

Yale University

## EliScholar – A Digital Platform for Scholarly Publishing at Yale

---

Public Health Theses

School of Public Health

---

January 2023

### The Genetic Generalizability Gap Of Colorectal Cancer: Assessment By Polygenic Risk Score

Lauren R. Lautermilch  
laurenlautermilch@gmail.com

Follow this and additional works at: <https://elischolar.library.yale.edu/ysphtdl>

---

#### Recommended Citation

Lautermilch, Lauren R., "The Genetic Generalizability Gap Of Colorectal Cancer: Assessment By Polygenic Risk Score" (2023). *Public Health Theses*. 2286.  
<https://elischolar.library.yale.edu/ysphtdl/2286>

This Open Access Thesis is brought to you for free and open access by the School of Public Health at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Public Health Theses by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact [elischolar@yale.edu](mailto:elischolar@yale.edu).

THE GENETIC GENERALIZABILITY GAP OF COLORECTAL CANCER: ASSESSMENT  
BY POLYGENIC RISK SCORE

By

Lauren Lautermilch

May, 2023

A Thesis

Submitted to the Department of Chronic Disease Epidemiology

Yale School of Public Health

In Partial Fulfilment of the Requirements

For the Degree of Masters of Public Health

Advisor: Dr. Andrew Dewan, Ph.D., MPH  
Committee Member: Dr. Tormod Rogne, MD, Ph.D.

**ABSTRACT:**

Colorectal cancer (CRC) is one of the most common incident cancers in developed countries. Though the majority of colorectal cancer cases do not arise in individuals with inherited pathogenic mutations or a family history of CRC, there is a substantial heritable component of colorectal cancer. Thus, there is a critical need to incorporate genetic risk factors into overall CRC risk assessment to ultimately improve the efficacy of current screening strategies and optimize precision medicine approaches.

Presented here is a polygenic risk score (PRS) utilizing 79 genetic variants for colorectal cancer. The PRS was generated using genome-wide significant single nucleotide polymorphisms (SNPs) as identified from a meta-analysis of genome-wide association studies (GWAS) among individuals of European and east Asian genetic ancestry (100,204 cases and 154,587 controls). This PRS was then applied to 9,304 colorectal cancer cases and 477,902 controls from the UK Biobank (UKB) of multiple different races/ethnicities.

There was a statistically significant difference in mean PRS value in white, black, and Asian subpopulations, indicating acceptable discriminatory accuracy with regards to CRC case-control status. The greatest discriminatory accuracy was demonstrated among white participants, followed by Asian participants. Due to the reference population from which the PRS was constructed consisting only of European and east Asian genetic ancestry, the PRS demonstrated reduced predictive power among black participants. The combination of a significant difference in mean PRS value, but relatively poor predictive power among this subpopulation is indicative of the genetic generalizability gap of colorectal cancer. Therefore, it is essential to explore other risk-associated SNPs specific to African ancestry to further elucidate the unique genetic etiology of CRC.

**TABLE OF CONTENTS**

1. ABSTRACT .....	P.2
2. STATEMENT OF THE PROBLEM .....	P.4
a. Epidemiology of colorectal cancer (CRC)	
b. Genetics of CRC	
c. Polygenic risk scores (PRS): utility and shortcomings	
d. Framework of investigation	
3. MATERIALS AND METHODS .....	P.7
a. UK Biobank (UKB) study sample	
b. Generation of PRS	
c. Statistical analyses	
4. RESULTS .....	P.9
5. DISCUSSION.....	P.11
a. Novelty of the investigation	
b. Limitations of analysis	
c. Future Directions	
6. REFERENCES .....	P.16
7. APPENDIX .....	P.18

**INTRODUCTION:** Colorectal cancer (CRC) is the third most common cancer in the United States<sup>1</sup> and fourth most common cause of cancer death globally.<sup>2</sup> Though CRC mortality has declined significantly in the United States due to the efficacious implementation of screening programs, this cancer remains a burden that accounts for 8.3% of U.S. cancer deaths.<sup>1</sup> Colorectal cancer is defined as malignant cancer that originates in the colon and/or the rectum. More than 90% of colorectal cancers are adenocarcinomas, which develop from the glandular epithelial cells of the colorectum.<sup>3</sup> Approximately two-thirds (60-65%) of all CRC cases can be considered sporadic,<sup>3</sup> meaning they arise in individuals without family history and/or inherited pathogenic genetic mutations. These sporadic cases, therefore, are a result of acquired somatic genome alterations that may be caused by a variety of behavioral and biological factors. Specifically, risk increases with age (particularly above age 50), male sex, and black race.<sup>2</sup> Among modifiable risk factors are obesity, low physical activity, type 2 diabetes, irritable bowel disease, smoking, alcohol consumption, and diet (including consumption of red and processed meat, and low fiber and whole grain consumption).<sup>3,4</sup> Emerging evidence additionally suggests that colorectal cancer risk may be associated with the abundance of certain gut microbiota, indicating a possible role of antibiotic use in CRC risk.<sup>4</sup>

Colorectal cancer has a uniquely large heritable component as compared to other common malignant neoplasms. It is estimated that CRC has a heritability of 35%.<sup>5</sup> Approximately 5% of colorectal cancers are due to well-defined, familial cancer syndromes, including familial adenomatous polyposis, Lynch syndrome, MUTYH-associated polyposis, and some hamartomatous polyposis conditions.<sup>6</sup> The remaining 25% of inherited CRC cases occur in individuals with a family history of CRC, but lack any of the obvious aforementioned cancer syndromes.<sup>3,6</sup> These cancers are likely to be caused by less penetrant, but possibly more common

genetic alterations, or single nucleotide polymorphisms (SNPs). Furthermore, it is probable that these inherited cases of CRC are caused by alterations at multiple genetic loci that are collectively carcinogenic. In aggregate, such genetic loci that have been identified from genome-wide association studies (GWAS) explain 0.65% of all known heritability.<sup>7</sup> Specifically, GWAS have identified approximately 70 common genetic variants associated with of colorectal cancer.<sup>8</sup>

Despite the large burden of colorectal cancer in the United States, it is among the most preventable cancers due the utilization of effective screening techniques (e.g., fecal-based tests including fecal immunochemical test and fecal occult blood test, flexible sigmoidoscopy, and colonoscopy). As of 2016, the 5-year relative survival for distant, late-stage CRC cases is below 15%, while regional, early-stage cases have a 5-year survival that is slightly below 90%.<sup>9</sup> It is obvious that early detection is essential to reducing colorectal cancer mortality. A meta-analysis of current global (mainly North America, Europe, and Asia) CRC screening guidelines suggests that screening recommendations are largely based upon age and family history.<sup>10</sup> Therefore, it emerges that genetic risk prediction could not only inform screening recommendations to improve survival, but also guide targeted interventions through precision medicine (e.g., specific chemotherapy and chemo-preventive drugs, diet and lifestyle modifications).<sup>11</sup>

The information gained from GWAS has been utilized to develop predictive genetic tools including genetic risk scores (GRS) and polygenic risk scores (PRS). Both scores are single-value estimates of an individual's genetic liability to a trait or disease. While a GRS measures genetic predisposition to a particular outcome through examination of a set of genome-wide significant SNPs, a PRS extends this to include many SNPs, including those beyond genome-wide significance that tend to have smaller effects.<sup>12</sup> Thus, the PRS – with enhanced predictive power – is useful when examining an outcome that is influenced by either rare variants or many variants

with small effect sizes. In consideration of the high heritability and genetic etiology of colorectal cancer, polygenic risk scores may be apt to predict genetic CRC risk.

Previous approaches that have utilized genetic susceptibility in predicting colorectal cancer have successfully incorporated genetic variants as identified by GWAS.<sup>20-23</sup> These GRS have utilized upwards of 63 SNPs, with the most recent yielding moderate predictive power [area under receiver operating curve (AUC) = 0.59].<sup>20</sup> It has been demonstrated that PRS approaches for CRC risk prediction, including linkage disequilibrium (LD) clumping PRS and LDpred PRS (a Bayesian approach for risk prediction), yield better predictive performance, which is likely due to the inclusion of significant disease-associated SNPs across the entirety of the genome.<sup>11</sup>

However, current PRS for a variety of chronic conditions have far greater predictive value among individuals of European ancestry compared to other ancestries as PRS tend to be generated from GWAS of mostly or entirely white European populations.<sup>12,13</sup> Previous approaches have demonstrated poor predictive performance of European-generated PRS in small Asian, Hispanic, and African cohorts as compared to larger European cohorts.<sup>11</sup> The disparities in prediction accuracy for different genetic ancestries can be attributed to the following statistical observations: (1) GWAS will tend to discover genetic variants that are common to the study population, which tend to be white and/or of European genetic ancestry; (2) linkage disequilibrium differentiates marginal effect size for polygenic traits across populations, despite shared causal mechanisms; and (3) the environment, lifestyle factors, and demography differ across populations.<sup>13</sup> Despite these noted shortcomings of the PRS in addressing population structure, some studies have demonstrated suitable predictive power of a PRS generated from a white Europeans when applied to non-European populations for breast cancer,<sup>14,15</sup> prostate cancer,<sup>16,17</sup> pancreatic cancer,<sup>18</sup> and melanoma.<sup>19</sup> Yet, it is evident that PRS may have limited generalizability across ancestral groups.

This may be particularly problematic for outcomes demonstrating great variability in risk across race and/or genetic ancestry, such as CRC, where risk is elevated with black race.

Here, this study aims to assess this genetic generalizability gap for colorectal cancer by assessing the predictive power of a European and east Asian CRC PRS across racial/ethnic groups. A PRS was generated utilizing 79 SNPs that have reached genome-wide significance among 100,204 CRC cases and 154,587 controls of both European and East Asian genetic ancestry. The CRC PRS was then applied to racial subpopulations (White, Black, Asian) of the UK Biobank cohort, consisting of a total of 9,304 CRC cases and 477,902 controls.

**METHODS:** *Study Sample.* The UK Biobank is a large prospective cohort that, from 2006 to 2010, has recruited over 500,000 participants between the ages 40 and 69.<sup>24</sup> Eligible adults living in the United Kingdom were invited to visit one of twenty-two assessment centers for enrollment where, after providing consent, they completed questionnaires and interviews, completed physical and functional measurements, and provided biospecimens (including blood, saliva, and urine). Study follow-up was completed via linkage to available national datasets (i.e., death certificates, electronic medical records). Cancer diagnosis is informed by linkage to national cancer registries from the Research Information Service of the National Health Service and the Information Services of NHS Scotland.

The UK Biobank remains a valuable asset in studying genetic associations with disease prevention, diagnosis, and treatment for a variety of diseases common in middle and later life due to its prospective nature, large cohort size, and extensive collection of both genetic and clinical information. All participants who provided blood samples have been genotyped, and genome-wide SNP data is available for these individuals.<sup>24,25</sup> Eligible UKB participants were those active as of March, 2022 with available imputed genotypes, totaling to 487,206 participants. This includes a



total of 9,304 cases of colorectal cancer and 477,902 controls. Further details of the study sample can be found in Table 1.

Table 1. Characteristics of eligible UKB sample population.

Characteristic	Cases (Incident and Prevalent): 9,304	Controls: 477,902
	<b>N (%) Cases</b>	<b>N (%) Controls</b>
Sex		
Male	5,220 (56.1)	217,736 (45.6)
Female	4,084 (43.9)	260,166 (54.4)
Age of enrollment (years), mean $\pm$ SD	61.0 $\pm$ 6.4	56.5 $\pm$ 8.1
Race/ethnicity		
White	8,988 (96.6)	450,161 (94.2)
Black	82 (0.9)	7,559 (1.6)
Asian	98 (1.1)	10,818 (2.3)
Mixed race	50 (0.5)	2,789 (0.6)
Other ethnic group	45 (0.5)	4,308 (0.9)
Prefer not to answer	32 (0.3)	1,550 (0.3)
Do not know	1 (0.01)	203 (0.04)
Missing	8 (0.09)	514 (0.1)
CRC Documentation		
ICD-10 code from cancer registry	6,685 (71.9)	-
Self-report	787 (8.5)	-
Both	1,832 (19.7)	-

*Outcome.* The outcome of interest was invasive colorectal cancer diagnosis at or after baseline assessment. The UK Biobank receives cancer diagnoses via linkage to cancer registries. From this, cases are categorized by ICD-10 code, with the relevant CRC codes being C18-C20. UK Biobank also includes self-reported cases of colorectal cancer from questionnaire or verbal interview with study staff. A total of 8,517 cases were retrieved from linkage with 1,832 of these also being reported via verbal interview and/or questionnaire. An additional 787 cases were retrieved from only self-report. Both prevalent ( $n = 2,293$ ) and incident cases ( $n = 7,011$ ) of colorectal cancer were included as those SNPs included in the PRS are germline variants. Controls are defined as any UKB participant without a self-reported or linked CRC diagnosis and available genotype data.

*Generation of PRS.* First, genome-wide significant SNPs were selected using fixed-effects meta-analyzed CRC GWAS data in both European (73%) and east Asian (27%) ancestry, comprising 100,204 cases and 154,587 controls.<sup>27</sup> From this analysis, a total of 205 SNPs was associated with CRC, including 50 novel variants, after adjustment by principal components (PC) to account for ancestral differences across individual GWAS. It should be noted that this GWAS data included some UK Biobank participants, among other European and east Asian cohorts. Of the 205 previously identified SNPs taken forward to the PRS, 79 were available among the imputed genotypes of the UKB participants. Resultantly, the PRS was calculated for each of the 502,413 UKB participants with 79 SNPs. PRS development was performed in PLINK v.1.90b. The PRS is calculated as the sum of risk alleles at each locus, weighted by the risk allele effect, and divided by the total number of SNPs used (N):

$$PRS = \frac{\sum_i^N \beta_i X_i}{N}$$

**Equation 1:** where  $X_i$  is the number of risk alleles observed within the individual sample and  $\beta_i$  is the log odds ratio estimate of the variant association with colorectal cancer from meta-analysis

*Statistical Analysis:* Mean PRS among cases and controls for each racial/ethnic UKB subpopulation were compared using Welch's t-test. Logistic regression analysis for which the observed colorectal cancer status was the dependent variable and the individual polygenic risk score was the independent variable was performed. Models were adjusted for both sex and age. From these models, discriminatory power of the PRS with regards to CRC case-control status was determined using the area under the receiver operating characteristics curve (AUC). All statistical analyses were performed using R 4.2.2.

**RESULTS:** Individual-level genetic risk from 5,028 variants was determined for 9,304 colorectal cancer cases and 477,902 controls in the UKB sample population by use of imputed genotype data. From visual examination of the distribution of scores between cases and controls, cases

consistently demonstrated higher PRS values (Figure S1). Overall, PRS values were greater in white individuals as compared to Asian and black individuals (Figure S1, Table 2). In all racial/ethnic subpopulations, there was a statistically significant difference in mean polygenic risk score between cases and controls as determined by Welch's t-test (Table 2).

Table 2. Polygenic risk score (PRS) summary statistics.

<b>UKB Population</b>	<b>Cases (n)</b>	<b>Controls (n)</b>	<b>Case Mean PRS</b>	<b>Control Mean PRS</b>	<b>p<sup>1</sup></b>
White	8,988	450,161	0.00253 ± 0.00158	0.00158 ± 0.00260	<2.2e-16
Black	82	7,559	0.00165 ± 0.00197	0.00120 ± 0.00232	0.042
Asian	98	10,818	0.00184 ± 0.00257	0.000921 ± 0.00262	6.2e-4

<sup>1</sup>Welch's t-test

As constructed, the PRS demonstrated the most significant discriminatory power among self-identified white individuals ( $p < 2.2e-16$ ), and the least significant discriminatory power among the self-identified black individuals ( $p = 0.042$ ). Among the Asian UKB subpopulation, there was moderately significant discriminatory power between CRC case-control status ( $p = 6.02e-04$ ). Further validation by ROC curve indicated acceptable discrimination for all racial/ethnic subpopulations (Table 3). Again, the greatest predictive power was seen among white individuals (AUC = 0.699), followed by Asian (AUC = 0.698) and black individuals (AUC = 0.662).

Table 3. PRS prediction of colorectal cancer in UKB racial/ethnic subpopulations.

<b>UKB Population</b>	<b>Cases (n)</b>	<b>Controls (n)</b>	<b>AUC</b>
White	8,988	450,161	0.699
Black	82	7,559	0.662
Asian	98	10,818	0.698

**DISCUSSION:** It is essential that genetic tools be developed to better identify individuals at high risk of colorectal cancer to not only enhance screening recommendations, but also precisely target preventative and therapeutic interventions for patients. Equally important is the need to distinguish those individuals at low risk for CRC to prevent unnecessary screening due to costs and possible complications. Due to the considerable heritability of CRC and its polygenic nature, the polygenic risk score is a promising method to evaluate genetic colorectal cancer risk.

Presented here is a PRS comprised of 79 SNPs for colorectal cancer from a cohort of European and east Asian genetic ancestry, and its application 9,304 CRC cases and 477,902 controls from UK Biobank. The PRS demonstrates predictive power ( $AUC = 0.699$ ) similar to previous PRS. This PRS demonstrates acceptable discriminatory accuracy with regards to colorectal cancer among white, black, and Asian target subpopulations. As the reference population from which the PRS was constructed is comprised of a majority of individuals of European ancestry, it is to be expected that the PRS would perform best among a white subpopulation. For similar reasons, the PRS was expected to accurately predict CRC status among Asian participants, too. Compared to performance among white participants, the reduced predictive power may be attributed to both the smaller number of ancestral east Asians in the reference population, and the issue that some risk variants included in the PRS are monomorphic in east Asian participants.<sup>27</sup> Additionally, the reduced frequency of CRC among non-white subpopulations will diminish the power of the PRS. The relatively poor predictive power of this PRS among self-identified black participants, relative to other subpopulations, indicates that some population-specific risk-SNPs may be missing. Those 79 SNPs that were included are sufficient to detect case-control status among this subpopulation, but are not comprehensive of ancestry-specific risk loci. Thus, it is evident that some genetic generalizability across ancestral groups may

exist for colorectal cancer; however, further analysis is necessary to determine the dynamics of CRC genetic etiology across genetic ancestry, particularly for non-European ancestry.

*Study Limitations.* This study has several limitations in both its conception and its analysis. The first among these is the choice of reference population and the study population. The meta-analysis of GWAS from which summary statistics were used to generate the PRS consisted of majority European ancestry subjects. A sizable minority of participants were of East Asian ancestry. As previously noted, PRS have historically underperformed in populations of genetic ancestry different from which the PRS was generated. This PRS is further evidence of this phenomenon as it manifests for colorectal cancer, as demonstrated by the PRS's reduced predictive power among black participants, especially. Additionally, the small number of Asian and black participants in the sample population, as compared to the number of white participants, dramatically reduced both power and precision. Inclusion of a greater number of non-white participants may improve estimates of the predictive power of a polygenic risk score, but it also may better represent the genetic profiles of CRC among non-white individuals.

Another limitation of this analysis is its nondiscriminatory outcome definition. All cases of colorectal cancer – incident or prevalent, and regardless of age of diagnosis or familial history of disease – were included in the analysis to maximize power. Future exploration of colorectal cancer risk, genetic and otherwise, in the UK Biobank cohort should consider the role of both familial cancer syndromes and early-onset colorectal cancer cases. Removal of CRC cases with familial history may improve predictive power as implicated loci are not included in this PRS. Despite overall decreases in the incidence of CRC in the United States, there has been a discouraging increase in the incidence of the disease among those aged 50 years or younger.<sup>30</sup>

Approximately half of all early-onset cases can be linked to either familial CRC or other hereditary cancer syndromes. Considering this heightened heritability, application of the PRS to early-onset cases and age-matched controls may illuminate novel genetic risks of this heterogeneous cancer.

Some limitations exist within the UK Biobank dataset. In generating the PRS, approximately 120 of the 205 genome-wide significant SNPs were lost. Due to the costly and time-intensive nature of whole-genome sequencing, imputation is a useful method to infer unobserved genotypes. Yet, the imputation method utilized for UKB participants resulted in imputed genotypes that did not include all SNPs of interest. In this case, imputation of the individual UKB participant genomes could be redone to increase genomic resolution, with particular attention paid to the loci of interest for this PRS.

It is also important to note that, due to the availability of racial/ethnic and genetic ancestry data within the UK Biobank dataset, this study is forced to conflate race with genetic ancestry. While race and ethnicity are social, and generally man-made concepts, genetic ancestry refers to the genetic architecture and genetic variation that is shared among a population. Though these concepts may align (e.g., self-identified “whiteness” and European ancestry), this is not always the case. Genetic ancestry is a measurable biological parameter that is better suited for conceptions of genetic risk of disease, particularly because race is not biologic in nature. The UK Biobank dataset has determined European genetic ancestry for many participants ( $n = 409,522$ ); however, it has not determined ancestry via principal components analysis of genotypes for any other ancestral group. With the current presentation of the dataset, it is not possible to fully assess the genetic generalizability gap of disease across ancestries; therefore, race/ethnicity is used as an unsatisfactory proxy.

*Future Directions.* From this analysis, it is apparent that the genetic contributions to colorectal cancer risk among non-European populations warrants further investigation. Evaluation of the existing PRS in a multi-ethnic cohort with an acceptable number of non-European (e.g., Asian and African) could further elucidate whether genetic risk for CRC in different ancestral groups can be predicted by the same European-generated polygenic risk score. As suggested by this study's preliminary results, it may be necessary to utilize multi-ethnic GWAS data to develop a generalizable risk score. One such dataset that could address this gap is the All of Us cohort. The All of Us Research Program seeks to recruit participants of demographic characteristics that have been historically underrepresented in biomedical and epidemiological research.<sup>31</sup> This program includes more than 300,000 Americans with available genetic data on approximately 165,000 participants, making it a promising dataset for further investigation into the genetic factors that confer disease risk in minority populations.

Furthermore, this PRS employs the simplest of approaches to generation of polygenic risk scores: utilization of an established list of genome-wide significant SNPs as identified by GWAS. However, SNPs could alternatively be selected from broader GWAS data based on LD clumping. This method is useful in minimizing the number of highly correlated SNPs entered into the PRS, but has demonstrated only minimally improved predictive power as compared to selection of only known GWAS variants.<sup>11</sup> Another LD-clumping approach is LDpred.<sup>29</sup> LDpred is a Bayesian genetic risk prediction model that accounts for familial polygenic components to risk that may actually have diminished predictive power at the individual level. Assuming that all SNPs are causal, this method infers the posterior mean effect size of each variant by using a prior on effect sizes and linkage disequilibrium from an external reference panel. For colorectal cancer risk specifically, LDpred-derived PRS have considerably greater predictive power over other

clumping-based approaches.<sup>11</sup> Creation of an LDpred-derived PRS will likely improve upon the genetic risk prediction presented here.

Finally, to further explore the utility of polygenic risk score for the purposes of personalized medicine and development of screening recommendations for colorectal cancer, future efforts should be made to optimize the number of variants included in CRC PRS. In considering current guidelines, genetic testing for CRC is only recommended for those with known or suspected familial cancer syndromes.<sup>32</sup> To prepare for future implementation, a CRC PRS should be optimized. That is, investigation into the fewest number of SNPs necessary to accurately discriminate colorectal cancer, while also maximizing the positive predictive value of the PRS, is necessary. Such a PRS could readily be incorporated into current CRC risk prediction and screening practices in a manner that is both efficient and cost-effective for individual patients and at the population level.

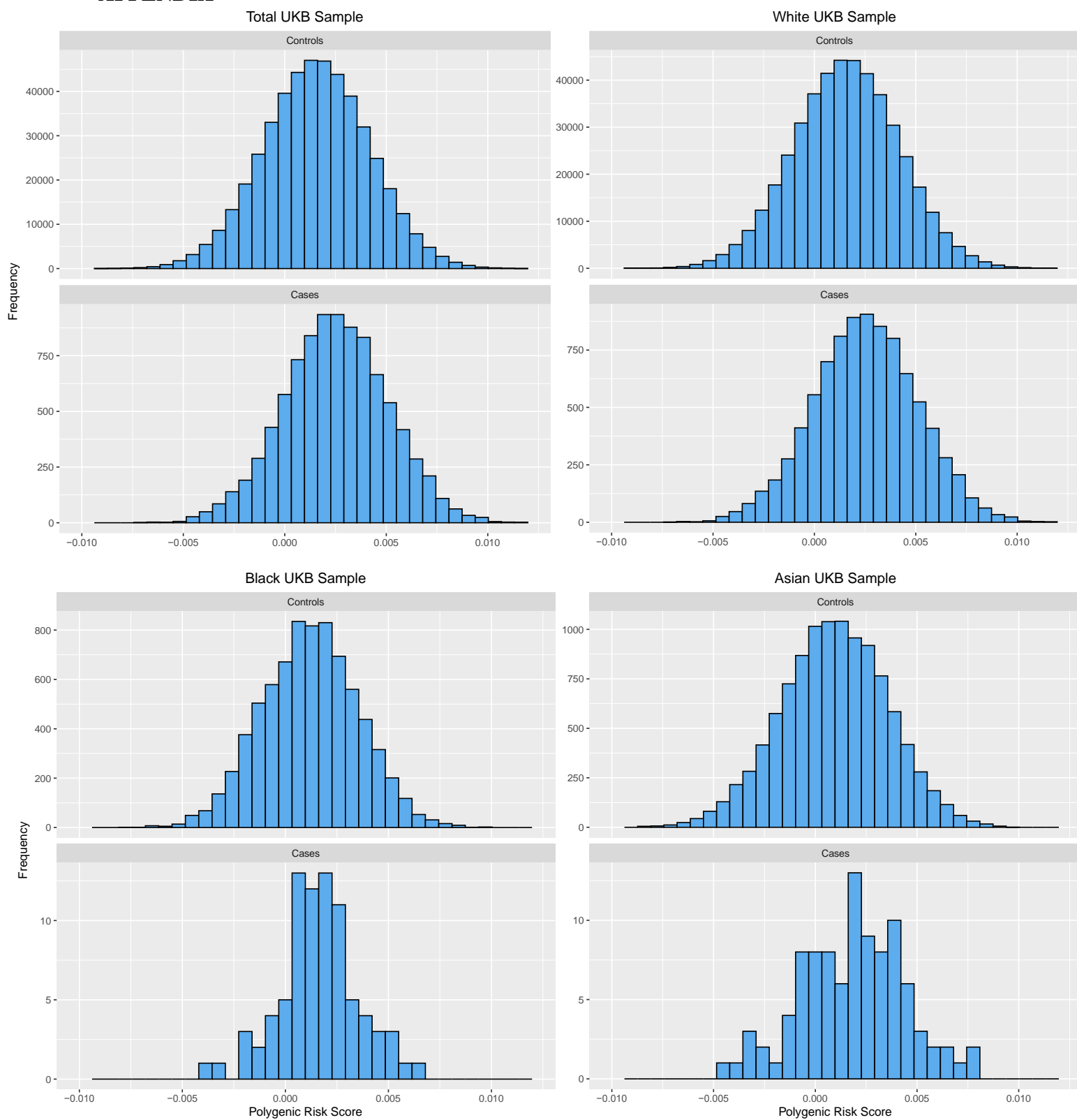


## REFERENCES

1. Ohri A, Robinson A, Liu B, Bhuket T, Wong R. Updated Assessment of Colorectal Cancer Incidence in the U.S. By Age, Sex, and Race/Ethnicity. *Digestive Diseases and Sciences*. 2020;65:1838-1849.
2. Brenner H, Kloor M, Pox CP. Colorectal cancer. *The Lancet*. 2014;383(9927):1490-1502.
3. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat Rev Gastroenterol Hepatol*. 2019 Dec;16(12):713-732.
4. World Cancer Research Fund International/American Institute for Cancer Research. Continuous update project report: Diet, nutrition, physical activity, and colorectal cancer. *American Institute for Cancer Research* <https://www.wcrf.org/wp-content/uploads/2021/02/Colorectal-cancer-report.pdf> (2018).
5. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*. 2000 Jul 13;343(2):78-85.
6. Jasperson, KW, Tuohy, TM, Neklason DW, Burt, RW. Hereditary and familial colon cancer. *Gastroenterology*. 2010;138:2044-2058.
7. Jiao S, Peters U, Berndt S, Brenner H, Butterbach K, Caan BJ, Carlson CS, Chan AT, Chang-Claude J, Chanock S, Curtis KR, Duggan D, Gong J, Harrison TA, Hayes RB, Henderson BE, Hoffmeister M, et al. Estimating the heritability of colorectal cancer. *Hum Mol Genet*. 2014 Jul 15;23(14):3898-3905.
8. Schmit SL, Edlund CK, Schumacher FR, Gong J, Harrison TA, Huyghe JR, Qu C, Melas M, Van Den Berg DJ, Wang H, Tring S, Plummer SJ, Albanes D, Alonso MH, Amos CI, Anton K, Aragaki AK, et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *J Natl Cancer Inst*. 2019 Feb 1;111(2):146-157.
9. Howlader N, Noone AM, Krapcho M, Miller D. National Cancer Institute; Bethesda, MD: 2019. SEER Cancer Statistics Review, 1975-2016. [https://seer.cancer.gov/archive/csr/1975\\_2016/](https://seer.cancer.gov/archive/csr/1975_2016/).
10. Bénard F, Barkun AN, Martel M, von Renteln D. Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. *World J Gastroenterol*. 2018;24(1):124-138.
11. Thomas M, Sakoda LC, Hoffmeister M, Rosenthal EA, Lee JK, van Duijnhoven FJB, Platz EA, Wu AH, Dampier CH, de la Chapelle A, Wolk A, Joshi AD, Burnett-Hartman A, Gsur A, Lindblom A, Castells A, Win AK, Namjou B, Van Guelpen B, Tangen CM, He Q, Li CI, Schafmayer C, Joshi CE, et al. Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. *Am J Hum Genet*. 2020 Sep 3;107(3):432-444.
12. Osterman MD, Kinzy TG, Cooke Bailey JN. Polygenic Risk Scores. *Current Protocols*. 2021;1.
13. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51:584-591.
14. Shieh Y, Fejerman L, Lott PC, Marker K, Sawyer SD, Hu D, Huntsman S, Torres J, Echeverry M, Bohórquez ME, Martínez-Chéquer JC, Polanco-Echeverry G, Estrada-Flórez AP; et al. A Polygenic Risk Score for Breast Cancer in US Latinas and Latin American Women. *J Natl Cancer Inst*. 2020 Jun 1;112(6):590-598.
15. Ho WK, Tan MM, Mavaddat N, Tai MC, Mariapun S, Li J, Ho PJ, Dennis J, Tyrer JP, Bolla MK, Michailidou K, Wang Q, Kang D, Choi JY, Jamaris S, Shu XO, Yoon SY, Park SK, Kim SW, et al. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nat Commun*. 2020 Jul 31;11(1):3833.
16. Du Z, Lubmawa A, Gundell S, Wan P, Nalukenge C, Muwanga P, Lutalo M, Nansereko D, Ndaruhutse O, Katuku M, Nassanga R; African Ancestry Prostate Cancer Consortium (AAPC); Asiimwe F, et al. Genetic risk of prostate cancer in Ugandan men. *Prostate*. 2018 Apr;78(5):370-376.
17. Du Z, Hopp H, Ingles SA, Huff C, Sheng X, Weaver B, Stern M, Hoffmann TJ, John EM, Van Den Eeden SK, Strom S, Leach RJ, Thompson IM Jr, Witte JS, Conti DV, Haiman CA. A genome-wide association study of prostate cancer in Latinos. *Int J Cancer*. 2020 Apr 1;146(7):1819-1826.
18. Nakatochi M, Lin Y, Ito H, Hara K, Kinoshita F, Kobayashi Y, Ishii H, Ozaka M, Sasaki T, Sasahira N, Morimoto M, Kobayashi S, Ueno M, Ohkawa S, Egawa N, Kuruma S, Mori M, Nakao H, et al. Prediction model for pancreatic cancer risk in the general Japanese population. *PLoS One*. 2018 Sep 7;13(9):e0203386.

19. Cust AE, Drummond M, Kanetsky PA; Australian Melanoma Family Study Investigators; Leeds Case-Control Study Investigators; Goldstein AM, Barrett JH, MacGregor S, Law MH, Iles MM, Bui M, et al. Assessing the Incremental Contribution of Common Genomic Variants to Melanoma Risk Prediction in Two Population-Based Studies. *J Invest Dermatol.* 2018 Dec;138(12):2617-2624.
20. Jeon J, Du M, Schoen RE, Hoffmeister M, Newcomb PA, Berndt SI, Caan B, Campbell PT, Chan AT, Chang-Claude J, Giles GG, Gong J, Harrison TA, Huyghe JR, Jacobs EJ, Li L, Lin Y, Le Marchand L, et al. Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. *Gastroenterology.* 2018 Jun;154(8):2152-2164.e19.
21. Hsu L, Jeon J, Brenner H, Gruber SB, Schoen RE, Berndt SI, Chan AT, Chang-Claude J, Du M, Gong J, Harrison TA, Hayes RB, Hoffmeister M, Hutter CM, Lin Y, Nishihara R, Ogino S, Prentice RL, et al. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology.* 2015 Jun;148(7):1330-9.e14.
22. Smith T, Gunter MJ, Tzoulaki I, Muller DC. The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. *Br J Cancer.* 2018 Oct;119(8):1036-1039.
23. Dunlop MG, Tenesa A, Farrington SM, Ballereau S, Brewster DH, Koessler T, Pharoah P, Schafmayer C, Hampe J, Völzke H, Chang-Claude J, Hoffmeister M, Brenner H, von Holst S, Picelli S, Lindblom A, et al. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals. *Gut.* 2013 Jun;62(6):871-81.
24. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779.
25. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203–209.
26. UK Biobank. *Cancer data: linkage from national cancer registries* from <https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/CancerLinkage.pdf>. 2013;1.4.
27. Fernandez-Rozadilla, C, Timofeeva, M, Chen, Z. *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).
28. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 2013 Mar;9(3):e1003348.
29. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, Genovese G, Loh PR, Bhatia G, Do R, Schizophrenia Working Group of the Psychiatric Genomics Consortium. Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 2015;97:576–592.
30. Mauri G, Sartore-Bianchi A, Russo AG, Marsoni S, Bardelli A, Siena S. Early-onset colorectal cancer in young individuals. *Mol Oncol.* 2019 Feb;13(2):109-131.
31. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The "All of Us" Research Program. *N Engl J Med.* 2019;381(7):668-676.
32. American Cancer Society. (2020, June 29). *Colorectal cancer genetic testing: Lynch syndrome testing.* Colorectal Cancer Genetic Testing | Lynch Syndrome Testing. Retrieved April 22, 2023, from <https://www.cancer.org/cancer/colon-rectal-cancer/causes-risks-prevention/genetic-tests-screening-prevention.html>.

## APPENDIX



**Figure S1.** Distribution of polygenic risk scores among case and control samples among the total UKB sample and among racial/ethnic subpopulations.