HHS Public Access

Author manuscript

Gastroenterology. Author manuscript; available in PMC 2019 June 01.

Published in final edited form as:

Gastroenterology. 2018 June ; 154(8): 2152–2164.e19. doi:10.1053/j.gastro.2018.02.021.

Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors

Jihyoun Jeon, PhD[1], **Mengmeng Du, ScD**[2], **Robert E. Schoen, MD, MPH**[3], **Michael Hoffmeister, PhD**[4,5,6], **Polly A. Newcomb, PhD**[7], **Sonja I. Berndt, PharmD, PhD**[8], **Bette Caan, DrPH**[9], **Peter T. Campbell, PhD**[10], **Andrew T. Chan, MD**[11,12,13], **Jenny Chang-Claude, PhD**[4,5,6], **Graham G. Giles, PhD**[14], **Jian Gong, PhD**[7], **Tabitha A. Harrison, MPH**[7], **Jeroen R. Huyghe, PhD**[7], **Eric J. Jacobs, PhD**[10], **Li Li, MD, PhD**[15], **Yi Lin, MS**[7], **Loïc Le Marchand, MD, PhD**[16], **John D. Potter, MD, PhD**[7], **Conghui Qu, MS**[7], **Stephanie A. Bien, PhD**[7], **Niha Zubair, PhD**[7], **Robert J. Macinnis, PhD**[14], **Daniel D. Buchanan, PhD**[17,18,19], **John L. Hopper, PhD**[18,20], **Yin Cao, DSc**[11,12], **Reiko Nishihara, PhD**[11], **Gad Rennert, MD, PhD**[21], **Martha L. Slattery, PhD**[22], **Duncan C. Thomas, PhD**[23], **Michael O. Woods, PhD**[24], **Ross L. Prentice, PhD**[7], **Stephen B. Gruber, MD, PhD**[25], **Yingye Zheng, PhD**[7], **Hermann Brenner, MD**[4,5,6], **Richard B. Hayes, DDS, PhD**[26], **Emily White, PhD**[7], **Ulrike Peters, PhD**[7], and **Li Hsu, PhD**[7] **on behalf of the Colorectal Transdisciplinary (CORECT) Study and Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)**

[1]Department of Epidemiology, University of Michigan, Ann Arbor, USA

[2]Memorial Sloan Kettering, New York, USA

[3]Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, USA

[4]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

[5]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

Corresponding Authors: Jihyoun Jeon: 1415 Washington Heights, Ann Arbor, MI 48109-2029. jihjeon@umich.edu; (P) 734-936-1442; (F) 734-764-3192, Ulrike Peters: 1100 Fairview Ave. N., M4-B402, Fred Hutchinson Cancer Research Center, Seattle, WA 98109. upeters@fredhutch.org; (P) 206-667-2450; (F) 206-667-7850, Li Hsu: 1100 Fairview Ave. N., M2-B500, Fred Hutchinson Cancer Research Center, Seattle, WA 98109. lih@fredhutch.org; (P) 206-667-2854; (F) 206-667-7004.

Disclosures: There are no conflicts of interest to disclose.

Part of this study was presented as an oral presentation at 2016 AACR meeting in Florida.

Author Contributions:

Conceived and designed the experiments: JJ, MD, RS, MH, PN, SB, PC, AC, JCC, GG, JG, TH, EJ, LL, LM, JP, RM, DB, JH, GR, MS, DT, MW, RP, SG, YZ, HB, RH, EW, UP, LH. Collected phenotype data and biological samples and contributed these as investigators for their respective study: MH, PN, SB, PC, AC, JCC, GG, LL, LM, DB, GR, MS, MW, HB. Analyzed and interpreted the data: JJ, MD, JG, RP, YZ, UP, LH. Contributed reagents/materials/analysis tools: JJ, MD, YZ, LH. Wrote the manuscript: JJ, MD, UP, LH. Critically reviewed the manuscript drafts and approved the final manuscript: All authors.

[6]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

[7]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, USA

[8]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, USA

[9]Division of Research, Kaiser Permanente Medical Care Program, Oakland, USA

[10]Epidemiology Research Program, American Cancer Society, Atlanta, USA

[11]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA

[12]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, USA

[13]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, USA

[14]Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, and Centre for Epidemiology and Biostatistics, School of Global and Population Health, University of Melbourne, Melbourne, Australia

[15]Case Western Reserve University, Cleveland, USA

[16]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, USA

[17]Colorectal Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Parkville, Victoria, Australia

[18]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria, Australia

[19]Genetic Medicine and Family Cancer Clinic, The Royal Melbourne Hospital, Parkville, Victoria, Australia

[20]Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea

[21]Carmel Medical Center, Haifa, Israel

[22]Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, USA

[23]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA

[24]Memorial University of Newfoundland, St. John's, Canada

[25]USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, USA

[26]Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, USA

Abstract

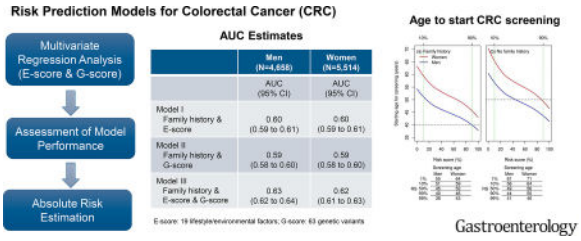
Background & Aims—Guidelines for initiating colorectal cancer (CRC) screening are based on family history but do not consider lifestyle, environmental, or genetic risk factors. We developed models to determine risk of CRC, based on lifestyle and environmental factors and genetic variants, and to identify an optimal age to begin screening.

Methods—We collected data from 9748 CRC cases and 10,590 controls in the Genetics and Epidemiology of Colorectal Cancer Consortium and the Colorectal Transdisciplinary study, from 1992 through 2005. Half of the participants were used to develop the risk determination model and the other half were used to evaluate the discriminatory accuracy (validation set). Models of CRC risk were created based on family history, 19 lifestyle and environmental factors (E-score), and 63 CRC-associated single-nucleotide polymorphisms identified in genome-wide association studies (G-score). We evaluated the discriminatory accuracy of the models by calculating area under the receiver operating characteristic curve (AUC) values, adjusting for study, age, and endoscopy findings for the validation set. We used the models to project the 10-year absolute risk of CRC for a given risk profile and recommend ages to begin screening, in comparison to CRC risk for an average individual at 50 years of age, using external population incidence rates for non-Hispanic whites from the Surveillance, Epidemiology, and End Results Program registry.

Results—In our models, E-score and G-score each determined risk of CRC with greater accuracy than family history. A model that combined both scores and family history estimated CRC risk with an AUC value of 0.63 (95% CI, 0.62–0.64) for men and 0.62 (95% CI, 0.61–0.63) for women; AUC values based on only family history ranged from 0.53 to 0.54 and those based only E-score or G-score ranged from 0.59 to 0.60. Although screening is recommended to begin at age 50 years for individuals with no family history of CRC, starting ages calculated based on combined E-score and G-score differed by 12 years for men and 14 for women, for individuals with the highest vs the lowest 10% of risk.

Conclusions—We used data from 2 large international consortia to develop CRC risk calculation models that included genetic and environmental factors along with family history. These determine risk of CRC and starting ages for screening with greater accuracy than the family history only model, which is based on the current screening guideline. These scoring systems might serve as a first step toward developing individualized CRC prevention strategies.

Graphical abstract



Keywords

colon cancer; GECCO; CORECT; colonoscopy

Introduction

Despite progress in reducing colorectal cancer (CRC) incidence and mortality in recent decades in the US, CRC remains the third leading cause of cancer death.[1] CRC is one of the most preventable and treatable cancers if detected early.[2] Though screening for CRC is recommended for adults between age 50 and 75,[3] in 2013, only 58% were in compliance.[4] Currently screening guidelines are based only on age and family history; however, over 80% of CRC cases have no family history. By evaluating the influence of multiple lifestyle, environmental[5] and genetic risk factors, especially as genetic information will increasingly become a routine part of the medical record,[6,7] risk prediction models can be used to more accurately define low- and high-risk populations, which is the core of precision medicine. Improved risk stratification may also increase screening adherence and uptake, particularly for those at higher risk, as these individuals may be more likely to follow recommendations for prevention when aware of their heightened risk.[8–11] Furthermore, it can optimize the appropriate use of invasive technology.

Several models have been developed to determine the risk of CRC,[9,12–17] adenoma,[18,19], or colorectal neoplasia including both CRC and adenoma,[20] most of which included only clinical, lifestyle and environmental risk factors, while a few models have accounted for the then-known genetic variants[9,13] and one model included a limited number of lifestyle, environmental and genetic factors.[17] However, thus far, no models have attempted to incorporate broad established and putative lifestyle risk factors with the growing number of common genetic variants.

While substantial progress has been made in understanding CRC risk factors, translating lifestyle, environmental and genetic risk factor information into actionable clinical information is the next step in developing personalized prevention. We developed risk prediction models for CRC based on 19 lifestyle and environmental factors and 63 common genetic variants known to be associated with CRC risk using data from 14 population-based studies. We expanded the risk prediction analysis to define the optimal starting age for screening, demonstrating the potential utility of using a model to tailor screening recommendations according to one’s personal risk profile.

Methods and Materials

Data from two large consortia (9,748 CRC cases and 10,590 controls): the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) and the Colorectal Transdisciplinary study (CORECT) were randomly split into two equal halves, with one half for building risk prediction models and the other for evaluating the models. The data consists of 6 case-control studies and 8 cohort-based nested case-control studies. A description of study design and characteristics of each study populations is provided in the Supplementary Note Tables S1 and S2. Only individuals of European descent were included due to insufficient sample size for other ancestry groups. All participants gave written informed consent, and studies were approved by their respective Institutional Review Boards.

Environmental Risk Score

Lifestyle and environmental risk factor information including demographics, behavioral factors, anthropometric traits, diet, pharmacological factors, and medical history was collected by in-person interviews and/or structured questionnaires as detailed previously (references listed in the Supplementary Note). All factors were collected at the study reference time, which was defined as study entry or blood collection for cohort studies and one to two years before sample ascertainment for case-control studies to ensure exposures assessed before cancer diagnoses. We used a multi-step data harmonization procedure as described in detail by Hutter et al.[21] and in the Supplementary Note to harmonize each risk factor across studies. All factors collected under study protocols had generally low missing rates (<10%) except for hormone replacement therapy (16%) and physical activity (16%). Given the relatively low missing rates, we replaced the missing values with the study- and sex-specific mean.

To model the harmonized lifestyle and environmental risk factors, we calculated a score (E-score) based on 19 factors: height (cm), body mass index (kg/m^2), education (less than high school graduate, high school graduate or completed GED, some college or technical school, college graduate or more), history of type 2 diabetes mellitus (no/yes), smoking status (ever/never), alcohol consumption (< 1g/day, 1–28 g/day, >28 g/day; one standard drink is 14g), regular aspirin use (no/yes), regular NSAIDs use (no/yes), regular use of post-menopausal hormones (no/yes, women only), sex- and study-specific quartiles of smoking pack-years and dietary factors (intake of fiber, calcium, folate, processed meat, red meat, fruit, vegetable), total-energy, and physical activity (no/yes). The physical activity was defined as “yes” if sum of vigorous and moderate physical activities is >= 1hour/week, and “no” otherwise. A more detailed description for each variable is provided in the Supplementary Note.

We aimed to create a risk score that summarized an individual’s overall lifestyle and environmental risk profile. Separately for men and women, we created a weighted risk score by summing all the risk factors weighted by their log-odds ratio estimates (Table S4) obtained from a multivariable logistic regression model that included all of the risk factors listed above, adjusting for study, age, family history, and endoscopy history. Family history was coded as a yes/no variable for presence or absence of a first-degree relative with CRC, and endoscopy history was coded as yes, no, or missing, depending on whether a participant had sigmoidoscopy or colonoscopy screening before the study reference time, or such information was missing. For three studies (MECC, Kentucky, NFCCR), we set endoscopy history as entirely missing to avoid any potential bias due to lack of sufficient information to differentiate between screening and diagnostic testing.

As PLCO is a screening trial, we coded endoscopy history as yes if the participant was randomized to the screening arm or had a prior history of endoscopy regardless of trial arms, and no otherwise.[9] Further, to account for reduced protective effect of endoscopy beyond 10 years, we excluded from our analysis CRC cases with follow-up more than 10 years in cohort-based nested case-control studies. Once we obtained the E-score based on 19 lifestyle and environmental risk factors, we then re-coded it to sex- and study-specific percentiles based on cut points in controls and modeled as an ordinal variable.

We also performed a sensitivity analysis using a group lasso regression method[22] to evaluate if any variable selection could improve the model performance, but there was no discernable difference. We used the R-package 'grpreg' for this analysis.

Genetic Risk Score

A total of 63 single-nucleotide polymorphisms (SNPs) at 49 known CRC loci have been identified by GWAS.[23–31] The detailed characteristics of these SNPs are provided in Table S5. Each SNP variable was coded as dosage: for directly genotyped SNPs, we coded it as 0,1, or 2 copies of the risk allele, and for imputed SNPs, as the expected number of copies of the risk allele. Details for the genotyping, imputation, and quality controls have been previously published.[30,32]

Similar to E-score, we created a weighted genetic risk score (G-score) that accounted for the strength of CRC-association with each SNP. The weights were estimated regression coefficients obtained from a multivariable logistic regression that included all 63 SNPs adjusting for age, sex, genotype platform, and up to 6 principal components (PCs) that were previously determined by the genome-wide association studies for each genotype platform[30] to account for population substructure (Table S5). We constructed a G-score for each individual by taking the weighted sum of risk alleles over all 63 SNPs, recoded the G-score as percentile based on cut points in controls, and modeled as an ordinal variable.

Statistical Analysis

We estimated the odds ratios (ORs) and 95% confidence intervals (CIs) for the association of CRC with E-score, G-score, and family history using a multivariable logistic regression analysis adjusting for study, age, endoscopy history, genotype platform, and PCs (whenever appropriate). We stratified the model by sex to allow for variation in risk factor effects for men and women.

In addition, we evaluated how well individual lifestyle and environmental risk factors included in the E-score discriminate cases from controls by estimating odds per adjusted standard deviation (OPERA),[33,34] which measures the risk association with change in the risk factor per its standard deviation after adjusting for all the other variables in the model.

We evaluated the discriminatory accuracy of risk prediction models by calculating the AUC adjusting for study, age, and endoscopy history using the validation dataset. The AUCs were estimated based on family history, E-score, and G-score. We did not include age and endoscopy history in the AUC calculations as predictors but adjusted for both to account for potential confounding. Briefly, we calculated AUC estimates stratified by study, age, and endoscopy history and combined these estimates weighted by the proportion of cases in each stratum.[35] A total of 100 bootstrap samples were used to obtain the 95% CIs of the AUC estimate. Tests with two-sided p-values less than 0.05 are considered statistically significant. All analyses were performed using R.

We also performed three sensitivity analyses. First, we excluded 9 SNPs (rs10911251, rs11903757, rs812481, rs35360328, rs11190164, rs3217810, rs3184504, rs73208120, rs6066825) that were discovered by several consortia (GECCO, CCFR, and CORECT), of

which our data are a subset. Second, we excluded three studies (MECC, Kentucky, NFCCR) for which we set the endoscopy history entirely missing. Finally, as a subset of our data has more detailed family information such as the number and youngest age at diagnosis of first-degree relatives with CRC, we built risk prediction models including these detailed family variables and evaluate the potential gain compared to the yes/no family history variable.

Absolute Risk of CRC

We estimated the 10-year absolute risk of developing CRC and corresponding 95% CIs for a given risk profile, as previously described.[9] We used external population incidence rates for non-Hispanic whites during 1992–2005 (reflective of the time period in which lifestyle and environmental risk factors were assessed across studies) from the Surveillance, Epidemiology, and End Results Program (SEER) registry[36] (Table S6) to calculate the baseline hazard function. This is achieved by multiplying the external incidence rate with one minus population attributable risk (PAR), which is estimated by taking the average of the inverse exponential of risk scores among cases.[37] The sex-specific PAR estimates are largely consistent across studies (Figure S1); we, therefore, pooled all studies to estimate the overall PAR to improve efficiency. We also accounted for competing risks from death in the absolute risk estimation, where the mortality rates were obtained from the National Center for Health Statistics (Table S4). We obtained the 95% CIs of the 10-year absolute risk estimates of CRC with 100 bootstrap samples.

To illustrate the potential utilization of the models, we estimated the recommended age to start screening by assessing when the person's risk exceeds a pre-specified risk threshold. We set the risk threshold, the average of the 10-year CRC risk for a 50-year-old man (1.25%) and woman (0.68%), i.e., (1.25%+0.68%)/2=0.97%, who have not previously received an endoscopy, to the current guideline of starting CRC screening at age 50 for people at average risk. We estimated the age at which the 10-year CRC risk becomes greater than this risk threshold based on sex, family history status, E-score and G-score, and set this age as the recommended starting age for the first screening.

Results

For both men and women, as expected, cases had significantly higher E-scores (Men: mean=59.9, standard deviation (sd)=28.5; Women: mean=59.9, sd=27.9) than controls (Men: mean=49.6, sd=29.1; Women: mean=49.5 sd=29.0), with p-value $< 10^{-15}$. Similarly, cases had significantly higher G-scores (Men: mean=58.7, sd=28.4; Women: mean=58.2, sd=28.2) than controls (Men: 48.9, sd=29.2; women: mean=50.0, sd=28.6), with p-value $< 10^{-15}$. Compared with controls, cases were also more likely to have a positive family history and no prior endoscopy (Table 1).

Association of E-score, G-score, and Family history with CRC risk

All E-score, G-score, and family history were significantly associated with CRC risk after adjusting for study, age, endoscopy history, genotype platform, and PCs (Table 2). The E-score was associated with about 1.36 fold increase risk (Men: OR=1.36 per quartile, 95% CI, 1.29 to 1.44, Women: OR=1.35 per quartile, 95%CI, 1.28 to 1.42). Interestingly, the G-

score increases the CRC risk with a similar magnitude as the E-score (Men: OR= 1.34 per quartile, 95% CI, 1.27 to 1.42, Women: OR=1.30 per quartile, 95% CI, 1.23 to 1.36). A positive family history increased the CRC risk by about 1.5 fold (Men: OR=1.67, 95% CI, 1.38 to 2.03, Women: OR=1.46, 95% CI, 1.24 to 1.72). All three factors had slightly stronger associations for men compared with women.

We evaluated pairwise interaction effects between family history, E-score, G-score, and endoscopy history. Most were not significant at level $\alpha = 0.05$. Our sensitivity analysis was done by excluding 9 SNPs discovered by several consortia, of which our data is part. Using a G-score based on the remaining 54 SNPs, the association with CRC risk remained largely the same with a slight attenuation (Table S7). Another sensitivity analysis by excluding three studies (MECC, Kentucky, NFCCR) also provided similar results compared to the main analysis (Table S8).

We also estimated OPERA for 19 individual lifestyle and environmental risk factors included for the E-score (Table S9) to compare discrimination power between risk factors. In particular, body mass index and regular aspirin use for both men and women, education, alcohol consumption, and vegetable consumption for men, and diabetes, physical activity, red meat consumption and regular use of post-menopausal hormones for women contribute the most to discrimination between cases and controls.

Assessment of the Model Performance

The AUC estimates for the model including only family history were 0.53 (95% CI, 0.52 to 0.54) for men. Including either E-score or G-score into the model improved the AUC significantly (E-score AUC = 0.60, G-score AUC = 0.59; both p-values $<10^{-12}$). Including both E-score and G-score further improved the models that included either the E-score or the G-score alone (AUC=0.63, 95% CI, 0.62 to 0.64; p-value $< 10^{-11}$) (Table 3). A similar pattern was also observed for women.

Estimation of Probability of Developing CRC Given Specific Risk Profiles

The absolute risk of developing CRC varies substantially depending on an individual's risk profile at selected quantiles. Figure 1 presents the 10-year absolute risk of CRC for a 50-year-old who had never had an endoscopic screening exam by varying his or her risk score under various models. As we expected, using either E-score or G-score gives a better risk stratification compared with using only the family history information. Models including both E-score and G-score further improve risk stratification compared with models including only one of these scores for both men and women.

Table 4 shows the estimated 10-year risk of CRC for a 50-year old man or woman given selected profiles for endoscopy history, family history, E-score, and G-score. As reference, we calculated the average 10-year CRC risks for 50-year old men and women based on the SEER CRC incidence rates (Men=0.69%; Women=0.49%).[36]

As an illustration, let's consider a 50-year old man who did not have any first-degree relatives with CRC and had not had a prior endoscopy, but was in a high-risk 90th percentile category for both E-score and G-score. His probability of developing CRC in the next 10

years is 2.90% (95% CI, 2.21% to 3.58%). In contrast, had he previously received an endoscopy, his 10-year risk of CRC could have been reduced to 0.84% (95% CI, 0.64% to 1.04%), which gets close to the average 10-year risk in the general population (0.69%), but lower than the average risk for a man with no prior endoscopy (1.25%). For a 50-year old woman with the same risk profile, her 10-year risk of CRC is 1.41% (95% CI, 1.17% to 1.64%). Had she received a prior endoscopy, her risk could have been reduced to 0.76% (95% CI, 0.63% to 0.88%).

For a 50-year-old person with a low risk profile, such as no first-degree relatives with CRC and being in the low-risk category for both E-score and G-score (lowest 10^{th} percentile for each), the 10-year risk of CRC is much lower than the average 10-year risk in the general population, even without a prior endoscopy (Men=0.42%, 95% CI, 0.32% to 0.52%; Women=0.24%, 95% CI, 0.20% to 0.28%). The risk would be reduced further had they received a prior endoscopy (Men=0.12%, 95% CI, 0.09% to 0.15%; Women=0.13%, 95% CI, 0.11% to 0.15%).

Using risk prediction model to guide risk-stratified CRC screening recommendations

By utilizing our risk prediction model (including family history, E-score, and G-score), we estimated the recommended age for initiating screening for individuals without having a prior endoscopy according to their risk profiles (Figure 2).

For those with high-risk of CRC as determined by a positive family history and 90^{th} percentile of the combined risk score of E-score and G-score, the recommended age to start screening is 40 for men and 46 for women, respectively. On the other hand, for those with a positive family history but in the 10^{th} percentile of the combined risk score, the recommended age to begin in men is 51 and in women 59 (i.e., 11 years later for men and 13 years later for women). Most people with positive family history do not reach the risk threshold until well after the age of 40, when screening is currently recommended to begin in those with a positive family history. Under our model, about 62% of women and 15% of men with a positive family history do not reach the risk threshold until the age of 50 years.

The recommended starting ages for screening in people who do not have any 1^{st} degree relatives with CRC show consistent patterns (Figure 2b), but they are shifted upwards due to the overall lower risk in those with no family history. Based on the combined E-score and G-score, for people with no family history and in the 90^{th} percentile of the score, the starting age is 44 for men and 50 for women, respectively. But for those with no family history but in the 10^{th} percentile of the score, the starting age in men is 56 and in women 64 (i.e., 12 years and 14 years later for men and women, respectively). If we compare the recommended ages for the first screening in those with no family history at the extremes (i.e., 1^{st} percentile versus 99^{th} percentile of the risk score), the difference in first screening age is 20 years or more (See the sub-table under Figure 2). Further, there is a fraction of the population that reaches the risk threshold for starting screening well before age 50. For example, about 15% of men with no family history would reach the risk threshold before age 45.

Discussion

We built sex-specific risk prediction models by including an E-score based on 19 lifestyle and environmental CRC risk factors and a G-score based on 63 common GWAS variants associated with CRC risk. Our analyses show that both scores are independent risk predictors for CRC and yield similar AUC estimates. Incorporating both scores significantly improves the discriminatory accuracy compared with using only family history, which is the current basis for US screening guidelines for CRC. It is worth noting that many existing models[12,14] also include age, endoscopy history, and/or the results of endoscopy examination as predictors, yielding seemingly high AUC estimates. However, models including endoscopy history and/or the results of endoscopy exams as predictors are not appropriate for recommending age for the first screening. Further, as these recommendations are provided based on individual risk profiles at a given age, it is important to ensure that the models based on risk factors other than age perform well. Our sex-specific models that include both environmental/lifestyle variables and common genetic variants without age and endoscopy history have improved discriminatory accuracy.

A few models were previously developed to determine the risk of CRC by including common genetic variants.[9,13,17] Adding genetic information into a model increases the discriminatory accuracy significantly over a model using only family history. Dunlop et al. developed a model based on family history and 10 common genetic variants (AUC=0.56). Hsu et al. developed sex- and site-specific models by using family history and 27 common genetic variants with adjustment for endoscopy history (Men: AUC=0.59 and Women: AUC=0.56). Ibáñez-Sanz et al. included 21 genetic variants in addition to 6 environmental risk factors (AUC=0.63), although the bias due to variable selection may not be corrected. In this study, we included 63 common genetic variants known to be associated with CRC risk without variable selection and observed that AUC is further improved by including more variants (Men: AUC=0.59 and Women: AUC=0.59). Incorporating established lifestyle and environmental factors led to a further improvement in the discriminatory accuracy (Men: AUC=0.63 and Women: AUC=0.62).

Employing our risk prediction model can identify very high and low risk groups, which may have practical implications for screening recommendations. For example, for men with no family history, our previous model[9] using 27 known GWAS variants showed that the recommended age for the first screening ranges from 47 to 52 years of age in the top 10% versus the bottom 10% of the G-score. In contrast, by adding an E-score and expanding the G-score to include more recently identified GWAS variants yielded a substantially wider range for recommending screening from 44 to 56 years of age. Similar improvement was observed for women and individuals with positive family history. Importantly, although the overall rate of CRC declined by about 27% from 1992 to 2009, the rate for the population aged less than 50 years has increased 29% over the same time period,[38] and this sub-population will not undergo screening under current guidelines. Under our risk prediction model, about 10% of men with no family history but at high risk by the combined risk score of E-score and G-score reach the risk threshold before the age of 44 years, suggesting that they may benefit from earlier screening. On the other hand, most people with positive family history do not reach the risk threshold until well after age 40 when the current guidelines

recommend to begin screening, and hence may be able to postpone screening. For example, a sizable fraction of women (62%) and men (15%) with positive family history but at low risk by the combined risk score do not reach the risk threshold until the age of 50 years. Using an E-score only model, the recommended age for the first screening among people with no family history ranges between 11 and 13 years between the top 1% and the bottom 1% of risk score for men and women, respectively (Figure S3). When adding a G-score in the model, the range increases to between 20 and 25 years for men and women, respectively. This indicates that even with small improvement in discriminatory accuracy, the improvement in risk stratification based on combined family history, E- and G-scores can be substantial.

Through our broad risk assessment and determination of the starting age of screening based on a personal risk profile, we are taking the next step towards precision medicine for prevention of CRC, which remains one of the leading causes of cancer death. As genetic information is increasingly incorporated in electronic health records in both the research field (such as, eMERGE, CSER, Geisinger) and commercial field (e.g., Illumina's BioBank chip at $25/subject),[39–41] and online tools to assess lifestyle and environmental risk factors are available,[42,43] personalizing the decision for starting age of screening may be advisable, as screening uptake may be increased for those at higher risk. People with positive family history are twice as likely to undergo screening, suggesting that awareness of individual risk is associated with increased uptake.[9] We used colonoscopy as an example for recommended age to start screening. In other settings, e.g., an organized FIT screening program, the precise recommended age to start may be less critical. However, the USPSTF screening guideline for CRC recommends adults aged 50 to 75 years at average risk to start screening with no preferred screening test. When to start screening, regardless of screening modality, remains an important issue. Our model including both E- and G-scores shows that the range of recommended age to start can be as wide as 20 years for individuals at the top 1% and bottom 1%, suggesting the potential application of our model even for FIT screening. While we validated our model through AUC estimation using a half dataset via a 50–50 random split of data, it would benefit from external cohorts to evaluate other measures such as calibration, clinical efficacy, cost effectiveness, and potential ethical, social and legal concerns.

Our risk prediction models are sex-specific because of the following considerations. Of the lifestyle and environmental factors included to construct an E-score, some risk factors showed sex differences. Although pack years of smoking and dietary variables were coded to sex- and study-specific quartiles to harmonize those across sexes and studies, other variables (e.g., height, alcohol consumption, family history, endoscopy history) showed different distributions by sex (Table 1). Post-menopausal hormone use is only applicable to women. Further, there is a potential sex-specific disparity in CRC risk, with men having a higher CRC incidence rate than women.

The overall missing rates for the most variables were relatively low (<10%) except for a few variables such as hormone replacement therapy and physical activity (~16%). We explored various approaches for handling missingness when building the E-score. These approaches include the complete-case analysis, missing indicator approach, multiple imputation, and

mean imputation. The results were consistent despite all approaches taken. Therefore, we adopted the most straightforward mean imputation approach to handle the missing values.

Our study has several strengths. First, we used a standardized protocol to harmonize the lifestyle and environmental factors leading to consistent and robust associations across all studies. Second, our E-score and G-score are broad, including 19 lifestyle and environmental factors and all known GWAS loci discovered to date. Our machine learning approaches selected almost all of these factors into the model, confirming a role for each of these factors in risk determination. Third, our sample size is very large and, hence, the parameter estimates are reasonably precise. Lastly, our estimates for the association of endoscopy history and family history with CRC risk are consistent with the expected effects reported in other literature[44,45], which render credibility for our model. In summary, our model led to robust risk prediction.

There are limitations. There was lack of detailed information on CRC endoscopic screening (colonoscopy vs. sigmoidoscopy) and other screening tests such as fecal immunochemical test in some studies. As screening may confound the association of E-score and CRC risk, it is conceivable that including more detailed information on screening may provide more accurate estimates for the effect of E-score, and therefore a more precise estimation of CRC risk. Second, family history is a binary variable of whether or not a person had one or more first-degree relatives with CRC. Including additional information, such as number of first-degree relatives with CRC and the age at diagnosis of CRC among the relatives may provide more accurate estimates for the strength of CRC-association with family history.[46] However, our sensitivity analysis using a subset of our data did not show much improvement in discriminatory accuracy by incorporating the more detailed family history information (Tables S10 and S11). Third, our G-score included only the known common genetic susceptibility variants for CRC. However, a sizable fraction of CRC could occur by familial heritability, and only a small fraction of this heritability is currently explained by known common genetic variants.[47] Thus there is potential to further improve risk detremination by incorporating genetic variants that have not reached the GWAS significance level.[48] Also highly penetrant genes, such as DNA mismatch repair genes[49] or APC, could improve the risk prediction. However, carriers of highly penetrant mutations are undergoing substantially different testing than the general population with sporadically occurring cancer, and thus are not the focus of our risk prediction effort. Fourth, the E-score is based on questionnaire data and the measurements for lifestyle and environmental variables, in particular, dietary intake and physical activity, may not be reliable.[50] As such it may lead to a reduced discriminatory accuracy. Recent developments in objective measurements of environmental and lifestyle variables (e.g., wearable devices and metabolomics) have the potential to improve risk prediction and discrimination. Fifth, the E-score, G-score, family history and endoscopy history were established based on the information collected at the reference time in each study, and treated as fixed characteristics that do not change over lifetime. This is true for the G-score, however, other variables may change over the course of the person's life. Our models did not account for such changes due to lack of information. Lastly, our data were restricted to only European descendants and consequently our models may not provide accurate CRC risk prediction for other ancestry groups.

In conclusion, we demonstrate that both lifestyle and environmental factors and common GWAS variants are independent risk predictors for CRC and improved the discriminatory accuracy significantly compared with models that used only family history information. The models yield a wide range of clinically actionable variation in risk stratification as demonstrated by the recommendation on when to start screening. These models may be useful to prioritize those at high risk for targeted prevention or intervention and to reduce emphasis on those at low risk of developing CRC, thereby optimizing utilization of screening in clinical practice with individually tailored prevention strategies. Various aspects, of course, should be considered when utilizing the models for real clinical use such as difficulty of collecting data for some variables like total caloric intake. Our model incorporating the most established or likely associated environmental/lifestyle risk factors as well as known common genetic variants discovered to date could serve as the reference for any subsequent parsimonious model development, which may exclude some variables that are more difficult to assess and would make clinical implication of the model less feasible. Models that incorporate both environmental/lifestyle risk factors and common genetic variants may serve as the first step toward developing individually tailored CRC prevention strategies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grant Support: COLO2&3: National Institutes of Health (R01 CA60987).

CPS 2: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort.

DACHS: German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4 and CH 117/1-1), and the German Federal Ministry of Education and Research (01KH0404 and 01ER0814).

DALS: National Institutes of Health (R01 CA48998 to M. L. Slattery).

NHS and HPFS: HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, R01 CA137178, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003), NHS by the National Institutes of Health (R01 CA137178, P01 CA087969, UM1 CA186107, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003).

Kentucky: This work was supported by the following grant support: 1) Clinical Investigator Award from Damon Runyon Cancer Research Foundation (CI-8) and 2) NCI R01CA136726.

MCCS Axiom & OncoArray: MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR) and the Australian Institute of Health and Welfare (AIHW), including the National Death Index and the Australian Cancer Database.

MEC: National Institutes of Health (R37 CA54281, P01 CA033619, and R01 CA63464).

MECC: This work was supported by the National Institutes of Health, U.S. Department of Health and Human Services (R01 CA81488 to SBG and GR).

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 Services (U01 CA74783); 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.

PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics was supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Additionally, a subset of control samples were genotyped as part of the Cancer Genetic Markers of Susceptibility (CGEMS) Prostate Cancer GWAS (Yeager, M et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007 May;39(5):645–9), CGEMS pancreatic cancer scan (PanScan) (Amundadottir, L et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet. 2009 Sep;41(9):986–90, and Petersen, GM et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet. 2010 Mar;42(3):224–8), and the Lung Cancer and Smoking study (Landi MT, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. 2009 Nov;85(5):679–91). The prostate and PanScan study datasets were accessed with appropriate approval through the dbGaP online resource (http://cgems.cancer.gov/data/) accession numbers phs000207.v1.p1 and phs000206.v3.p2, respectively, and the lung datasets were accessed from the dbGaP website (http://www.ncbi.nlm.nih.gov/gap) through accession number phs000093.v2.p2. Funding for the Lung Cancer and Smoking study 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. For the lung study, the GENEVA Coordinating Center provided assistance with genotype cleaning and general study coordination, and the Johns Hopkins University Center for Inherited Disease Research conducted genotyping.

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 HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

GECCO: National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088; R01 CA059045; U01 CA164930).

Abbreviations

CRC	Colorectal cancer
AUC	Area Under the receiver operating characteristic Curve
GECCO	Genetics and Epidemiology of Colorectal Cancer Consortium
CORECT	Colorectal Transdisciplinary study
CCFR	Colorectal Cancer Family Registry
SEER	Surveillance, Epidemiology, and End Results Program registry
Colo2&3	Hawaii Colorectal Cancer Studies 2 and 3
CPS2	Cancer Prevention Study II
DACHS	Darmkrebs: Chancen der Verhütung durch Screening
DALS	Diet, Activity and Lifestyle Study
HPFS	Health Professionals Follow-up Study
Kentucky	Kentucky Case-Control Study
MCCS	Melbourne Collaborative Cohort Study
MEC	Multi-Ethnic Cohort

MECC	Molecular Epidemiology of Colorectal Cancer
NFCCR	Newfoundland Case-Control Study
NHS	Nurses' Health Study
PLCO	Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial
VITAL	VITamins And Lifestyle
WHI	Women's Health Initiative
SNP	Single-nucleotide polymorphism
OR	Odds ratios
CI	Confidence interval
PAR	Population attributable risk
GWAS	Genome-wide association study

References

1. American Cancer Society. Colorectal Cancer Facts & Figures 2014–2016. 2014
2. Bibbins-Domingo K, Grossman DC, et al. US Preventive Services Task Force. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2016; 315:2564–2575. [PubMed: 27304597]
3. US Preventive Services Task Force. [Accessed 10/5, 2016] Final Recommendation Statement: Colorectal Cancer: Screening. 2016. Available at: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/colorectal-cancer-screening2
4. National Center for Health Statistics. [Accessed 10/5, 2016] Table 72 (page 1 of 2). Use of colorectal tests or procedures among adults aged 50–75, by selected characteristics: United States, selected years 2000–2013. 2016. Available at: http://www.cdc.gov/nchs/hus/contents2015.htm#072
5. Inadomi JM. Screening for Colorectal Neoplasia. N Engl J Med. 2017; 376:149–156. [PubMed: 28076720]
6. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011; 3:79re1.
7. Chute CG, Kohane IS. Genomic medicine, health information technology, and patient care. JAMA. 2013; 309:1467–1468. [PubMed: 23571583]
8. Ramsey S, Blough D, McDermott C, et al. Will knowledge of gene-based colorectal cancer disease risk influence quality of life and screening behavior? Findings from a population-based study. Public Health Genomics. 2010; 13:1–12. [PubMed: 20160979]
9. 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. [PubMed: 25683114]
10. Drescher CW, Beatty JD, Resta R, et al. The effect of referral for genetic counseling on genetic testing and surgical prevention in women at high risk for ovarian cancer: Results from a randomized controlled trial. Cancer. 2016; doi: 10.1002/cncr.30190
11. Lieberman D, Ladabaum U, Cruz-Correa M, et al. Screening for Colorectal Cancer and Evolving Issues for Physicians and Patients: A Review. JAMA. 2016; 316:2135–2145. [PubMed: 27893135]
12. Win AK, Macinnis RJ, Hopper JL, Jenkins MA. Risk prediction models for colorectal cancer: a review. Cancer Epidemiol Biomarkers Prev. 2012; 21:398–410. [PubMed: 22169185]
13. 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; 62:871–881. [PubMed: 22490517]

14. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. Cancer Prev Res. 2016; 9:13–26.
15. Jeon J, Meza R, Hazelton WD, Renehan AG, Luebeck EG. Incremental benefits of screening colonoscopy over sigmoidoscopy in average-risk populations: a model-driven analysis. Cancer Causes Control. 2015; 26:859–870. [PubMed: 25783458]
16. Knudsen AB, Zauber AG, Rutter CM, et al. Estimation of Benefits, Burden, and Harms of Colorectal Cancer Screening Strategies: Modeling Study for the US Preventive Services Task Force. JAMA. 2016; 315:2595–2609. [PubMed: 27305518]
17. Ibanez-Sanz G, Diez-Villanueva A, Alonso MH, et al. Risk Model for Colorectal Cancer in Spanish Population Using Environmental and Genetic Factors: Results from the MCC-Spain study. Sci Rep. 2017; 7:43263. [PubMed: 28233817]
18. Murchie B, Tandon K, Hakim S, Shah K, O'Rourke C, Castro FJ. A New Scoring System to Predict the Risk for High-risk Adenoma and Comparison of Existing Risk Calculators. J Clin Gastroenterol. 2017; 51(4):345–351. [PubMed: 27322531]
19. Cao Y, Rosner BA, Ma J, et al. Assessing individual risk for high-risk colorectal adenoma at first-time screening colonoscopy. Int J Cancer. 2015; 137:1719–1728. [PubMed: 25820865]
20. Yeoh KG, Ho KY, Chiu HM, et al. The Asia-Pacific Colorectal Screening score: a validated tool that stratifies risk for colorectal advanced neoplasia in asymptomatic Asian subjects. Gut. 2011; 60:1236–1241. [PubMed: 21402615]
21. Hutter CM, Chang-Claude J, Slattery ML, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. Cancer Res. 2012; 72:2036–2044. [PubMed: 22367214]
22. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J.R. Statist. Soc. B. 2006; 68:49–67.
23. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet. 2007; 39:989–994. [PubMed: 17618283]
24. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet. 2008; 40:631–637. [PubMed: 18372901]
25. 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; 42:973–977. [PubMed: 20972440]
26. Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, et al. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. PLoS Genet. 2011; 7:e1002105. [PubMed: 21655089]
27. Zhang B, Jia WH, Matsuda K, et al. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat Genet. 2014; 46:533–542. [PubMed: 24836286]
28. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. Gut. 2015; 64:1623–1636. [PubMed: 26187503]
29. 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. [PubMed: 25990418]
30. Schumacher FR, Schmit SL, Jiao S, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun. 2015; 6:7138. [PubMed: 26151821]
31. Zeng C, Matsuda K, Jia WH, et al. Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. Gastroenterology. 2016; 150:1633–1645. [PubMed: 26965516]
32. Peters U, Jiao S, Schumacher FR, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology. 2013; 144:799–807.e24. [PubMed: 23266556]
33. Hopper JL. Odds per adjusted standard deviation: comparing strengths of associations for risk factors measured on different scales and across diseases and populations. Am J Epidemiol. 2015; 182:863–867. [PubMed: 26520360]

34. Krishnan K, Baglietto L, Apicella C, et al. Mammographic density and risk of breast cancer by mode of detection and tumor size: a case-control study. Breast Cancer Res. 2016; 18 63-016-0722-4.
35. Janes H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am J Epidemiol. 2008; 168:89–97. [PubMed: 18477651]
36. Surveillance Epidemiology and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER 9 Regs Research Data, Nov 2011 Sub (1973–2010). - Linked To County Attributes - Total U.S., 1969–2010 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2013, based on the November 2012 submission. (http://www.seer.cancer.gov)
37. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989; 81:1879–1886. [PubMed: 2593165]
38. Rahman R, Schmaltz C, Jackson CS, Simoes EJ, Jackson-Thompson J, Ibdah JA. Increased risk for colorectal cancer under age 50 in racial and ethnic minorities living in the United States. Cancer Med. 2015; 4:1863–1870. [PubMed: 26471963]
39. Ginsburg G. Medical genomics: Gather and use genetic data in health care. Nature. 2014; 508:451–453. [PubMed: 24765668]
40. Shirts BH, Salama JS, Aronson SJ, et al. CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record. J Am Med Inform Assoc. 2015; 22:1231–1242. [PubMed: 26142422]
41. Hartzler A, McCarty CA, Rasmussen LV, et al. Stakeholder engagement: a key component of integrating genomic information into electronic health records. Genet Med. 2013; 15:792–801. [PubMed: 24030437]
42. National Cancer Institute. [Accessed 11/12/2017] Colorectal Cancer Risk Assessment Tool. 2014. Available at: https://www.cancer.gov/colorectalcancerrisk/
43. Siteman Cancer Center. [Accessed 11/12/2017] Your Disease Risk. 2013. Available at: http://yourdiseaserisk.wustl.edu/YDRDefault.aspx?ScreenControl=YDRGeneral&ScreenName=YDRcolon
44. Brenner H, Stock C, Hoffmeister M. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. BMJ. 2014; 348:g2467. [PubMed: 24922745]
45. Weigl K, Jansen L, Chang-Claude J, Knebel P, Hoffmeister M, Brenner H. Family history and the risk of colorectal cancer: The importance of patients' history of colonoscopy. Int J Cancer. 2016; 139:2213–2220. [PubMed: 27459311]
46. Taylor DP, Burt RW, Williams MS, Haug PJ, Cannon-Albright LA. Population-based family history-specific risks for colorectal cancer: a constellation approach. Gastroenterology. 2010; 138:877–885. [PubMed: 19932107]
47. Jiao S, Peters U, Berndt S, et al. Estimating the heritability of colorectal cancer. Hum Mol Genet. 2014; 23:3898–3905. [PubMed: 24562164]
48. Wei Z, Wang W, Bradfield J, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet. 2013; 92:1008–1012. [PubMed: 23731541]
49. Zheng, Y., Hua, X., Win, AK., Jenkins, M., Macinnis, R., Newcomb, P. Does a comprehensive family history of colorectal cancer improve risk prediction?. In: Colditz, G.Gapstur, SM.Muir, KR., Sherman, ME., editors. Proceedings of An AACR Special Conference on Improving Cancer Risk Prediction for Prevention and Early Detection. Orlando, Florida, USA: Nov 16–19, 2016
50. Zheng C, Beresford SA, Van Horn L, et al. Simultaneous association of total energy consumption and activity-related energy expenditure with risks of cardiovascular disease, cancer, and diabetes among postmenopausal women. Am J Epidemiol. 2014; 180:526–535. [PubMed: 25016533]

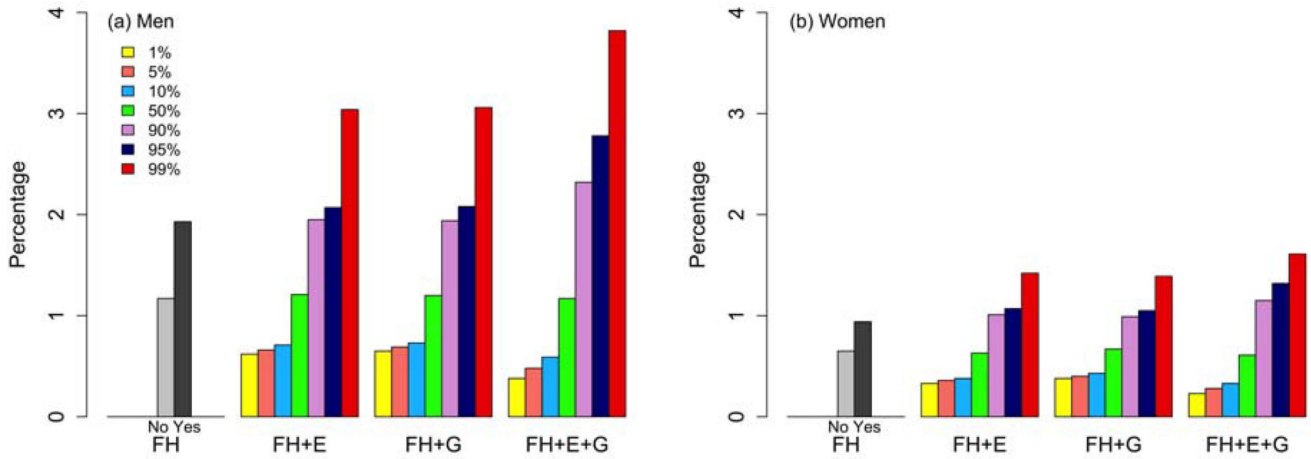
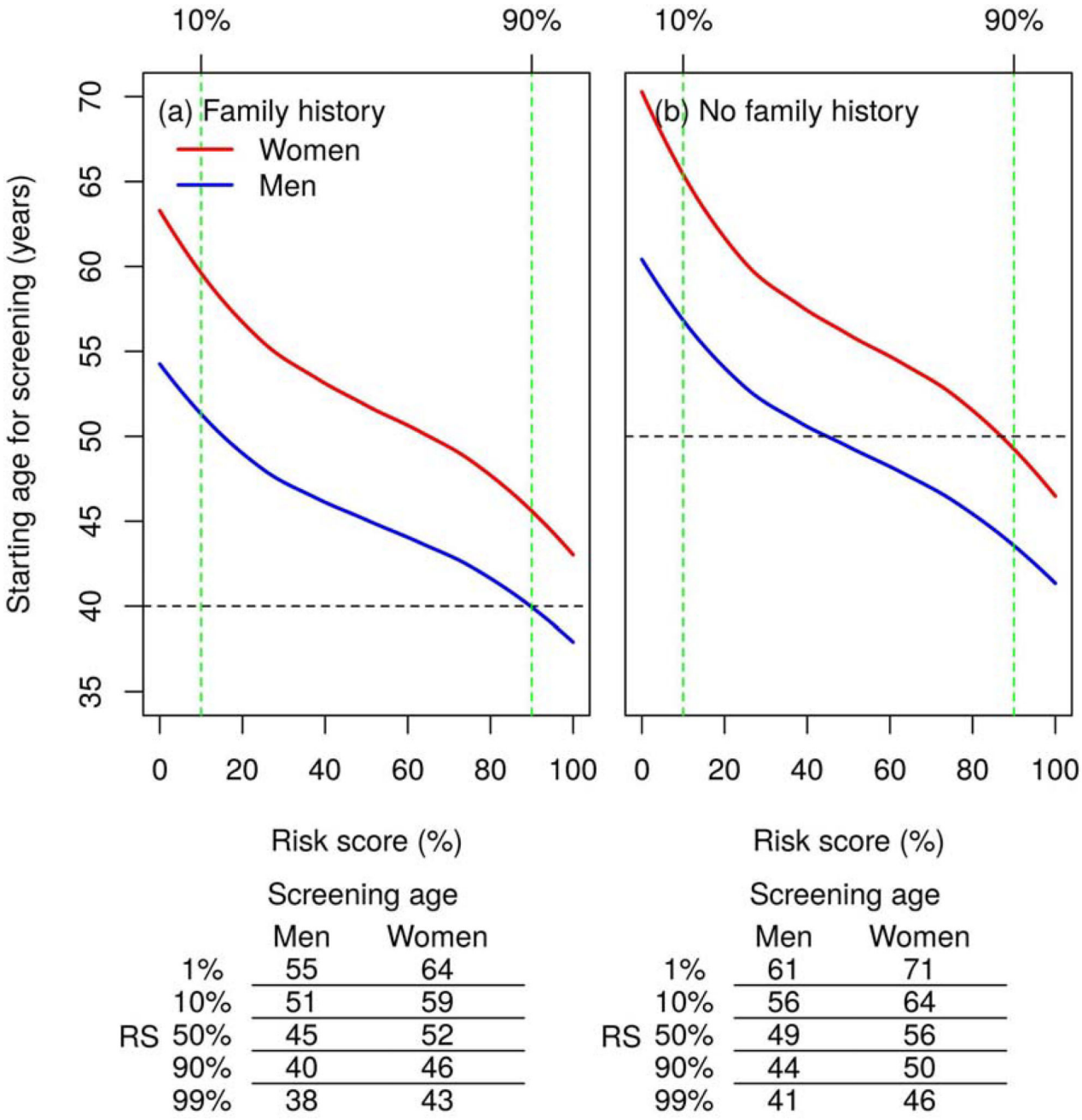


Figure 1.

Estimate of 10-year absolute risk of CRC at 1st, 5th, 10th, 50th, 90th, 95th, 99th percentiles of risk score under various models for an individual aged 50 years who did not have a prior endoscopy. FH: family history only model, FH+E: family history and environmental risk score (E-score) model, FH+G: family history and genetic risk score (G-score) model, FH+E +G: family history and E-score and G-score model.



Screening age

RS		Men	Women
	1%	55	64
	10%	51	59
RS	50%	45	52
	90%	40	46
	99%	38	43

Screening age

RS		Men	Women
	1%	61	71
	10%	56	64
RS	50%	49	56
	90%	44	50
	99%	41	46

Figure 2.

Recommended age to start CRC screening by various risk scores, which was based on both environmental risk score (E-score) and genetic risk score (G-score). The horizontal lines represent the recommended age for the first endoscopy depending on family history in the current screening guideline for CRC. The risk threshold to determine the age for the first screening was set as the average of 10-year CRC risks for a 50-year-old man (1.25%) and woman (0.68%), i.e., (1.25%+0.68%)/2=0.97%, who have not previously received an endoscopy.

Table 1

Descriptive characteristics of environmental and lifestyle risk factors in study population of the training dataset.

	Men		Women	
Variable	**Cases (N=2307)**	**Controls (N=2359)**	**Cases (N=2568)**	**Controls (N=2932)**
Age				
Mean (SD)	67.8 (9.7)	68.0 (10.0)	68.8 (9.7)	69.7 (8.8)
Height (cm)				
Mean (SD)	175.8 (7.4)	176.1 (7.4)	162.5 (6.3)	162.4 (6.4)
BMI (kg/m^2)				
Mean (SD)	27.8 (4.3)	27.1 (3.8)	27.4 (5.3)	26.6 (4.9)
Family history				
Yes (%)	333 (14.4)	250 (10.6)	401 (15.6)	368 (12.6)
Endoscopy history[1]				
Yes (%)	451 (19.5)	838 (35.5)	692 (26.9)	1071 (36.5)
Education[2]				
Cat1 (%)	355 (15.4)	306 (13.0)	438 (17.1)	440 (15.0)
Cat2 (%)	699 (30.3)	597 (25.3)	744 (29.0)	717 (24.5)
Cat3 (%)	525 (22.8)	588 (24.9)	650 (25.3)	755 (25.8)
Cat4 (%)	699 (30.3)	830 (35.2)	705 (27.5)	989 (33.7)
Diabetes				
Yes (%)	310 (13.4)	269 (11.4)	269 (10.5)	204 (7.0)
Lifestyle factors				
Physical activity				
Yes (%)	1111 (48.2)	1212 (51.4)	1037 (40.4)	1116 (38.1)
Smoking status				
Ever smoker (%)	1526 (66.1)	1445 (61.3)	1124 (43.8)	1292 (44.1)
Smoking pack-years[3]				
Mean (SD)	1.7 (1.5)	1.6 (1.4)	1.1 (1.5)	1.1 (1.4)
Alcohol consumption				
< 1g/day (%)	742 (32.2)	764 (32.4)	1427 (55.6)	1556 (53.1)
1–28 g/day (%)	918 (39.8)	1086 (46.0)	831 (32.4)	1080 (36.8)
>28 g/day (%)	419 (18.2)	329 (13.9)	115 (4.5)	138 (4.7)
Pharmaceutical factors				
Aspirin use				
Yes (%)	667 (28.9)	822 (34.8)	572 (22.3)	783 (26.7)
NSAIDs use				
Yes (%)	171 (7.4)	217 (9.2)	440 (17.1)	596 (20.3)
Post-menopausal hormone use				
Yes (%)	-	-	693 (27.0)	1085 (37.0)
Dietary factors				

	Men		Women	
Variable	**Cases (N=2307)**	**Controls (N=2359)**	**Cases (N=2568)**	**Controls (N=2932)**
Fiber[4]				
Mean (SD)	1.6 (0.9)	1.5 (0.9)	1.6 (1.0)	1.5 (1.0)
Calcium[4]				
Mean (SD)	1.6 (1.1)	1.5 (1.1)	1.7 (1.1)	1.5 (1.1)
Folate[4]				
Mean (SD)	1.6 (0.9)	1.5 (0.9)	1.6 (1.0)	1.5 (1.0)
Processed meat[4]				
Mean (SD)	1.5 (1.0)	1.4 (1.0)	1.5 (1.0)	1.4 (1.0)
Red meat[4]				
Mean (SD)	1.5 (1.0)	1.3 (1.0)	1.5 (1.1)	1.4 (1.1)
Fruit[4]				
Mean (SD)	1.7 (0.9)	1.7 (0.9)	1.7 (1.0)	1.6 (1.0)
Vegetable[4]				
Mean (SD)	1.7 (0.9)	1.6 (0.9)	1.6 (1.0)	1.5 (1.0)
Total energy				
Mean (SD)	2265.1 (702.7)	2249.7 (678.2)	1679.6 (549.8)	1720.5 (587.1)
Combined risk scores				
E-score				
Mean (SD)	59.9 (28.5)	49.6 (29.1)	59.9 (27.9)	49.5 (29.0)
G-score				
Mean (SD)	58.7 (28.4)	48.9 (29.2)	58.2 (28.2)	50.0 (28.6)

[1]Endoscopy history was entirely missing in five studies (MCCS, MECC, Kentucky, NFCCR, Colo2&3), so these studies were excluded when we compute the summary statistic for endoscopy history.

[2]Education variable has four categories. Cat1: less than high school graduate, Cat2: high school graduate or completed GED, Cat3: some college or technical school, Cat4: college graduate or more.

[3]Smoking pack-years among ever smokers was harmonized across studies by sex- and study-specific quartiles, and assigned values 1,2,3,4. For never smokers, it was assigned as “0”. This variable was treated as continuous variable in the analysis.

[4]Dietary variables (fiber, calcium, folate, processed meat, red meat, fruit, vegetable) were harmonized across studies by sex- and study-specific quartiles, and assigned values 0,1,2,3 in the order of increasing risk marginally. These variables were treated as continuous variables in the analysis.

Table 2

Odds ratio (95% CI) of risk factors associated with CRC risk in the model building dataset.

	Men (N= 4,666)	Women (N= 5,500)
Variable	**OR (95% CI)**	**OR (95% CI)**
E-score *	1.36 (1.29 to 1.44)	1.35 (1.28 to 1.42)
G-score *	1.34 (1.27 to 1.42)	1.30 (1.23 to 1.36)
Family history		
No	1.00	1.00
Yes	1.67 (1.38 to 2.03)	1.46 (1.24 to 1.72)
Endoscopy history		
No	1.00	1.00
Yes	0.29 (0.24 to 0.34)	0.53 (0.47 to 0.61)
Missing	0.46 (0.30 to 0.71)	0.67 (0.49 to 0.91)

The logistic regression model includes study, age, E-score, G-score, family history, endoscopy history, genotype platform, and PCs.

* ORs for E-score and G-score per quartile increase

Table 3

AUC comparisons between risk prediction models in the validation dataset.

	Men (N= 4,658)		Women (N= 5,514)	
	AUC (95% CI)		AUC (95% CI)	
Model I				
Family history	0.53 (0.52 to 0.54)		0.54 (0.52 to 0.55)	
Model II				
Family history & E-score	0.60 (0.59 to 0.61)		0.60 (0.59 to 0.61)	
		$P_{II\ vs.\ IV} = 1.0\times10^{-11}$		$P_{II\ vs.\ IV} = 6.4\times10^{-12}$
Model III				
Family history & G-score	0.59 (0.58 to 0.60)		0.59 (0.58 to 0.60)	
		$P_{III\ vs.\ IV} \sim 0$		$P_{III\ vs.\ IV} \sim 0$
Model IV				
Family history & E-score & G-score	0.63 (0.62 to 0.64)		0.62 (0.61 to 0.63)	

The analyses were adjusted for study, age, and endoscopy history.

Table 4

Examples of 10-year absolute risk estimates for CRC with different risk factor profiles.

	Endoscopy	Family history	E-score	G-score	Men		Women	
					Risk (%)	95% CI	Risk (%)	95% CI
Age=50								
Reference	Average risk[1]				0.69		0.49	
	Average risk (no endoscopy)[2]				1.25	1.03 to 1.47	0.68	0.60 to 0.76
	No	No	1%	1%	0.34	0.26 to 0.43	0.20	0.16 to 0.24
	No	No	10%	10%	0.42	0.32 to 0.52	0.24	0.20 to 0.28
	No	No	50%	50%	1.11	0.89 to 1.34	0.58	0.51 to 0.65
	No	No	90%	90%	2.90	2.21 to 3.58	1.41	1.17 to 1.64
	No	No	99%	99%	3.59	2.68 to 4.50	1.72	1.40 to 2.03
	No	Yes	1%	1%	0.57	0.40 to 0.74	0.29	0.22 to 0.36
	No	Yes	10%	10%	0.71	0.51 to 0.91	0.35	0.27 to 0.43
	No	Yes	50%	50%	1.86	1.38 to 2.33	0.85	0.69 to 1.01
	No	Yes	90%	90%	4.80	3.42 to 6.19	2.05	1.61 to 2.49
	No	Yes	99%	99%	5.93	4.15 to 7.72	2.49	1.92 to 3.06
	Yes	No	1%	1%	0.10	0.07 to 0.12	0.11	0.08 to 0.13
	Yes	No	10%	10%	0.12	0.09 to 0.15	0.13	0.11 to 0.15
	Yes	No	50%	50%	0.32	0.25 to 0.39	0.31	0.27 to 0.35
	Yes	No	90%	90%	0.84	0.64 to 1.04	0.76	0.63 to 0.88
	Yes	No	99%	99%	1.04	0.77 to 1.31	0.92	0.75 to 1.10
	Yes	Yes	1%	1%	0.16	0.12 to 0.21	0.15	0.12 to 0.19
	Yes	Yes	10%	10%	0.20	0.15 to 0.26	0.19	0.15 to 0.23
	Yes	Yes	50%	50%	0.53	0.40 to 0.67	0.45	0.37 to 0.54
	Yes	Yes	90%	90%	1.40	1.00 to 1.80	1.10	0.87 to 1.33
	Yes	Yes	99%	99%	1.73	1.21 to 2.26	1.34	1.04 to 1.64

[1] Average risks in general population were calculated based on SEER incidence rates for men and women separately.

[2] Average risks among people who have not had prior endoscopy were calculated based on a model including endoscopy only and the SEER incidence rates for men and women separately.