Supplemental Methods. Details of Study Population, Risk Model Development, Calculation of the Polygenic Risk Score, Statistical Analysis, Data Availability and Funding of Individual Studies in GECCO and CORECT

Genetic Epidemiology Research on Adult Health and Aging (GERA)

The Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort comprised study participants (survey respondents) from the Research Program on Genes, Environment, and Health (RPGEH) and California Men's Health Study (CMHS), both nested within the member population of Kaiser Permanente Northern California (KPNC), a large integrated healthcare delivery system. The KPNC member population includes 30-40% of the general population in a 22-county geographic area of northern California and is broadly representative of the general population,[1] although extremes of income are underrepresented. The RPGEH is a contemporary cohort. Construction of this cohort began in 2007, when a six-page survey was mailed to 1.9 million individuals in the member population who were ages 18 and older and who had been previously enrolled health plan members for at least two years. This survey was designed to ascertain data on demographic and lifestyle characteristics, including race/ethnicity, education, income, marital status, self- and family history for 35 selected conditions, diet and physical activity, smoking and alcohol consumption, and reproductive history and health. In July 2008, approximately 400,000 survey respondents were asked to sign a consent form authorizing broad use of their survey data, longitudinal electronic health record data, and biospecimens in research on genetic and environmental factors associated with health and disease. Those who provided consent were mailed saliva DNA collection (Oragene) kits. In 2009, over 40,000 men ages 45 to 69 years, who were KPNC health plan members and had enrolled in the CMHS in 2002–2003, were similarly asked to provide saliva samples. At CMHS enrollment, men completed mailed surveys to ascertain data on demographic and lifestyle factors, akin to that of RPGEH.

In total, 110,266 consenting participants who provided saliva samples were selected for genotyping. All racial and ethnic minority participants with saliva samples (n = 20,935, 19%) were included to maximize diversity and a random subset were selected from the approximately 140,000 available non-Hispanic White participants with biospecimens, including participants in both the RPGEH and CMHS who provided consent and a biospecimen. The demographic characteristics of the GERA cohort are generally representative of the RPGEH and CMHS survey respondents, as well as the KPNC member population, although GERA cohort members are on average older (average age at time of sample collection: 63 years); the proportion of minority participants is lower, and the average level of education and income is slightly higher than in the KPNC member population. However, the GERA cohort does include representation of all sociodemographic groups included in the adult KPNC member population (personal communication with Dr. Catherine Schaefer). Four custom Affymetrix Axiom arrays were designed for genotyping, one for each major ancestral group represented in the GERA cohort: African American, Asian, European, and Latinx. As detailed elsewhere,[2] the selected number of SNPs and SNP content varied by array in order to maximize coverage of the whole genome, along with common and low frequency SNPs specific to race/ethnicity and known SNPs associated with disease phenotypes. Details on the calling and quality control have been described previously.[3]

For the present study, colorectal cancer status was determined for study participants from study entry to December 31, 2016, by linkage to the KPNC Cancer Registry, a current database of all patients with newly diagnosed cancers at KPNC facilities that adheres to the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program standards. Personal history of cancer was ascertained from cancer registry and electronic health record data. A family history of CRC was determined by integrating data from baseline surveys and electronic health records (i.e., diagnosis codes, family history documentation). Endoscopy history information for all participants was also extracted from medical records. As a

cohort unselected on any disease phenotype, GERA participants were not asked to engage in specific medical or screening tests for research purposes. However, since the majority (69%) of GERA participants were age 55 and older at baseline, most but not all had undergone screening for colorectal cancer, either by fecal immunochemical testing (FIT) or endoscopy (sigmoidoscopy or colonoscopy). At their baseline questionnaire, 70% were up to date for colorectal cancer screening. All study participants provided written informed consent, and the study was approved by the KPNC Institutional Review Board.

Identification of 140 CRC Known Loci

A total of 140 CRC known loci were identified from our large collaborative network comprised of the Colon Cancer Family Registry (CCFR), CORECT and GECCO (58,131 CRC cases and 67,347 controls) as well as previously published genome-wide association studies.[4–6] Among them, 13 loci were discovered in a large East Asian cohort[6] while the rest in cohorts predominantly of European ancestry.[4,5] These consist of 115 known loci that reached genome-wide significance at $5x10^{-8}$, and 25 secondary independent loci in known GWAS regions with p-values < $1x10^{-5}$ in conditional association analyses adjusted for the known loci in the region.

Estimating Effect Sizes and Baseline CRC Hazard Rate

GECCO and CORECT included both population-based case-control and cohort-based nested case-control studies, where the risk factors were collected at the time of case-control ascertainment for the case-control studies and at the cohort entry for nested case-control studies. To account for the potential difference between the two study designs so that the model can be properly validated in a prospective cohort setting such as the GERA, we incorporated the interaction effects between the cohort indicator and the two risk factors, endoscopy history and family history.

The baseline CRC hazard rates were estimated by the age- and sex-specific SEER CRC-incidence rates in Europeans (SEER registry 18, 2007 to 2015, https://seer.cancer.gov/) multiplied by 1 - population attributable risk. The population attributable risk was estimated by taking the average of inverse exponential risk scores among all cases in the GECCO and CORECT dataset for model development.[7,8] Competing risks from death of other causes were accounted for in the calculation of absolute CRC risk by incorporating other-cause age-specific mortality rates in Europeans using SEER 18 (2007-2015) data.

Calculation of Polygenic Risk Score in Validation Dataset

Out of 140 CRC known loci, 3 were not available in the GERA (rs755229494, rs373585858, and rs556532366), and we included the most correlated surrogates in Europeans (rs112334046, R2=0.40, minor allele frequency [MAF] =0.0026; rs187032491, R2=1, MAF=0.0043; and rs181781440, R2=0.45, MAF=0.0015). Similar to the PRS of GECCO and CORECT participants, we calculated the PRS of GERA participants as a weighted sum of numbers of effective alleles of the 140 known loci with their estimated marginal log-odds ratios as weights.[4]

Statistical Analysis

We define the observed time-to-event from study entry to the earliest of the following events: CRC diagnosis, death, or last follow-up. Because colonoscopy may remove precursor lesions, participants who received colonoscopy screening after study entry were censored at 6 months post first colonoscopy screening; the 6-month interval was to allow for the work-up of colonoscopy results or repeating an inadequate examination.

Calibration We compared the expected (E) and observed (O) numbers of CRC cases in t years since study entry to evaluate the calibration of predicted absolute risk based on the proposed

risk prediction model. The expected number of CRC cases at t-year was calculated by summing the projected absolute risk estimate for each participant, given their baseline risk profiles, at either the observed time-to-event or t year, whichever comes first. We assessed the calibration overall and within subgroups defined by age (40-49, 50-59, ≥60), sex (men, women), endoscopy history within 10 years prior to study entry (yes, no), and family history (yes, no). The 95% confidence intervals (CIs) for the t-year E/O ratio were calculated by using the normal approximation to Poisson distributions as E/O $exp(\pm 1.96\sqrt{\frac{1}{O}})$.

We further evaluated t-year calibration across the CRC-risk spectrum. Participants were grouped into 10 equal risk strata by their t-year absolute risk estimates. The observed t-year absolute risk in the j[th] risk stratum is determined by $1 - \hat{S}_j(t)$, where $\hat{S}_j(t)$ is the Kaplan-Meier (KM) estimator for CRC-free probability accounting for right-censoring and competing risk[9,10] based on the individuals in that stratum. The 95% CIs were obtained for the KM estimator. Within each risk stratum, we compared the observed absolute risk to the expected absolute risk which is defined as the average of the model-predicted absolute risk. We plotted the observed and expected t-year absolute risk values across the CRC-risk spectrum and assessed any potential under/over-estimation of the risk model on the CRC-risk over the entire risk range than overall calibration.

As a secondary analysis, we evaluated the calibration of the PRS alone on relative risk (RR) in Europeans and other racial and ethnic groups, including African Americans, Asians, and Latinx. The purpose is to examine how well the calibration for PRS by itself is across ethnicities. Participants in each racial and ethnic group were equally divided into 7 strata by their PRS, and the middle stratum that include the 50[th] percentile of the PRS is set as the reference. The number of strata, 7, was chosen to ensure a fair number of CRC cases within each stratum for reliable estimation of RR across all ethnicities. The expected RR for a PRS stratum is the ratio

of the within-stratum geometric average of individuals' model-based RR, defined as individuals' PRS times the estimated PRS effect, between that stratum and the reference stratum. The observed RR and its 95% CI were obtained by fitting a Cox model with a 0-1 stratum indicator as a covariate including only the specific stratum and the reference stratum. We plotted the observed and the expected RRs in a log-scale across PRS strata for each racial/ethnic group.

Development of The Two Reduced Models (Model 1 and 2) Model 1 included age and family history. The association of family history was estimated by a logistic regression model with family history adjusting for sex, age and an interaction between study type (cohort vs. case-control) and family history fitted to GECCO and CORECT as described above. The baseline CRC hazard rates were derived using age-specific CRC incidence rates in Europeans in SEER 18 (2007-2015). Model 2 included age, family history, sex, and endoscopy history in prior 10 years. We referred to the section of "Estimating Effect Sizes and Baseline CRC Hazard Rate" above for detailed model development as Model 2 was developed using the same approach for the proposed PRS-enhanced model.

References

1. Gordon N, Lin T. The Kaiser Permanente Northern California Adult Member Health Survey. *Perm J*. 2016;20(4):15-225. doi:10.7812/TPP/15-225

2. Hoffmann TJ, Kvale MN, Hesselson SE, et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics*. 2011;98(2):79-89. doi:10.1016/j.ygeno.2011.04.005

3. Kvale MN, Hesselson S, Hoffmann TJ, et al. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics*. 2015;200(4):1051-1060. doi:10.1534/genetics.115.178905

4. Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet*. 2019;51(1):76-87. doi:10.1038/s41588-018-0286-6

5. The PRACTICAL consortium, Law PJ, Timofeeva M, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun*. 2019;10(1):2154. doi:10.1038/s41467-019-09775-w

6. Lu Y, Kweon SS, Tanikawa C, et al. Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer. *Gastroenterology*. 2019;156(5):1455-1466. doi:10.1053/j.gastro.2018.11.066

7. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol*. 1985;122(5):904-914. doi:10.1093/oxfordjournals.aje.a114174

8. Jeon J, Du M, Schoen RE, et al. Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. *Gastroenterology*. 2018;154(8):2152-2164.e19. doi:10.1053/j.gastro.2018.02.021

9. Aalen O. Nonparametric Estimation of Partial Transition Probabilities in Multiple Decrement Models. *Ann Statist*. 1978;6(3). doi:10.1214/aos/1176344198

10. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data: Kalbfleisch/The Statistical*. John Wiley & Sons, Inc.; 2002. doi:10.1002/9781118032985

Acknowledgement Section

Funding of individual studies in GECCO and CORECT

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA164930, U01 CA137088, R01 CA059045, R01201407). Genotyping/Sequencing services were provided by the Center for Inherited Disease Research (CIDR) (X01-HG008596 and X-01-HG007585). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268201200008I. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704.

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

Colorectal Cancer Transdisciplinary (CORECT) Study: The CORECT Study was supported by the National Cancer Institute, National Institutes of Health (NCI/NIH), U.S. Department of Health and Human Services (grant numbers U19 CA148107, R01 CA81488, P30 CA014089, R01 CA197350; P01 CA196569; R01 CA201407) and National Institutes of Environmental Health Sciences, National Institutes of Health (grant number T32 ES013678).

CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. This study was conducted with Institutional Review Board approval.

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 and 01ER1505B).

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

GERA: 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 [R01CA206279 (Peters, Corley, and Hayes) and UM1CA222035 (Corley and Lee)]. Access to the GERA 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 NIH/dbGaP: phs000674.

Harvard cohorts (HPFS, NHS): HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, U01 CA167552, R01 CA137178, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003) and NHS by the National Institutes of Health (R01 CA137178, P01 CA087969, UM1 CA186107, R01 CA151993, R35 CA197735, K07CA190673, and P50 CA127003).

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

MCCS: Cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 509348, 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 CA063464).

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 Serivces (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 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.

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.

Additional Contributions

CPS-II: The authors thank the CPS-II participants and Study Management Group for their invaluable contributions to this research. The authors would also like to acknowledge the

contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention National Program of Cancer Registries, and cancer registries supported by the National Cancer Institute Surveillance Epidemiology and End Results program.

DACHS: We thank all participants and cooperating clinicians, and Ute Handte-Daub, Utz Benscheid, Muhabbet Celik and Ursula Eilber for excellent technical assistance.

Harvard cohorts (HPFS, NHS): 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 would like to thank the participants and staff of the HPFS and NHS for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data.

Kentucky: We would like to acknowledge the staff at the Kentucky Cancer Registry.

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.

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: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf