scientific reports

Check for updates

OPEN

Multiple polygenic score approach in colorectal cancer risk prediction

Shangqing Joyce Jiang[1,2,54], Minta Thomas[2,54], Elisabeth A. Rosenthal[3], Amanda I. Phipps[2,4], Lori C. Sakoda[5,6], Franzel J. B. van Duijnhoven[7], Andrew J. Pellatt[8], Christy L. Avery[9], Sonja I. Berndt[10], D. Timothy Bishop[11], Sergi Castellví-Bel[12], Andrew T. Chan[13,14,15,16,17,18], Robert C. Grant[19], Chris Gignoux[20], Andrea Gsur[21], Marc J. Gunter[22,23], Christopher A. Haiman[24], Michael Hoffmeister[25], Gail P. Jarvik[3,26], Mark A. Jenkins[27], Temitope O. Keku[28], Sébastien Küry[29], Jeffrey K. Lee[5,30,31], Loic Le Marchand[32], Victor Moreno[33,34,35,36], Polly A. Newcomb[2,37], Christina C. Newton[38], Shuji Ogino[16,17,39,40], Julie R. Palmer[41], Rachel Pearlman[42], Conghui Qu[2], Robert E. Schoen[43], Caroline Y. Um[38], Bethany Van Guelpen[44,45], Kala Visvanathan[46], Veronika Vymetalkova[47,48], Emily White[2,4], Michael O. Woods[49], Elizabeth A. Platz[46,50], Hermann Brenner[25,51], Douglas A. Corley[5,30], Iris Landorp Vogelaar[52], Li Hsu[2,53,55] & Ulrike Peters[2,4,55]✉

Recent studies have demonstrated that for various diseases, incorporating polygenic risk scores (PRSs) for other traits and diseases into the PRS-based risk prediction model may improve predictive performance – known as Multiple Polygenic Score (MPS) approach. We aimed to examine whether the MPS approach improves colorectal cancer (CRC) risk prediction. We included 2,187 non-CRC PRSs from the polygenic Score (PGS) Catalog and used machine learning (ML) models to select the most predictive non-CRC PRSs, utilizing individual-level data from 31,257 CRC cases and 33,408 controls. An independent dataset from the Genetic Epidemiology Research in Adult Health and Aging (GERA) cohort (4,852 cases and 67,939 controls) was randomly split into subsets for model estimation and validation. The model combined MPS with two existing CRC-PRSs based on known loci and genome-wide genotyping. We then assessed model performance by calculating the area under the receiver operating curve (AUC) in the validation set and performed 1,000 bootstrapped iterations to evaluate AUC improvements. The ML model selected 337 non-CRC PRSs predictive of CRC risk. Adding MPS to the CRC-PRSs significantly improved AUC by 0.017 (95% CI: 0.011–0.022, $p < 0.0001$) when combined with known-loci CRC-PRS, 0.005 (95% CI: 0.002–0.007, $p = 0.0005$) with genome-wide CRC-PRS, and 0.004 (95% CI: 0.002–0.006, $p = 0.0005$) with both the known loci and genome-wide CRC-PRSs. These findings demonstrate MPS's potential to refine CRC risk prediction models and highlight opportunities for further advancements in risk prediction.

Keywords Polygenic risk score, Multi-trait PRS, Colorectal cancer

[1]University of Washington, Seattle, WA, USA. [2]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [3]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA, USA. [4]Department of Epidemiology, University of Washington, Seattle, WA, USA. [5]Kaiser Permanente Division of Research, Oakland, CA, USA. [6]Department of Health Systems Science, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, USA. [7]Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands. [8]Intermountain Health, Salt Lake City, UT, USA. [9]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [10]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [11]Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK. [12]Gastroenterology Department, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. [13]Division of Gastroenterology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [14]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [15]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [16]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [17]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA. [18]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA. [19]Princess Margaret Cancer Centre, University Health Network, Toronto, Canada. [20]Colorado Center for Personalized Medicine, University of Colorado - Anschutz

Medical Campus, Aurora, CO, USA. [21]Center for Cancer Research, Medical University of Vienna, Vienna, Austria. [22]Nutrition and Metabolism Branch, International Agency for Research on Cancer, World Health Organization, Lyon, France. [23]Cancer Epidemiology and Prevention Research Unit, School of Public Health, Imperial College London, London, UK. [24]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [25]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. [26]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [27]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC, Australia. [28]Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, NC, USA. [29]Service de Génétique médicale, Nantes Université, CHU de Nantes, Nantes F-44000, France. [30]Department of Gastroenterology, Kaiser Permanente San Francisco Medical Center, San Francisco, CA, USA. [31]Division of Gastroenterology, University of California, San Francisco, San Francisco, CA, USA. [32]University of Hawaii Cancer Center, Honolulu, HI, USA. [33]Oncology Data Analytics Program (ODAP), Unit of Biomarkers and Suceptibility (UBS), Catalan Institute of Oncology (ICO), L'Hospitalet del Llobregat, Barcelona 08908, Spain. [34]ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona 08908, Spain. [35]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid 28029, Spain. [36]Department of Clinical Sciences, Faculty of Medicine and health Sciences, Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona (UB), L'Hospitalet de Llobregat, Barcelona 08908, Spain. [37]School of Public Health, University of Washington, Seattle, WA, USA. [38]Department of Population Science, American Cancer Society, Atlanta, Georgia. [39]Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [40]Institute of Science Tokyo, Tokyo, Japan. [41]Slone Epidemiology Center, at Boston University, Boston, MA, USA. [42]Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, OH, USA. [43]Departments of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA. [44]Department of Diagnostics and Intervention, Oncology Unit, Umeå University, Umeå, Sweden. [45]Wallenberg Centre for Molecular Medicine, Umeå University, Umeå, Sweden. [46]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [47]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic. [48]Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Prague, Czech Republic. [49]Discipline of Genetics, Memorial University of Newfoundland, St. John's, Canada. [50]Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, USA. [51]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [52]Department of Public Health, Erasmus MC, University Medical Center, Rotterdam, The Netherlands. [53]Department of Biostatistics, University of Washington, Seattle, WA, USA. [54]Shangqing Joyce Jiang and Minta Thomas contributed equally to this work. [55]Li Hsu and Ulrike Peters: These authors jointly supervised this work. ✉email: upeters@fredhutch.org

Colorectal cancer (CRC) is one of the most common causes of cancer death[1]. In 2024, more than 150,000 individuals are projected to be diagnosed in the US, with CRC and more than 50,000 deaths are expected to be attributed to the disease[2]. However, CRC screening can effectively reduce the incidence of CRC, diagnose patients at an earlier stage, remove precursor lesions (polyps and adenomas), and ultimately improve health outcomes[1]. One promising approach to guiding personalized CRC screening is the use of risk prediction models.

Several CRC risk prediction models have been developed, primarily based on environmental and lifestyle risk factors, such as smoking, alcohol, diet, obesity, and diabetes[3–5]. In recent years, prediction scores based on common genetic variants across the human genome have shown promise in identifying individuals with a higher genetic risk of CRC[6,7]. These predictors are known as Polygenic Risk Scores (PRSs). PRSs leverage information based on either known loci from genome-wide association studies (GWAS) or genome-wide risk prediction models including a large number of common variants[8]. PRS tools developed for CRC have been shown to identify individuals at a higher risk for CRC, and adding such PRS into existing clinical risk prediction scores improves risk prediction[6,9].

There have been many advances in the development of PRSs. With more data available in GWAS, more known loci have been identified and their associations with CRC are becoming better understood[7,10]. In addition, machine learning (ML) models that use genome-wide information have demonstrated further improvements in PRS development[11,12]. In independent validations, these CRC-PRSs developed from genome-wide information using ML models had a high area under the receiver operating curve (AUROC, AUC), a measure for predictive performance, ranging from 0.63 to 0.65, demonstrating their promise in predicting CRC risk using genetic data[6,7].

Another recent advancement in genetic risk prediction is the Multiple PRS (MPS) approach, which incorporates PRSs for multiple related traits into a single risk prediction model. The motivation of our study is two-fold. First, there are many risk factors for CRC, such as type 2 diabetes, smoking, obesity, and chronic inflammation[13–15], have genetic components. For example, type 2 diabetes is a known risk factor for CRC, and PRSs developed for diabetes have demonstrated strong predictive performance for type 2 diabetes[16]. Including PRSs for these risk factors may add complementary information beyond CRC-PRSs. Second, traditional penalization models used in developing PRSs, which shrink the effect of less predictive variants, may lose valuable information, that could potentially be captured by PRSs for other traits[8].

Previous studies have shown that including PRSs from other traits can improve risk prediction for mental health outcomes and diabetes among others[17–19]. However, this approach has not yet been tested for CRC. As the goal is to improve risk prediction, a more comprehensive approach using a broad set of PRSs, including those not directly associated with CRC, is needed to evaluate whether this MPS approach is effective for CRC risk prediction[17,18,20].

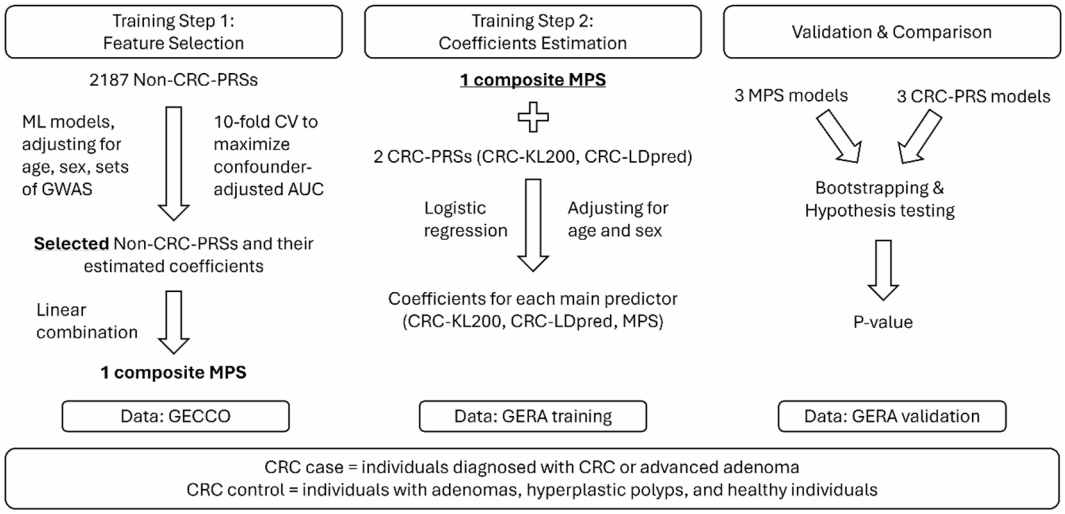

Fig. 1. Overall approach: The first step was to select the most predictive non-CRC PRSs among 2,187 non-CRC PRSs using ML models (Lasso, Ridge and Elastic Net) with 10-fold CV. This step was performed using GECCO data. We adjusted for age, sex, and GWAS platforms. We selected the best-performing model with the highest confounder-adjusted AUC. Afterwards, we constructed a composite MPS by taking the linear combination of selected non-CRC PRSs, weighted by their coefficient estimates. The second step was to construct risk prediction models by adding MPS to CRC-PRSs-only models. We adjusted for age and sex in this step. The data were the training subset of GERA cohort. The last step was to validate the confounder-adjusted AUC using the validation dataset. We used bootstrapping to compare confounder-adjusted AUC in CRC-PRS-only models and MPS models and calculated the two-sided p-values.

GECCO and CCFR					
	Cases (*N* = 31,257)			**Controls (*N* = 33,408)**	**Overall (*N* = 64,665)**
	CRC (*N* = 26,884)	**Advanced adenoma (*N* = 4,373)**	**Total (*N* = 31,257)**	**Total**	**Total**
Age, mean (SD)	65.7 (11.4)	64.3 (8.0)	65.5 (11.0)	64.2 (11.6)	64.8 (11.4)
Female, n (%)	12,811 (47.7%)	1,858 (42.5%)	14,669 (46.9%)	17,780 (53.2%)	32,449 (50.2%)
GERA					
	Cases (*N* = 4,852)			**Controls (*N* = 67,939)**	**Overall (*N* = 72,791)**
	CRC (*N* = 1,311)	**Advanced adenoma (*N* = 3,541)**	**Total (*N* = 4,852)**	**Total**	**Total**
Age, mean (SD)	70.9 (11.7)	68.5 (9.1)	69.2 (9.9)	71.3 (13.3)	71.1 (13.1)
Female, n (%)	674 (51.4%)	1,643 (46.4%)	2,317 (47.8%)	40,203 (59.2%)	42,520 (58.4%)

Table 1. Characteristics of the datasets: GECCO and CCFR and GERA. Abbreviations. CRC: colorectal cancer. GECCO: Genetic and Epidemiology of CRC Consortium. CCFR: Colon Cancer Family Registry. GERA: Genetic Epidemiology Research in Adult Health and Aging (GERA) cohort. SD: standard deviation.

To address this gap, our study aims to assess whether a comprehensive MPS approach can enhance CRC risk prediction compared to a PRS risk prediction model developed solely for CRC = specific PRSs (i.e., CRC-PRS model).

Methods

The overall approach for developing and validating the MPS risk prediction model for CRC is illustrated in Fig. 1. Briefly, our approach involved three steps. The first step was to incorporate all PRSs for other traits (i.e., non-CRC PRSs) available in the polygenic score (PGS) Catalog[16] into our large GWAS dataset for CRC from the Genetic and Epidemiology of CRC Consortium (GECCO) and Colon Cancer Family Registry (CCFR). Detailed information on the imputation process, GWAS, and sub-studies can be found in our previous publications[6,10]. This dataset included 64,665 individuals of European ancestry, comprising 31,257 cases diagnosed with CRC or advanced adenomas (AA), and 33,408 controls. The mean age of participants was 64.8 years (SD = 11.4) and 50.2% were females (Table 1). To select the most predictive PRSs, we used ML models including Lasso, Ridge, and Elastic Net. The best performing model based on cross-validation (CV) performance returned a subset of selected non-CRC PRSs along with their estimated coefficients.

The second step was to generate a CRC risk prediction model using logistic regression, which include both CRC and selected non-CRC PRSs. The third step was to validate the performance of the proposed model. For steps two and three, we randomly split the Genetic Epidemiology Research in Adult Health and Aging

(GERA) cohort into subsets. The GERA cohort is a large-scale, community-based research project studying genetic factors related to age-related diseases in a demographically diverse group of adults. It integrates extensive longitudinal medical records with genetic information from Kaiser Permanente, Northern California, a multi-center integrated healthcare delivery system[21]. Among the 72,791 individuals, there were 4,852 cases and 67,939 controls. The mean age of participants was 71.1 years (SD = 13.1) and 58.4% were females (Table 1). Detailed descriptions of this dataset have been published previously[6]. All study protocols were approved by Fred Hutchinson Institutional Review Board (IRB) and Kaiser Permanente Northern California IRB, and informed consent was obtained from all participants in accordance with the Helsinki Accord.

The following sections provide a detailed description of each step.

CRC-PRSs

We used two CRC-PRSs, both of which we published previously. Briefly, the first CRC-PRS was based on 204 GWAS known loci (CRC-KL200); and the second one was based on genome-wide approach constructed using LDpred2 (CRC-LDpred). For more details of these two CRC-PRSs, please refer to previous publications[7,22].

Non-CRC PRSs

To examine whether incorporating non-CRC PRSs could enhance predictive performance, we comprehensively incorporated all available non-CRC PRSs without prior selection. We downloaded scoring files for all 2,724 PRSs' from the PGS Catalog as of November 8, 2022[16]. The vast majority of SNPs included in the non-CRC PRSs had minimal missingness, with a median of 2% missingness. PRSs with more than 20% missing SNPs were excluded to avoid substantial loss of predictive accuracy (Table S1). Additionally, we excluded PRSs specifically designed to predict CRC, colon cancer, and rectal cancer. To minimize overlap with CRC cases, we also excluded three PRSs related to two broader traits—gastrointestinal cancer and rectal/anal cancer. After these exclusions, 2,187 non-CRC PRSs remained for analysis in the ML model (Fig. 2).

Among these 2,187 non-CRC PRSs, some were associated with precursor lesions of CRC, including rectal polyps, benign neoplasms of the colon, and benign neoplasms of the digestive system. These PRSs were retained in the analysis, as they predict precursors of CRC rather than CRC itself.

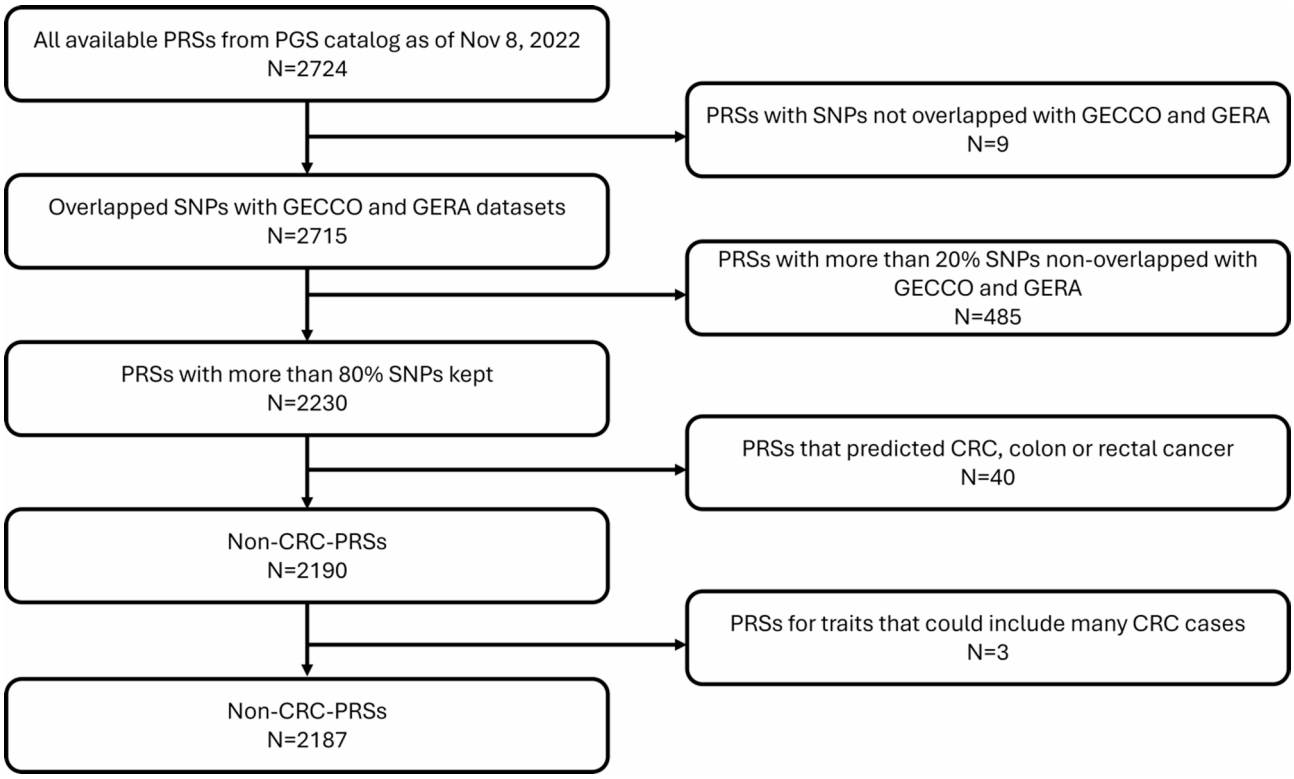

Fig. 2. Selection of non-CRC-PRSs: We downloaded all PRSs available as of Nov 8, 2022, from PGS Catalog. We first excluded 9 PRSs without overlapping SNPs with GECCO, CCFR, and GERA datasets. Then, we further excluded 485 PRSs with at least 20% SNPs unavailable in GECCO, and GERA datasets. Because our focus was non-CRC PRSs, we excluded 40 PRSs that predicted CRC, colon cancer, and rectal cancer. Additionally, three PRSs were excluded because they predicted the trait where many CRC cases were included. Eventually, we had 2,187 non-CRC PRSs.

Statistical analysis

PRS calculation

We calculated all PRSs for each individual using the genotype data of GECCO and CCFR for step one and of GERA for steps two and three. For each PRS, we calculated the weighted sum of effect alleles across all common variants included in that specific PRS. Scoring files for each PRS were obtained from the PGS Catalog, which provided information on SNPs, effect alleles, reference alleles, and corresponding effect weights[16]. All PRSs were standardized to have a mean of zero and a standard deviation (SD) of one, based on the overall sample mean.

Model Building and validation

Step 1. Selection of non-CRC PRSs.

The first step involved identifying the most predictive non-CRC PRSs for CRC risk. We developed ML models, including Ridge, Lasso, and Elastic Net models on non-CRC PRSs derived from the GECCO and CCFR datasets. To determine the optimal penalization tuning parameter, we performed 10-fold CV, selecting the value that maximized the confounder-adjusted AUC. This metric standardizes confounders distribution in the case group to match that of the control group, ensuring that the predictive performance is attributed solely to primary predictors of interest.

Each ML model was adjusted for age, sex, and genotyping platform, with penalization applied only to PRSs, while age, sex, and genotyping platform remained unpenalized. After performing 10-fold CV, we calculated the mean confounder-adjusted AUC and selected the model with the highest value. Using the optimal tuning parameters, we re-fitted the model on the full GECCO and CCFR dataset to estimate the coefficients for non-CRC PRSs. At the end of this step, we generated an MPS score using best performing ML model. This score represents a linear combination of selected non-CRC PRSs weighted by the estimated coefficients.

Step 2. Developing risk prediction models.

The second step was to construct risk prediction models including both CRC-PRS and MPS. We considered six models in total: three CRC-PRS models (CRC-KL200 only, CRC-LDpred only, and a combination of CRC-KL200 and CRC-LDpred), and three MPS + CRC-PRS models where MPS was added to each of the three CRC-PRSs models. Since we had significantly reduced the number of predictors using ML methods and had a sufficiently large sample size for model development, we applied logistic regression model to combing PRSs. CRC-PRSs and MPS were normalized to have a mean of zero and a standard deviation (SD) of one, and all models were adjusted for age, and sex. We used the GERA dataset for this step, where 72,791 adults of European ancestry based on genetic data were included.

We randomly split GERA into two equally sample sized. We used one set for estimating the parameters for each of the six models (step 2), and the second set for validation (see step 3. Validation and Comparison). We developed risk prediction models for combined and sex-stratified analysis (i.e., separate models for males and for females). Our primary outcome was advanced neoplasia, including CRC or advanced adenomas, and with all other individuals classified as controls. As a secondary analysis, we examined the CRC as outcome and defining controls as all individuals, except those with CRC or advanced adenoma.

Step 3. Validation and Comparison.

We evaluated the performance of the six risk prediction models using the validation subset of the GERA data sets. For each model, we estimated the confounder-adjusted AUC, accounting for age and sex, and obtained the standard errors (SE) and the 95% confidence intervals (CI) based on 1000 bootstrap samples.

We conducted three pairs of comparisons: (1) combined CRC-KL200 and MPS vs. CRC-KL200, (2) combined CRC-LDpred and MPS vs. CRC-LDpred only, and (3) combined CRC-KL200, CRC-LDpred and MPS vs. CRC-KL200 and CRC-LDpred. To evaluate the improvement in model performances, we first calculated the difference in confounder-adjusted AUC by subtracting the AUC of each CRC-PRS model from the corresponding MPS and CRC-PRS model in the GERA validation dataset (for example: CRC-KL200 + MPS minus CRC-KL200), We then repeated this process across 1,000 bootstrapped samples to estimate the SE of the AUC difference, and derive the 95% CI.

To determine statistical significance, we computed the z-score for the AUC difference and obtained two-sided p-values. A p-value below 0.05 was considered statistically significant.

Results

Step 1. Model selection of non-CRC-PRSs.

The best performing ML model was an Elastic Net model. In 10-fold CV, the model maximized the AUC when the alpha value was around 0.75 and the lambda value was around 0.003. This model achieved a mean confounder-adjusted AUC of 0.607 across 10-fold CV in GECCO data (Figure S1) and identified 337 non-CRC PRSs with non-zero coefficient estimates. The list of selected 337 non-CRC PRSs along with their coefficient estimates can be found in Table S2. Non-CRC PRSs with the highest absolute coefficient values were those for benign neoplasm of colon, any cancer, college education, hemorrhoids, and body mass index (BMI).

Step 2. Developing risk prediction models.

First, we evaluated the performance of all models on datasets where cases included advanced neoplasia, including CRC and advanced adenomas, with all other individuals classified as controls. Across all six models, both CRC-related PRSs and MPS were statistically significant (Table 2). For example, in Model 6, the odds ratio (OR) with 95% confidence interval (CI) for MPS, CRC-KL200, and CRC-LDpred were 1.11 (95% CI 1.05–1.16 with p-value < 0.0001), 1.12 (95% CI 1.05–1.19 with p-value 0.0001), and 1.45 (95% CI 1.36–1.54 with p-value < 0.0001), respectively (Table 2). The results of the sex-stratified analysis were given in Table S3, and Table S4. In the risk prediction models for females, most predictors had statistically significant ORs. However, CRC-KL200 became non-significant when CRC-LDpred was also included as a predictor (Table S3). All CRC-PRS and MPS + CRC-PRS remained significant in all models for males. However, unlike in models for females,

Model No.	Predictors	OR estimates	95% CI: Lower limit	95% CI: Upper limit	*P*-value
1	CRC-KL200	1.49	1.43	1.55	< 0.0001
2	CRC-LDpred	1.63	1.57	1.72	< 0.0001
3	CRC-KL200	1.15	1.08	1.21	< 0.0001
	CRC-LDpred	1.51	1.42	1.58	< 0.0001
4	Composite MPS	1.22	1.16	1.28	< 0.0001
	CRC-KL200	1.36	1.30	1.42	< 0.0001
5	Composite MPS	1.13	1.07	1.19	< 0.0001
	CRC-LDpred	1.54	1.46	1.62	< 0.0001
6	Composite MPS	1.11	1.05	1.16	< 0.0001
	CRC-KL200	1.12	1.05	1.19	0.0001
	CRC-LDpred	1.45	1.36	1.54	< 0.0001

Table 2. Main analysis: odds ratio (OR) estimates, 95% CI and p-values for each predictor in logistic regression model. Abbreviations. CRC: colorectal cancer. CI: confidence interval. MPS: multiple PRS.

Main analysis			
Model No	Main predictors	AUC (95% CI)	Point estimate of difference in AUC (95% CI) and P-value
1	CRC-KL200	0.600 (0.589–0.612)	
2	CRC-LDpred	0.631 (0.620–0.643)	
3	CRC-KL200, CRC-LDpred	0.632 (0.621–0.644)	
4	CRC-KL200, MPS	0.617 (0.606–0.629)	0.017# (0.011–0.022) pval < 0.0001
5	CRC-LDpred, MPS	0.636 (0.625–0.648)	0.005* (0.002–0.007) pval 0.0005
6	CRC-KL200, CRC-LDpred, MPS	0.636 (0.625–0.648)	0.004^ (0.002–0.006) pval 0.0005

Table 3. Main analyses: AUC (95% CI) estimates of risk prediction models and point estimate of differences in AUCs between models with and without MPS. Abbreviations. AUC: area under the Receiver Operating Curve. CRC: colorectal cancer. CI: confidence interval. MPS: multiple PRS. # Model 4 AUC – Model 1 AUC; *Model 5 AUC-Model2 AUC; and ^ Model 6 AUC – Model 3 AUC.

CRC-KL200 remained significant even when CRC-LDpred was included in the model (CRC-KL200 in Model 3: OR 1.19 (95% CI 1.09–1.28 with p-value < 0.0001); in Model 6: OR 1.16 (95% CI 1.07–1.25 with p-value 0.0003). MPS continued to be statistically significant even when both CRC-KL200 and CRC-LDpred were included (MPS in Model 6: OR 1.12 (95% CI 1.05–1.20 with p-value 0.001) (Table S4). We also assessed the performance of the model when cases were restricted to individuals with CRC only. MPS and CRC-KL200 become non-significant predictors when the model included CRC-LDpred, and the results were included in Table S5.

Step 3. Validation and Comparison.

AUC analysis for advanced neoplasia.

When comparing individuals with advanced neoplasia to all others, the AUC was 0.600 (95%CI: 0.589–0.612), 0.631 (95%CI: 0.620–0.643), and 0.632 (95%CI: 0.621–0.644) for CRC-KL200, CRC-LDpred, and CRC-KL200 + CRC-LDpred, respectively (Table 3). The AUCs for the three MPS + CRC-PRS models are 0.617 (95%CI: 0.606–0.629), 0.636 (95%CI: 0.625–0.648), and 0.636 (95%CI: 0.625–0.648), respectively (Table 3).

Incorporating MPS into the known loci model increased the AUC 0.017 (95% CI: 0.011–0.022; $p < 0.0001$); the CRC-LDpred model 0.005 (95%CI: 0.002–0.007; $p = 0.0005$); and both the CRC-KL200 and CRC-LDpred model 0.0004 (95%CI: 0.002–0.006; $p = 0.0005$) (Table 3, Figure S2-S4).

The results of the sex-stratified analysis were given in Table S6 and Table S7. For females, incorporating MPS into the known loci model increased the AUC 0.012 (95% CI: 0.003–0.021); the CRC-LDpred model 0.004 (95%CI: 0.001–0.009); and both the CRC-KL200 and CRC-LDpred model 0.003 (95%CI: 0.0-, 0.007) (Table S6), whereas for males, incorporating MPS into the known loci model increased the AUC 0.016 (95% CI: 0.009–0.023); the CRC-LDpred model 0.005 (95%CI: 0.001–0.010); and both the CRC-KL200 and CRC-LDpred model 0.004 (95%CI: 0.001–0.008) (Table S7),

CRC analysis.

When we restricted the analysis to CRC only (excluding advanced adenoma) the addition of MPS improved the risk prediction performance of the CRC-KL200 model (AUC difference = 0.015; 95%CI: 0.007, 0.023; $p = 0.0002$), but did not improve the performance significantly when CRC-LDpred was the main predictor (AUC difference = 0.002; 95%CI: 0.000, 0.015; $p = 0.13$) or when CRC-LDpred and CRC-KL200 were both included (AUC difference = 0.002; 95%CI: 0.000, 0.004; $p = 0.098$) (Table S8).

Discussion

We found that MPS was an independent predictor beyond CRC-PRSs in both the combined and sex-stratified analysis. This suggests that the information captured by non-CRC PRSs contributed to advanced colorectal neoplasia prediction. However, when CRC-LDpred was included in the model, the OR estimate of MPS was rather small. This explained why MPS did not noticeably improve the AUC, especially when CRC-LDpred was included in the model.

Among the 337 non-CRC PRSs selected by the Elastic Net model, those with the largest OR estimates were for predicting benign neoplasm, precursor lesions of CRC. This may explain why in the analysis focusing on CRC, excluding advanced adenomas, MPS's contribution in risk prediction was somewhat weakened. Together, these findings indicate that the added value of MPS may primarily reflect risk for precursor lesions rather than invasive CRC, underscoring the importance of evaluating both endpoints in risk prediction.

Although MPS improved the AUC significantly, the increment in AUC was small. Consistent with our findings, Truong et al. found that incorporating more than 2000 PRSs to predict coronary artery disease, the increase in AUC ranged from 0.003 to 0.023 compared to a model without MPS, and the p-value for the increase in AUC were all highly significant ($p < 2e\text{-}16$)[19]. Additionally, Krapohl et al. found that the MPS approach improved the variance explained of educational achievement, general cognitive ability, and BMI by 0.011 to 0.016 and this increase was statistically significant ($p < 0.004$)[17]. The magnitude of the effect size of MPS in these findings is in line with ours[17,19].

There are at least two factors that may influence whether an MPS approach can contribute to risk prediction. The first one is the genetic architecture of a given trait, and the second factor is how much heritability has already been captured by the trait/disease-specific PRS (i.e., CRC-PRSs in our study). CRC is a complex trait, and both genetic variants and environmental risk factors contribute to its risk deposition[23]. CRC is highly polygenic and impacted by both common genetic risk factors as well as rare variants in high-penetrance genes, such as mismatch repair genes[23]. Rare variants in high-penetrance genes can increase the risk of CRC to up by 70%[23], and account for 3–5% of CRC cases[24,25]. To date, more than 200 common variants have been discovered in GWAS of CRC, which explain close to 20% of the familial risk; however, it is estimated that all common variants (including undiscovered) explain over 70% of the familial risk[7,22]. This suggests that in addition to the known GWAS loci, other common variants can contribute to the risk prediction. This is supported by our finding that the AUC was improved from 0.012 to 0.017 after we added MPS to CRC-KL200.

However, because CRC-LDpred was developed using genome-wide data from a large sample of individuals with European ancestry and Asian ancestry, and constructed by LDpred2, one of the state-of-the-art methods for developing PRS[11,26], it is possible that CRC-LDpred has captured the majority of the genome-wide information of CRC, especially undiscovered common variants. As a result, the incremental improvement from incorporating MPS may be limited. Instead, integrating non-genetic risk factors, such as lifestyle, comorbidities, and environmental exposures, has been shown to significantly improve CRC risk prediction when combining with CRC PRS[9,26,27]. Additionally, new methods for PRS development, such as incorporating functional information, may also help improve risk predictive performance[28–30]. Future studies should focus on developing more effective approaches to improve PRS risk prediction.

There are two potential mechanisms by which the MPS approach might improve risk prediction. The first one is pleiotropy, i.e., the genetic correlations among complex traits[31]. Recent pleiotropic analyses have found novel genetic risk factors associated with CRC[32–34]. These novel genetic risk factors might not have been included in the CRC-PRSs but may be captured by PRS for other traits. Secondly, methods of PRS development may over-shrink the effect sizes of common variants, leading to a loss of information of CRC-PRSs[35–37]. In this case, these common variants can be supplemented by PRSs of other traits in the MPS approach.

Overall, our study shows that combining multiple PRSs yields modest but statistically significant improvements in CRC prediction. The small increments are consistent with prior studies in CRC and other complex traits[17,19]. While these gains in AUC may have limited direct clinical impact, they represent refinements in risk modeling. Multi-PRS models capture complementary information and integrating them with established risk factors such as BMI, smoking, or diabetes may further improve risk prediction. Given the moderate predictiveness of genetic risk prediction models, the greatest utility of PRSs - including MPS - may lie in risk stratification: identifying individuals at elevated risk who might benefit from earlier or more frequent screening, or in enhancing the performance of existing screening tools such as fecal blood tests or blood-based biomarkers.

An important strength of our study is that we designed the MPS approach in a comprehensive way. First, instead of pre-selecting traits based on the current understanding of CRC risk factors, we considered all available non-CRC PRSs from the PGS Catalog[16]. Then, we used ML models (ridge, lasso, and elastic net) to identify those predictive of CRC. This ensured an unbiased model and allowed us to incorporate a broad set of 337 non-CRC PRSs while addressing collinearity through regularized regression. Second, all models were validated in a separate dataset, ensuring that AUC estimates reflect independent evaluation and avoiding overfitting. Third, we used confounder-adjusted AUC[38], which measures the predictive performance of CRC-PRSs and MPS only, excluding prediction attributable to factors known to predict CRC, such as age and sex. Using this metric, we were able to make fair comparison between CRC-PRS models and MPS models.

However, limitations are also noteworthy. First, our study only focused on individuals of European ancestry, as most PRSs were developed for this population[16]. Extending these analyses to diverse ancestral groups will

be important to assess generalizability. Second, we examined the linear relationship between CRC and PRSs. A previous study has used deep neural networks, which captures non-linearity, and found that it performed equally well or outperformed ridge regression[39]. However, a recent article found that XGBoost, another ML method that captures non-linearity, improved the predictive performance of MPS models compared to lasso, but the improvement was only via covariates rather than PRSs[40]. Future studies should examine how and why different ML models perform in the context of MPS in CRC and other diseases.

Conclusion

In conclusion, combining multiple PRSs yields modest but statistically significant improvements in CRC risk prediction. Although the absolute gains are limited, these refinements may still support risk stratification for individuals at elevated risk, potentially guiding earlier or more frequent screening and integration with other screening tools. Future work should focus on incorporating lifestyle and clinical factors, leveraging functional annotation, and validating models across diverse populations to advance personalized CRC prevention and early detection.

Data availability

GECCO have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession numbers [phs001078.v1.p1] (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001078.v1.p1), [phs001415.v1.p1] (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001415.v1.p1), and [phs001315.v1.p1] (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001315.v1.p1). Genotype data of GERA participants who consented to having their data shared with dbGaP are available from dbGaP under accession [phs000674.v2.p2] (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v2.p2). The complete GERA data are available upon successful application to the KP Research Bank.

Received: 7 May 2025; Accepted: 24 September 2025
Published online: 30 October 2025

References

1. Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A. & Jemal, A. Colorectal cancer statistics, 2023. *CA Cancer J. Clin.* **73**, 233–254 (2023).
2. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **74**, 12–49 (2024).
3. Usher-Smith, J. A., Walter, F. M., Emery, J. D., Win, A. K. & Griffin, S. J. Risk prediction models for colorectal cancer: A systematic review. *Cancer Prev. Res. (Phila Pa).* **9**, 13–26 (2016).
4. Peng, L., Balavarca, Y., Weigl, K., Hoffmeister, M. & Brenner, H. Head-to-Head comparison of the performance of 17 risk models for predicting presence of advanced neoplasms in colorectal cancer screening. *Am. J. Gastroenterol.* **114**, 1520–1530 (2019).
5. McGeoch, L. et al. Risk prediction models for colorectal cancer incorporating common genetic variants: A systematic review. *Cancer Epidemiol. Biomarkers Prev.* **28**, 1580–1593 (2019).
6. Thomas, M. et al. Genome-wide modeling of polygenic risk score in colorectal cancer risk. *Am. J. Hum. Genet.* **107**, 432–444 (2020).
7. Thomas, M. et al. Combining Asian and European genome-wide association studies of colorectal cancer improves risk prediction across Racial and ethnic populations. *Nat. Commun.* **14**, 6147 (2023).
8. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
9. Briggs, S. E. W. et al. Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK biobank data: population based cohort study. *BMJ* **379**, e071707 (2022).
10. Huyghe, J. R. et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76–87 (2019).
11. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
12. Elgart, M. et al. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Commun. Biol.* **5**, 856 (2022).
13. Sawicki, T. et al. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers (Basel)* **13**(9), 2025 (2021).
14. Ge, T. et al. Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Med.* **14**, 70 (2022).
15. Hahn, S. J., Kim, S., Choi, Y. S., Lee, J. & Kang, J. Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study. *EBioMedicine* **86**, 104383 (2022).
16. Lambert, S. A. et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).
17. Krapohl, E. et al. Multi-polygenic score approach to trait prediction. *Mol. Psychiatry*. **23**, 1368–1374 (2018).
18. Sinnott-Armstrong, N. et al. Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat. Genet.* **53**, 185–194 (2021).
19. Truong, B. et al. Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell. Genomics.* **4**, 100523 (2024).
20. Neumann, A. et al. Combined polygenic risk scores of different psychiatric traits predict general and specific psychopathology in childhood. *J. Child. Psychol. Psychiatry*. **63**, 636–645 (2022).
21. Pepe, M. S. & Cai, T. The analysis of placement values for evaluating discriminatory measures. *Biometrics* **60**, 528–535 (2004).
22. Fernandez-Rozadilla, C. et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and East Asian ancestries. *Nat. Genet.* **55**, 89–99 (2023).
23. Peters, U., Bien, S. & Zubair, N. Genetic architecture of colorectal cancer. *Gut* **64**, 1623–1636 (2015).
24. Argillander, T. E. et al. Features of incident colorectal cancer in Lynch syndrome. *United Eur. Gastroenterol. J.* **6**, 1215–1222 (2018).
25. Gordon, A. S. et al. Rates of actionable genetic findings in individuals with colorectal cancer or polyps ascertained from a community medical setting. *Am. J. Hum. Genet.* **105**, 526–533 (2019).
26. Su, Y. R. et al. Validation of a Genetic-Enhanced risk prediction model for colorectal cancer in a large Community-Based cohort. *Cancer Epidemiol. Biomarkers Prev.* **32**, 353–362 (2023).

27. Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 713–732 (2019).
28. Márquez-Luna, C. et al. Incorporating functional priors improves polygenic prediction accuracy in UK biobank and 23andMe data sets. *Nat. Commun.* **12**, 6052 (2021).
29. Zhuang, Y., Kim, N. Y., Fritsche, L. G., Mukherjee, B. & Lee, S. Incorporating functional annotation with bilevel continuous shrinkage for polygenic risk prediction. *BMC Bioinform.* **25**, 65 (2024).
30. Zheng, Z. et al. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nat. Genet.* **56**, 767–777 (2024).
31. Bien, S. A. & Peters, U. Moving from one to many: insights from the growing list of pleiotropic cancer risk genes. *Br. J. Cancer*. **120**, 1087–1089 (2019).
32. Cheng, I. et al. Pleiotropic effects of genetic risk variants for other cancers on colorectal cancer risk: PAGE, GECCO and CCFR consortia. *Gut* **63**, 800–807 (2014).
33. Sun, J. et al. Cross-cancer pleiotropic analysis identifies three novel genetic risk loci for colorectal cancer. *Hum. Mol. Genet.* **32**, 2093–2102 (2023).
34. Pardo-Cea, M. A. et al. Biological basis of extensive Pleiotropy between blood traits and cancer risk. *Genome Med.* **16**, 21 (2024).
35. Kim, H., Grueneberg, A., Vazquez, A. I. & Hsu, S. De Los Campos, G. Will big data close the missing heritability gap? *Genetics* **207**, 1135–1145 (2017).
36. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A. & Smoller, J. W. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
37. Baker, E. & Escott-Price, V. Polygenic risk scores in alzheimer's disease: current applications and future directions. *Front. Digit. Health*. **2**, 14 (2020).
38. Le Borgne, F. et al. Standardized and weighted time-dependent receiver operating characteristic curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors. *Stat. Methods Med. Res.* **27**, 3397–3410 (2018).
39. Klau, J. H. et al. AI-based multi-PRS models outperform classical single-PRS models. *Front. Genet.* **14**, 1217860 (2023).
40. Albiñana, C. et al. Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nat. Commun.* **14**, 4702 (2023).

Acknowledgements

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088, R01 CA059045, U01 CA164930, R21 CA191312, R01 CA244588, R01 CA206279, R01 CA201407, P50 CA285275, R01 CA273198, U01HG008657, U01CA261339). Genotyping/Sequencing services were provided by the Center for Inherited Disease Research (CIDR) contract number HHSN268201700006I and HHSN268201200008I. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704. Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685.ASTERISK: a Hospital Clinical Research Program (PHRC-BRD09/C) from the University Hospital Center of Nantes (CHU de Nantes) and supported by the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC). CLUE II funding was from the National Cancer Institute (U01 CA086308, Early Detection Research Network; P30 CA006973), National Institute on Aging (U01 AG018033), and the American Institute for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government. The Colon Cancer Family Registry (CCFR, www.coloncfr.org) is supported in part by funding from the National Cancer Institute (NCI), National Institutes of Health (NIH) (award U01 CA167551). Support for case ascertainment was provided in part from the Surveillance, Epidemiology, and End Results (SEER) Program and the following U.S. state cancer registries: AZ, CO, MN, NC, NH; and by the Victoria Cancer Registry (Australia) and Ontario Cancer Registry (Canada). The CCFR Set-1 (Illumina 1 M/1 M-Duo) and Set-2 (Illumina Omni1-Quad) scans were supported by NIH awards U01 CA122839 and R01 CA143237 (to GC). The CCFR Set-3 (Affymetrix Axiom CORECT Set array) was supported by NIH award U19 CA148107 and R01 CA81488 (to SBG). The CCFR Set-4 (Illumina OncoArray 600 K SNP array) was supported by NIH award U19 CA148107 (to SBG) and by the Center for Inherited Disease Research (CIDR), which is funded by the NIH to the Johns Hopkins University, contract number HHSN268201200008I. Additional funding for the OFCCR/ARCTIC was through award GL201-043 from the Ontario Research Fund (to BWZ), award 112746 from the Canadian Institutes of Health Research (to TJH), through a Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society (to SG), and through generous support from the Ontario Ministry of Research and Innovation. The SFCCR Illumina HumanCytoSNP array was supported in part through NCI/NIH awards U01/U24 CA074794 and R01 CA076366 (to PAN). The content of this manuscript does not necessarily reflect the views or policies of the NCI, NIH or any of the collaborating centers in the Colon Cancer Family Registry (CCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government, any cancer registry, or the CCFR.COLON: The COLON study was sponsored by Wereld Kanker Onderzoek Fonds, including funds from grants 2014/1179, IIG_FULL_2021_022, IIG_FULL_2021_023, and IIG_FULL_2023_0117 as part of the World Cancer Research Fund International Grant Programme, by Alpe d'Huzes and the Dutch Cancer Society (UM 2012–5653, UW 2013-5927, UW2015-7946), by ERA-NET on Translational Cancer Research (TRANSCAN via the Dutch Cancer Society (UW2013-6397, UW2014-6877) and the Netherlands Organization for Health Research and Development (ZonMw), the Netherlands), by the Regio Deal Foodvalley (162135), and by the Dutch Research Council (KICH1.LG01.22.009). The Nqplus study is sponsored by a ZonMW investment grant (98-10030); by PREVIEW, the project PREVention of diabetes through lifestyle intervention and population studies in Europe and around the World (PREVIEW) project which received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant no. 312057; by funds from TI Food and Nutrition (cardiovascular health theme), a public–private partnership on precompetitive research in food and nutrition; and by FOODBALL, the Food Biomarker Alliance, a project from JPI Healthy Diet for a Healthy Life.COLO2&3: National Institutes of Health (R01 CA060987).CORSA: The CORSA study was funded by Austrian Research Funding Agency (FFG) BRIDGE (grant 829675, to Andrea Gsur), the "Herzfelder'sche Familienstiftung" (grant to Andrea Gsur) and was supported by COST Action

BM1206.CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. Czech Republic CCS: This work was supported by the Czech Science Foundation (21-04607X), by the Grant Agency of the Ministry of Health of the Czech Republic (grants AZV NU22J-03-00028 and AZV NU22J-03-00033), and the project National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102) - Funded by the European Union – Next Generation EU. DACHS: This work was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HE 5998/2-1, KL 2354/3-1, RO 2270/8-1 and BR 1704/17-1), the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany, and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A, 01ER1505B and 01KD2104A). DALS: National Institutes of Health (R01 CA048998 to M. L. Slattery).EDRN: This work is funded and supported by the NCI, EDRN Grant (U01-CA152753).EPIC: The coordination of EPIC is financially supported by International Agency for Research on Cancer (IARC) and also by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC). The national cohorts are supported by: Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam- Rehbruecke (DIfE), Federal Ministry of Education and Research (BMBF) (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Compagnia di SanPaolo and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS) - Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO (Spain); Swedish Cancer Society, Swedish Research Council and and Region Skåne and Region Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk; MR/M012190/1 to EPIC-Oxford). (United Kingdom).EPICOLON: This work was supported by grants from Fondo de Investigación Sanitaria/FEDER (PI08/0024, PI08/1276, PS09/02368, PI11/00219, PI11/00681, PI14/00173, PI14/00230, PI17/00509, 17/00878, PI20/00113, PI20/00226, PI23/00189, Acción Transversal de Cáncer), Xunta de Galicia (PGIDIT07PXIB9101209PR), Ministerio de Economia y Competitividad (SAF07-64873, SAF 2010-19273, SAF2014-54453R), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST, PRYGN211085CAST), Beca Grupo de Trabajo "Oncología" AEG (Asociación Española de Gastroenterología), Fundación Privada Olga Torres, FP7 CHIBCHA Consortium, Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya, 2014SGR135, 2014SGR255, 2017SGR21, 2017SGR653, 2021SGR00716, 2021SGR01185), Catalan Tumour Bank Network (Pla Director d'Oncologia, Generalitat de Catalunya), PERIS (SLT002/16/00398, Generalitat de Catalunya), Marató TV3 (202008-10), CERCA Programme (Generalitat de Catalunya) and COST Actions BM1206 and CA17118. CIBERehd is funded by the Instituto de Salud Carlos III.Harvard cohorts: HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, U01 CA167552, R01 CA137178, R01 CA151993, R35 CA197735 and R35 CA253185), NHS by the National Institutes of Health (P01 CA087969, UM1 CA186107, R01 CA137178, R01 CA151993, R35 CA197735 and R35 CA253185), and NHS2 (NHSII) by the National Institutes of Health (U01 CA176726 and R35 CA197735). Related to these cohorts, S.O.'s work was in part supported by the Cancer Research UK Grand Challenge Award (C10674 / A27140) and the American Cancer Society Clinical Research Professor Award (Grant # CRP-24-1185864-01-PROF).Hawaii Adenoma Study: NCI grant R01 CA072520.LCCS: The Leeds Colorectal Cancer Study was funded by the Food Standards Agency and Cancer Research UK Programme Award (C588/A19167).MEC: National Institutes of Health (R37 CA054281, P01 CA033619, R01CA126895, and U01 CA164973).NCCCS I & II: We acknowledge funding support for this project from the National Institutes of Health, R01 CA066635 and P30 DK034987.NFCCR: This work was supported by an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (CRT 43821); the National Institutes of Health, U.S. Department of Health and Human Serivces (U01 CA074783); and National Cancer Institute of Canada grants (18223 and 18226). The authors wish to acknowledge the contribution of Alexandre Belisle and the genotyping team of the McGill University and Génome Québec Innovation Centre, Montréal, Canada, for genotyping the Sequenom panel in the NFCCR samples. Funding was provided to Michael O. Woods by the Canadian Cancer Society Research Institute.NSHDS: The research was supported by Biobank Sweden through funding from the Swedish Research Council (VR 2017-00650, VR 2017-01737), the Swedish Cancer Society (CAN 2017/581), Region Västerbotten (VLL-841671, VLL-833291), Knut and Alice Wallenberg Foundation (VLL-765961), and the Lion's Cancer Research Foundation (several grants) and Insamlingsstiftelsen, both at Umeå University.OSUMC: OCCPI funding was provided by Pelotonia and HNPCC funding was provided by the NCI (CA016058 and CA067941).PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Funding was provided by National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01 HG 004438.RPGEH: Data used in this study were generated by the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH), including the Genetic Epidemiology Research on Adult Health and Aging (GERA) data. The RPGEH has been funded by the National Institutes of Health [RC2 AG036607 (Schaefer and Risch)], the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, The Ellison Medical Foundation, and the Kaiser Permanente Community Benefit Program. This study has also been supported in part by a grant from the National Cancer Institute [R01 CA206279 (Peters, Corley, and Hayes)]. Access to RPGEH data used in this study may be obtained by application to the Kaiser Permanente Research Bank (KPRB) via ResearchBankAccess@kp.org. A subset of the GERA cohort consented for public use

can be found at NlH/dbGaP: phs000674.SELECT: Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Numbers U10 CA037429 (CD Blanke), and UM1 CA182883 (CM Tangen/IM Thompson). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.SMS: This work was supported by the National Cancer Institute (grant P01 CA074184 to J.D.P. and P.A.N., grants R01 CA097325, R03 CA153323, and K05 CA152715 to P.A.N., and the National Center for Advancing Translational Sciences at the National Institutes of Health (grant KL2 TR000421 to A.N.B.-H.)REACH: This work was supported by the National Cancer Institute (grant P01 CA074184 to J.D.P. and P.A.N., grants R01 CA097325, R03 CA153323, and K05 CA152715 to P.A.N., and the National Center for Advancing Translational Sciences at the National Institutes of Health (grant KL2 TR000421 to A.N.B.-H.)VITAL: National Institutes of Health (K05 CA154337).WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001,75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005Acknowledgements: ASTERISK: We are very grateful to those without whom this project would not have existed. We also thank all those who agreed to participate in this study, including the patients and the healthy control persons, as well as all the physicians, technicians and students.CCFR: The Colon CFR graciously thanks the generous contributions of their study participants, dedication of study staff, and the financial support from the U.S. National Cancer Institute, without which this important registry would not exist. The authors would like to thank the study participants and staff of the Seattle Colon Cancer Family Registry and the Hormones and Colon Cancer study (CORE Studies).CLUE II: We thank the participants of Clue I and Clue II and appreciate the continued efforts of the staff at the Johns Hopkins George W. Comstock Center for Public Health Research and Prevention in the conduct of the Clue Cohort Studies. Maryland Cancer Registry (MCR): Cancer data was provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data is also supported by the Cooperative Agreement NU58DP007114, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.COLON and NQplus: the authors would like to thank the COLON and NQplus investigators at Wageningen University & Research and the involved clinicians in the participating hospitals.CORSA: We kindly thank all individuals who agreed to participate in the CORSA study. Furthermore, we thank all cooperating physicians and students and the Biobank Graz of the Medical University of Graz.CPS-II: The authors express sincere appreciation to all Cancer Prevention Study-II participants, and to each member of the study and biospecimen management group. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries and cancer registries supported by the National Cancer Institute's Surveillance Epidemiology and End Results Program. The study protocol was approved by the institutional review boards of Emory University, and those of participating registries as required. The authors assume full responsibility for all analyses and interpretation of results. The views expressed here are those of the authors and do not necessarily represent the American Cancer Society or the American Cancer Society – Cancer Action Network.Czech Republic CCS: We are thankful to all clinicians in major hospitals in the Czech Republic, without whom the study would not be practicable. We are also sincerely grateful to all patients participating in this study.DACHS: We thank all participants and cooperating clinicians, and everyone who provided excellent technical assistance.EDRN: We acknowledge all contributors to the development of the resource at the University of Pittsburgh School of Medicine, Division of Gastroenterology, Hepatology and Nutrition, Department of Pathology, and Biomedical Informatics.EPIC: Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.EPICOLON: We are sincerely grateful to all patients participating in this study who were recruited as part of the EPICOLON project. We acknowledge the Spanish National DNA Bank, Biobank of Hospital Clínic–IDIBAPS and Biobanco Vasco for the availability of the samples. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona.Harvard cohorts: The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. We acknowledge Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital as home of the NHS. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries (NPCR) and/or the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program. Central registries may also be supported by state agencies, universities, and cancer centers. Participating central cancer registries include the following: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Indiana, Iowa, Kentucky, Louisiana, Massachusetts, Maine, Maryland, Michigan, Mississippi, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Puerto Rico, Rhode Island, Seattle SEER Registry, South Carolina, Tennessee, Texas, Utah, Virginia, West Virginia, Wyoming. The authors assume full responsibility for analyses and interpretation of these data.LCCS: We acknowledge the contributions of all who conducted this study which was originally reported as 10.1093/carcin/24.2.275. NCCCS I & II: We would like to thank the study participants, and the NC Colorectal Cancer Study staff.NSHDS investigators thank the Västerbotten Intervention Programme, the Northern Sweden MONICA study, the Biobank Research Unit at Umeå University and Biobanken Norr at Region Västerbotten for providing data and samples and acknowledge the contribution from Biobank Sweden, supported by the Swedish Research Council. PLCO: The authors thank the PLCO Cancer Screening Trial screening center investigators

and the staff from Information Management Services Inc and Westat Inc. Most importantly, we thank the study participants for their contributions that made this study possible. Cancer incidence data have been provided by the District of Columbia Cancer Registry, Georgia Cancer Registry, Hawaii Cancer Registry, Minnesota Cancer Surveillance System, Missouri Cancer Registry, Nevada Central Cancer Registry, Pennsylvania Cancer Registry, Texas Cancer Registry, Virginia Cancer Registry, and Wisconsin Cancer Reporting System. All are supported in part by funds from the Center for Disease Control and Prevention, National Program for Central Registries, local states or by the National Cancer Institute, Surveillance, Epidemiology, and End Results program. The results reported here and the conclusions derived are the sole responsibility of the authors.SEARCH: We thank the SEARCH teamSELECT: We thank the research and clinical staff at the sites that participated on SELECT study, without whom the trial would not have been successful. We are also grateful to the 35,533 dedicated men who participated in SELECT.WHI: The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: https://s3-us-west-2.amazonaws.com/www-whi-org/wp-content/uploads/WHI-Investigator-Long-List.pdf.

Author contributions

S.J., M.T., L.H. and U.P. made substantial contributions to the conception or design of the work, prepared the manuscript text, figures and tables. A.C., C.Y.U., F.J.B.vD., H.B., L.L.M., R.S.P., R.C.G., R.E.S., S.K., S.C-B., S.O., S.I.B., T.K., V.V., V.M. recruited patients and collected samples. S.C-B., S.I.B., T.K., V.V., V.M., U.P. prepared samples and performed QC analysis.E.A.P., G.P.J., L.L.M., L.C.S., S.J., M.T., S.O., S.I.B., V.M., L.H., U.P. analyzed and interpreted the data. All authors reviewed the manuscript and substantially revised the manuscript. L.H. and U.P. supervised the study. U.P. is the corresponding author.

Funding

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088, R01 CA059045, U01 CA164930, R21 CA191312, R01 CA244588, R01 CA206279, R01 CA201407, P50 CA285275, R01 CA273198, U01HG008657, U01CA261339). Genotyping/Sequencing services were provided by the Center for Inherited Disease Research (CIDR) contract number HHSN268201700006I and HHSN268201200008I. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704. Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685. ASTERISK: a Hospital Clinical Research Program (PHRC-BRD09/C) from the University Hospital Center of Nantes (CHU de Nantes) and supported by the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC). CLUE II funding was from the National Cancer Institute (U01 CA086308, Early Detection Research Network; P30 CA006973), National Institute on Aging (U01 AG018033), and the American Institute for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government. The Colon Cancer Family Registry (CCFR, www.coloncfr.org) is supported in part by funding from the National Cancer Institute (NCI), National Institutes of Health (NIH) (award U01 CA167551). Support for case ascertainment was provided in part from the Surveillance, Epidemiology, and End Results (SEER) Program and the following U.S. state cancer registries: AZ, CO, MN, NC, NH; and by the Victoria Cancer Registry (Australia) and Ontario Cancer Registry (Canada). The CCFR Set-1 (Illumina 1 M/1 M-Duo) and Set-2 (Illumina Omni1-Quad) scans were supported by NIH awards U01 CA122839 and R01 CA143237 (to GC). The CCFR Set-3 (Affymetrix Axiom CORECT Set array) was supported by NIH award U19 CA148107 and R01 CA81488 (to SBG). The CCFR Set-4 (Illumina OncoArray 600 K SNP array) was supported by NIH award U19 CA148107 (to SBG) and by the Center for Inherited Disease Research (CIDR), which is funded by the NIH to the Johns Hopkins University, contract number HHSN268201200008I. Additional funding for the OFCCR/ARCTIC was through award GL201-043 from the Ontario Research Fund (to BWZ), award 112746 from the Canadian Institutes of Health Research (to TJH), through a Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society (to SG), and through generous support from the Ontario Ministry of Research and Innovation. The SFCCR Illumina HumanCytoSNP array was supported in part through NCI/NIH awards U01/U24 CA074794 and R01 CA076366 (to PAN). The content of this manuscript does not necessarily reflect the views or policies of the NCI, NIH or any of the collaborating centers in the Colon Cancer Family Registry (CCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government, any cancer registry, or the CCFR. COLON: The COLON study was sponsored by Wereld Kanker Onderzoek Fonds, including funds from grants 2014/1179, IIG_FULL_2021_022, IIG_FULL_2021_023, and IIG_FULL_2023_0117 as part of the World Cancer Research Fund International Grant Programme, by Alpe d'Huzes and the Dutch Cancer Society (UM 2012–5653, UW 2013–5927, UW2015-7946), by ERA-NET on Translational Cancer Research (TRANSCAN via the Dutch Cancer Society (UW2013-6397, UW2014-6877) and the Netherlands Organization for Health Research and Development (ZonMw), the Netherlands), by the Regio Deal Foodvalley (162135), and by the Dutch Research Council (KICH1.LG01.22.009). The Nqplus study is sponsored by a ZonMW investment grant (98-10030); by PREVIEW, the project PREVention of diabetes through lifestyle intervention and population studies in Europe and around the World (PREVIEW) project which received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant no. 312057; by funds from TI Food and Nutrition (cardiovascular health theme), a public–private partnership on precompetitive research in food and nutrition; and by FOODBALL, the Food Biomarker Alliance, a project from JPI Healthy Diet for a Healthy Life. COLO2&3: National Institutes of Health (R01 CA060987). CORSA: The CORSA study was funded by Austrian Research Funding Agency (FFG) BRIDGE (grant 829675, to Andrea

Gsur), the "Herzfelder'sche Familienstiftung" (grant to Andrea Gsur) and was supported by COST Action BM1206. CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. Czech Republic CCS: This work was supported by the Czech Science Foundation (21-04607X), by the Grant Agency of the Ministry of Health of the Czech Republic (grants AZV NU22J-03-00028 and AZV NU22J-03-00033), and the project National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102) - Funded by the European Union – Next Generation EU. DACHS: This work was supported by the German Research Council (BR 1704/6 – 1, BR 1704/6 – 3, BR 1704/6 – 4, CH 117/1–1, HO 5117/2 – 1, HE 5998/2 – 1, KL 2354/3 – 1, RO 2270/8 – 1 and BR 1704/17 – 1), the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany, and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A, 01ER1505B and 01KD2104A). DALS: National Institutes of Health (R01 CA048998 to M. L. Slattery). EDRN: This work is funded and supported by the NCI, EDRN Grant (U01-CA152753). EPIC: The coordination of EPIC is financially supported by International Agency for Research on Cancer (IARC) and also by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC). The national cohorts are supported by: Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam- Rehbruecke (DIfE), Federal Ministry of Education and Research (BMBF) (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Compagnia di SanPaolo and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS) - Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO (Spain); Swedish Cancer Society, Swedish Research Council and and Region Skåne and Region Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk; MR/M012190/1 to EPIC-Oxford). (United Kingdom). EPICOLON: This work was supported by grants from Fondo de Investigación Sanitaria/FEDER (PI08/0024, PI08/1276, PS09/02368, PI11/00219, PI11/00681, PI14/00173, PI14/00230, PI17/00509, 17/00878, PI20/00113, PI20/00226, PI23/00189, Acción Transversal de Cáncer), Xunta de Galicia (PGIDIT07PXIB9101209PR), Ministerio de Economia y Competitividad (SAF07-64873, SAF 2010–19273, SAF2014-54453R), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST, PRYGN211085CAST), Beca Grupo de Trabajo "Oncología" AEG (Asociación Española de Gastroenterología), Fundación Privada Olga Torres, FP7 CHIBCHA Consortium, Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya, 2014SGR135, 2014SGR255, 2017SGR21, 2017SGR653, 2021SGR00716, 2021SGR01185), Catalan Tumour Bank Network (Pla Director d'Oncologia, Generalitat de Catalunya), PERIS (SLT002/16/00398, Generalitat de Catalunya), Marató TV3 (202008-10), CERCA Programme (Generalitat de Catalunya) and COST Actions BM1206 and CA17118. CIBERehd is funded by the Instituto de Salud Carlos III. Harvard cohorts: HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, U01 CA167552, R01 CA137178, R01 CA151993, R35 CA197735 and R35 CA253185), NHS by the National Institutes of Health (P01 CA087969, UM1 CA186107, R01 CA137178, R01 CA151993, R35 CA197735 and R35 CA253185), and NHS2 (NHSII) by the National Institutes of Health (U01 CA176726 and R35 CA197735). Related to these cohorts, S.O.'s work was in part supported by the Cancer Research UK Grand Challenge Award (C10674 / A27140) and the American Cancer Society Clinical Research Professor Award (Grant # CRP-24-1185864-01-PROF). Hawaii Adenoma Study: NCI grant R01 CA072520. LCCS: The Leeds Colorectal Cancer Study was funded by the Food Standards Agency and Cancer Research UK Programme Award (C588/A19167). MEC: National Institutes of Health (R37 CA054281, P01 CA033619, R01CA126895, and U01 CA164973). NCCCS I & II: We acknowledge funding support for this project from the National Institutes of Health, R01 CA066635 and P30 DK034987. NFCCR: This work was supported by an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (CRT 43821); the National Institutes of Health, U.S. Department of Health and Human Serivces (U01 CA074783); and National Cancer Institute of Canada grants (18223 and 18226). The authors wish to acknowledge the contribution of Alexandre Belisle and the genotyping team of the McGill University and Génome Québec Innovation Centre, Montréal, Canada, for genotyping the Sequenom panel in the NFCCR samples. Funding was provided to Michael O. Woods by the Canadian Cancer Society Research Institute. NSHDS: The research was supported by Biobank Sweden through funding from the Swedish Research Council (VR 2017 – 00650, VR 2017 – 01737), the Swedish Cancer Society (CAN 2017/581), Region Västerbotten (VLL-841671, VLL-833291), Knut and Alice Wallenberg Foundation (VLL-765961), and the Lion's Cancer Research Foundation (several grants) and Insamlingsstiftelsen, both at Umeå University. OSUMC: OCCPI funding was provided by Pelotonia and HNPCC funding was provided by the NCI (CA016058 and CA067941). PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Funding was provided by National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01 HG 004438. RPGEH: Data used in this study were generated by the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH), including the Genetic Epidemiology Research on Adult Health and Aging (GERA) data. The RPGEH has been funded by the National Institutes of Health [RC2 AG036607 (Schaefer and Risch)], the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, The Ellison Medical Foundation, and the Kaiser Permanente Community Benefit Program. This study has also been supported in part by a grant from the National Cancer Institute [R01 CA206279 (Peters, Corley, and Hayes)]. Access to RPGEH data used in this study may be obtained by application to the Kaiser

Permanente Research Bank (KPRB) via ResearchBankAccess@kp.org. A subset of the GERA cohort consented for public use can be found at NlH/dbGaP: phs000674. SELECT: Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Numbers U10 CA037429 (CD Blanke), and UM1 CA182883 (CM Tangen/IM Thompson). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. SMS: This work was supported by the National Cancer Institute (grant P01 CA074184 to J.D.P. and P.A.N., grants R01 CA097325, R03 CA153323, and K05 CA152715 to P.A.N., and the National Center for Advancing Translational Sciences at the National Institutes of Health (grant KL2 TR000421 to A.N.B.-H.) REACH: This work was supported by the National Cancer Institute (grant P01 CA074184 to J.D.P. and P.A.N., grants R01 CA097325, R03 CA153323, and K05 CA152715 to P.A.N., and the National Center for Advancing Translational Sciences at the National Institutes of Health (grant KL2 TR000421 to A.N.B.-H.) VITAL: National Institutes of Health (K05 CA154337). WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005.

Declarations

Competing interests

A.C. None with the current work. For work unrelated to this manuscript, getting consulting income from Boehringer Ingelheim and Pfizer Inc.Research support from Freenome Holdings.B.V.G Lecturer honorarium from AstraZeneca AB for educational activities unrelated to this work.L.C.S Research funding from AstraZeneca paid to institution and unrelated to work.R.C.G Graduate scholarship from Pfizer. Research funding from TD Bank. Paid consulting or advisory roles for Astrazeneca, Tempus, Eisai, Incyte, Knight Therapeutics, Guardant Health, and Ipsen. U.P. was a consultant with AbbVie and her husband is holding individual stocks for the following companies: Amazon, ARM Holdings PLC, BioNTech, BYD Company Limited, Crowdstrike Holdings Inc, CureVac, Google/Alphabet, Microsoft Corp, NVIDIA Corp, Stellantis. The remaining authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-21956-w.

Correspondence and requests for materials should be addressed to U.P.