variants with adjustment of endoscopy history and obtained a discrimination ability of 0.59 for men and 0.56 for women.[22] Similarly, Smith *et al* reported a c-statistic of 0.57 for genetic risk model combing 41 CRC susceptibility SNPs.[23] The most recent genetic model for CRC was developed by Jeon *et al* including 63 CRC susceptibility SNPs and achieved a slightly improved predictive accuracy with a c-statistic of 0.59.[24] This modest level of test performance is consistent across studies, suggesting that risk assessment algorithms based on independent SNPs reaching genome-wide significance level have similar performance characteristics in European populations. However, it should be kept in mind that our results pertain to the UK population of white ancestry only, and therefore generalization to any other ancestry need further evaluation.

With the expectation of improving the predictive power of common genetic variants, we additionally derived a set of SNPs associated with CRC risk with liberal p-value thresholds to allow the contribution of signals from additional susceptibility SNPs that have not been discovered or validated in previous GWAS efforts. Any correlation between SNPs was addressed by creating LD-adjusted regional scores. However, with inclusion of thousands of SNPs, the predictive capacity did not improve but showed a lower c-statistic in the range of 0.58 to 0.59, which is probably due to the cost of adding noise from SNPs that were not truly associated with CRC. To assess if the genetic susceptibility of known risk factors of CRC would further contribute to CRC prediction, we developed prediction models, which incorporated genetic information of several known risk factors, but the c-statistic remained close to 0.60.

Most previous efforts mainly focused on the predictive ability of PRS to capture the overall risk of CRC.[4, 5, 22-24] However, there is compelling evidence suggesting that genetic risk factors may differ by anatomic locations.[25] We therefore aimed to improve prediction of site-specific CRC by deconstructing the commonly used genomic risk score into several regional scores, allowing susceptibility signals through multiple/different mechanisms to influence genetic predisposition to site-specific CRC. Although we treated proximal, distal and rectal cancer as distinct endpoints to generate the best set of regional scores respectively, their predictive performance still showed modest discriminative ability. This might be limited by the fact that the weights used for regional score calculation were derived from the coefficient estimates for overall CRC instead of site-specific ones.

An extrapolation to the UK population led to the conclusion that 10% of the general population will have a 10-years absolute risk approaching 5% after 65 years old on the basis of quantifiable genetic risk alone and who will merit intensive screening. A 5% threshold of absolute risk has clinical and public health impact since it exceeds the highest risk at any age in the general population and it is 10-fold greater than the risk of a 50-year old person who is eligible to enter the population-based screening programs. Additionally, the modelling shows individuals at different levels of the $wPRS_{116}$ will reach the same risk estimate at different ages, supporting the notion that using genetic profiling in combination with age will lead to more effective risk stratification.

In conclusion, we show that prediction of CRC risk based on profiling with common genetic variants presents a moderate discriminability. Although the contribution of $wPRS_{116}$ to individualized risk profiling is limited, employing genetic risk profiling can achieve a moderate degree of risk discrimination that is helpful to identify a population subset with high genetic risk.

Ethics approval and informed consent

Ethics approval of SOCCS was obtained from the Multi-Centre Research Ethics committee for Scotland (approval number MREC/ 01/0/5) and informed consent was provided by all participants. UK Biobank has approval from the North West Multi-Centre Research Ethics Committee (11/NW/0382) and obtained written informed consent from all participants prior to the study.

Funding

E.T. is supported by a Cancer Research UK Career Development Fellowship (C31250/A22804).

The work was supported by Programme Grant funding from Cancer Research UK (C348/A12076) and by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre. This work was also funded by a grant to MGD as Project Leader with the MRC Human Genetics Unit Centre Grant (U127527198). At the Institute of Cancer Research, this work was supported by Cancer Research UK (C1298/A25514). Additional support was provided by the National Cancer Research Network. In Birmingham, funding was provided by Cancer Research UK (C6199/A16459).

Acknowledgements

We are grateful to all who contribute to recruitment, data collection and data curation. We acknowledge that these studies would not be possible without the patients and controls and their families. We acknowledge the expert support on sample preparation from the Genetics Core of the Edinburgh Wellcome Trust Clinical Research Facility.

Conflict of interest

All authors report no potential conflicts of interest.

Authors' contributions

ET and MD conceived the study; XL undertook data manipulations and statistical analysis with input from AS and PM; MT performed the meta-GWAS analysis. HC, RSH, IPMT, SMF, YH, XZ and MGD provided access to GWAS data. XL wrote the article with input from other authors. All authors critically reviewed the manuscript and contributed important intellectual content. All authors have read and approved the final manuscript as submitted.

Data accessibility

The data that support the findings of our study are available upon reasonable request from the corresponding authors. The data are not publicly available due to privacy or ethical restrictions.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68: 394-424.

2. Jiao S, Peters U, Berndt S, et al. Estimating the heritability of colorectal cancer. *Hum Mol Genet* 2014;23: 3898-905.

3. Kemp Z, Thirlwell C, Sieber O, et al. An update on the genetics of colorectal cancer. *Hum Mol Genet* 2004;13 Spec No 2: R177-85.

4. Dunlop MG, Tenesa A, Farrington SM, et al. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals. *Gut* 2013;62: 871-81.

5. Frampton MJ, Law P, Litchfield K, et al. Implications of polygenic risk for personalised colorectal cancer screening. *Ann Oncol* 2016;27: 429-34.

6. Peters U, Hutter CM, Hsu L, et al. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* 2012;131: 217-34.

7. Whiffin N, Hosking FJ, Farrington SM, et al. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum Mol Genet* 2014;23: 4729-37.

8. Al-Tassan NA, Whiffin N, Hosking FJ, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep* 2015;5: 10442.

9. Hofer P, Hagmann M, Brezina S, et al. Bayesian and frequentist analysis of an Austrian genome-wide association study of colorectal cancer and advanced adenomas. *Oncotarget* 2017;8: 98623-34.

10. He Y, Timofeeva M, Farrington SM, et al. Exploring causality in the association between circulating 25-hydroxyvitamin D and colorectal cancer risk: a large Mendelian randomisation study. *BMC Med* 2018;16: 142.

11. Weigl K, Chang-Claude J, Knebel P, et al. Strongly enhanced colorectal cancer risk stratification by combining family history and genetic risk score. *Clin Epidemiol* 2018;10: 143-52.

12. Orlando G, Law PJ, Palin K, et al. Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Hum Mol Genet* 2016;25: 2349-59.

13. Penegar S, Wood W, Lubbe S, et al. National study of colorectal cancer genetics. *Br J Cancer* 2007;97: 1305-9.

14. Paul J, Briggs A, Harkin A, et al. SCOT: Short Course Oncology Therapy—A comparison of 12 and 24 weeks of adjuvant chemotherapy in colorectal cancer. *J Clin Oncol* 2011: 29(15_suppl), e14145-e.

15. Houlston RS, Cheadle J, Dobbins SE, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* 2010;42: 973-7.

16. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* 2006;35: 34-41.

17. Vilhjalmsson BJ, Yang J, Finucane HK, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* 2015;97: 576-92.

18. McKeigue P. Quantifying performance of a diagnostic test as the expected information for discrimination: Relation to the C-statistic. *Stat Methods Med Res* 2018: 962280218776989.

19. Mavaddat N, Pharoah PD, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst* 2015;107.

20. Jo J, Nam CM, Sull JW, et al. Prediction of Colorectal Cancer Risk Using a Genetic Risk Score: The Korean Cancer Prevention Study-II (KCPS-II). *Genomics Inform* 2012;10: 175-83.

21. McGeoch L, Saunders CL, Griffin SJ, et al. Risk Prediction Models for Colorectal Cancer Incorporating Common Genetic Variants: A Systematic Review. *Cancer Epidemiol Biomarkers Prev* 2019;28: 1580-93.

22. Hsu L, Jeon J, Brenner H, et al. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology* 2015;148: 1330-9.e14.

23. Smith T, Gunter MJ, Tzoulaki I, et al. The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. *Br J Cancer* 2018;119: 1036-9.

24. Jeon J, Du M, Schoen RE, et al. Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. *Gastroenterology* 2018;154: 2152-64.e19.

25. Missiaglia E, Jacobs B, D'Ario G, et al. Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. *Ann Oncol* 2014;25: 1995-2001.