Race Corrections in Clinical Models: Examining Family History and Cancer Risk

Anna Zink, PhD*; Ziad Obermeyer, MD**; Emma Pierson, PhD***

August 11, 2023

Despite ethical and historical arguments for removing race corrections from clinical algorithms, the consequences of removal remain unclear. An important consideration in this debate is the varying quality of medical data across race groups. For example, family history of cancer is an essential predictor in cancer risk prediction algorithms but is less reliably documented for Black patients and may therefore be less predictive of cancer outcomes. In this study, we evaluated the impact of race correction when reported family history of colorectal cancer was less predictive of colorectal cancer risk for Black participants compared to White participants. We found that race correction correctly captured the differential value of reported family history, improved model fit, and increased the fraction of Black participants flagged as high risk, potentially improving racial equity in screening. Our study illustrates a much broader point: missing and erroneous data is ubiquitous in medicine, at rates which vary by race group, and race correction can help address this problem.

*University of Chicago
**University of California, Berkeley
***Cornell University

The medical community is locked in a consequential debate over when clinical algorithms should use *race corrections*, which include patient race as an input to the algorithm. This debate has immediate implications for patient care across domains including cardiology, pulmonology, nephrology, urology, obstetrics, oncology, and endocrinology, which have all historically relied on clinical algorithms containing race corrections.[1] For example, some health systems have recently removed race as an input when estimating glomerular filtration rate[2], a measure of kidney function which guides diagnosis of kidney disease, drug recommendations, and kidney transplants.[3] The reconsideration of race corrections is vital and long overdue: some race corrections rely on dubious data, exacerbate health disparities, or flow from false beliefs about race as a biological variable, when it is in fact a social construct.[1] At the same time, the consequences of removing race corrections remain unclear, and researchers have called for caution and careful study before doing so.[4,5]

Here, we study an important consideration in this debate: varying data quality across race groups. Differences in medical data quality by race group, and their consequences for health equity, occur frequently and have been documented in diverse domains.[6] As an example, consider family health history data. Family history of cancer is a known risk factor for breast, colorectal, ovarian and uterine cancer.[7] But *recorded* family health history information, often collected during medical interviews with a clinician, varies across race groups. A number of studies have found that family history of cancer is better known and documented in White patients.[8–11] These racial disparities in data quality mean that the predictive value of recorded family history could vary across race groups. In particular, the absence of *recorded* family history may be less reassuring in non-White patients, who may be incorrectly recorded as having no family history

either because the clinician does not ask, or the patient does not know. A race-blind risk prediction would fail to account for this, producing inappropriately low predicted risks for non-White patients without recorded family history; in contrast, a race correction could capture how the prognostic value of recorded family history varies by race.

We tested this hypothesis on a colorectal cancer risk score, where family history of colorectal cancer is an important risk factor and typically results in earlier and/or more frequent tests.[12] Using data from the Southern Community Cohort Study (SCCS)[13], established to study cancer disparities, we analyzed 77,836 adults (ages 40-74) with no history of colorectal cancer at baseline. Self-reported family health history was collected for all first-degree relatives of participants. Diagnosis of colorectal cancer was identified using follow-up survey, state cancer registry, and national death index data. See Table S1 for a summary of the sample.

First, we compared the prognostic value of self-reported family history for 10-year colorectal cancer risk for self-reported Black versus White participants (Figure 1). Running a logistic regression predicting 10-year colorectal cancer incidence, we found that family history was strongly predictive of cancer risk for white participants (OR: 1.74, 95% CI: 1.25-2.38, p-value 0.001), whereas for Black participants, it was not predictive (OR: 0.98, 95% CI: 0.72-1.29, p-value 0.887). In other words, the predictive value of reported family history varied across race groups, consistent with what would be expected if family history were less reliably reported in Black patients.

Then, we created two screening algorithms that modeled 10-year colorectal cancer risk as a function of age, sex, family history, screening history, and lifestyle habits based on the set of controls used in the NIH Colorectal Cancer tool.[14] The baseline algorithm was race-blind (i.e., did not include race as a predictive feature), whereas the race-corrected algorithm added an indicator for whether the participant was Black both as a main effect and as an interaction with family history.

Both race terms were statistically significant: the main effect (p-value: <0.001) captured the fact that, among participants who did not report known family history, Black participants had 1.38x higher odds than White participants of developing colorectal cancer. The interaction term between race and family history was also statistically significant (p-value: 0.010), indicating a differential predictive value of self-reported family history across race groups. These results were consistent across alternative model specifications, outcome definitions, and family history definitions (Tables S2-S5).

We performed a likelihood ratio test to assess whether the goodness of fit significantly improved using race correction and found that it did (p-value <0.001). Consistent with this, the race-corrected model achieved a higher Area Under the Receiving Operating Characteristic (AUROC) than the race-blind algorithm in the overall cohort (0.613 versus 0.606) and in individual racial groups (Table S6), though the increase in AUROC was not statistically significant (p-value: 0.149). (The fact that the increase in AUROC was not statistically significant is not surprising: we also observed a statistically insignificant increase in AUROC even when adding well-known

risk factors like smoking history[15], and this effect has also been documented in other risk prediction settings.[16])

We then compared how the race-blind and race-corrected algorithms assigned participants to colorectal cancer risk categories. The race-corrected algorithm included a larger share of Black participants among the predicted high-risk group (Figure 2). For example, in the race-corrected algorithm, 74.4% of participants flagged in the top 50% of predicted risk were Black compared to 66.1% in the race blind algorithm (p-value: <0.001). Similar results held across all high-risk cutoffs (e.g., top quartile, top decile, and top percentile).

Identifying individuals at high risk for colorectal cancer is an important component of prevention and screening practices in the U.S., where colorectal cancer remains the third leading cause of cancer-related death.[17] In 2021, the United States Preventive Services Taskforce changed the recommended age for colorectal screening from 50 to 45 in the hopes of increasing screening rates to counteract early-onset cancer, among Black men in particular[18,19], consistent with findings that colorectal cancer risk varies by race.[5] Removing race from colorectal screening predictors could reduce the number of Black patients recommended for screening, working against efforts to reduce disparities in colorectal cancer screening and outcomes.

Overall, our analysis shows that including race in a colorectal cancer screening algorithm improved model fit, captured the differential predictive value of family history across groups, and increased the fraction of Black participants flagged as high risk, potentially increasing their

access to screening. More broadly, our case study illustrates how race corrections can help allow models to capture varying data quality across race groups, a frequent phenomenon in health data.

This argument does not rely on treating race as a biological variable, which it is not. Rather, our argument is about the consequences of structural racism on the quality of medical data. The deep structural injustices which pervade healthcare and public health mean that critical inputs to medical algorithms -- like family history -- are more likely to be missing or misrecorded for some race groups. We do a disservice to patients if we fail to acknowledge this reality in our algorithms. But acknowledging that reality is not the same as accepting it. While race corrections can allow algorithms to capture current deficiencies in medical data, we need to look beyond that to what medical data could be, and must be, if we are to achieve health equity for all patients.

Methods

Data

Our data come from the Southern Community Cohort Study (SCCS) established in 2001 to study cancer disparities as well as other health conditions in the southeastern U.S.[20] SCCS enrollment began in 2002 and continued for eight years (until 2009). Participants were primarily recruited from community health centers in the following twelve states: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia. Data were collected from surveys administered at the time of enrollment and several follow-up periods. Data from the baseline survey were collected either through a self-administered survey or an in-person computer-assisted interview. Follow-up surveys were done

by telephone or self-administered: approximately 68% of participants completed the follow-up surveys. State cancer registry data were linked to participants when possible. Please refer to Table S1 for more information on the sample.

The primary outcome was whether the participant developed colorectal cancer in the ten years following enrollment. This variable was measured using the follow-up survey, cancer registry data and National Death Index reports of malignant neoplasms of the colon, rectum, and anus. We included all recorded cases of colorectal cancer from any of these three sources.

Family history of cancer was collected for participants' birth mother, birth father, full sisters, and full brothers. For each family member, respondents could select "yes", "no", or "don't know" for whether the person had cancer. Respondents who indicated that any of these family members had cancer then selected which type of cancer they had. We defined a participant as having a known family history of colorectal cancer if they indicated that one of their family members had colorectal cancer, consistent with previous work.[21] For the main analysis, we compared participants with a known family history of cancer to participants who did not have a known family history of cancer (grouping the "don't know" and "no family history" respondents together in the latter category). In a sensitivity analysis, we considered the effects of two alternate ways of coding family history: 1) analyzing family history as a 3-level categorical variable with "don't know" as a separate category and 2) grouping the "don't know" group with the "yes" group as opposed to with the "no" group.

All covariates were measured using data collected in the baseline survey. We defined race groups based on the participants' description of their race or ethnicity at baseline. Participants had six options to choose from (White, Black/African-American, Hispanic/Latino, Asian or Pacific Islander, American Indian or Alaska Native, and Other racial or ethnic group) and could mark all that apply. We defined Black participants as any participants who described themselves as Black/African-American. We defined White participants as any participants who described themselves as White only. More than 95% of the sample identified as either Black or White only, so we only included participants in these two groups for the analysis, following previous work.[22]

Analysis

We examined the prognostic value of family history by race group by plotting 10-year colorectal cancer rates by age, since age is an important risk factor for colorectal cancer that affects screening recommendations.[23] We also ran separate logistic regressions for Black versus White participants in which we predicted 10-year colorectal cancer incidence given age and family history and reported the odds ratio on the family history coefficient for each regression with 95% confidence intervals estimated using profile likelihood methods (see Table S2).

Next, we compared a race-corrected algorithm to a race-blind algorithm. The race-corrected algorithm included a race main effect term and an interaction between race and family history, in addition to the set of controls used by the NIH risk tool[14]: participant age at the time of enrollment, an indicator for female, BMI greater than 30, ever had a sigmoidoscopy, ever had a colonoscopy, ever had polyps, the age that the polyp was identified if ever, smoking status (current, former, never), drinking status (<=1 drink per day, >1 drink per day), whether they took

NSAIDs or Aspirin regularly, whether they did any vigorous activity, and whether they ate vegetables each day. The race-blind algorithm included the same set of NIH controls, but omitted the race main effect term and the interaction between race and family history.

To assess the impact and predictive performance of the two algorithms, data were randomly split into a training set (50% of the data) and a holdout test set (50% of the data). We fit logistic regression models on the training data and assessed the results on the holdout test data. To compute uncertainty on our estimates, we ran 5,000 iterations in which we reshuffled the dataset: at each iteration, we split the test/train data into a new random split, fit the clinical algorithm on the new train data, and predicted on the new test data. We used the distribution of performance measures across iterations to calculate confidence intervals and p-values.

To compare how the share of predicted high-risk participants who were Black differed between the race-blind and race-corrected algorithms, we defined predicted high-risk participants as those in the top k% percentile of predicted risk (where k = 50, 25, 10, 5, and 1), and looked at the share of Black participants among the predicted high-risk group (Figure 2). Then, we evaluated the predictive performance of the two algorithms, overall and by race group, on the holdout test data by measuring the Area Under the Receiving Operating Characteristic (AUROC), a standard measure of predictive performance (Table S6).[24]

We also performed a set of checks to ensure that our results held under different outcome definitions, model choices, and definitions of family history.

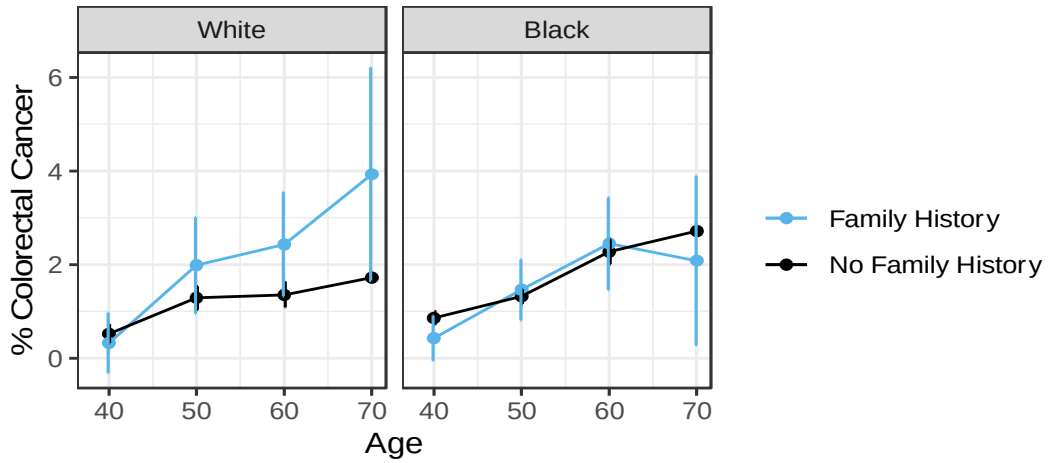
First, we checked that the interaction term between family history and race remained significant under two additional outcome definitions to check that the results were not sensitive to the source of reported data: (1) colorectal cancer cases reported in the follow-up surveys, and (2) colorectal cancer cases found in the state registry data. Most diagnosed cases (approx. 80%) were reported in either the follow-up surveys or the state registry data so we did not separate out national death index cases (Table S2).

Next, we repeated our examination of the relationship between family history and colorectal cancer using a Cox proportional hazards model, a common choice for modeling time to medical events (Table S3). For our analysis, the time to event was the diagnosis year minus the enrollment year for participants with a diagnosis of colorectal cancer and the censoring year minus the enrollment year for those without. The censoring year was the year of death (if applicable) or 2016, whichever occurred first.

Finally, we confirmed that our results were robust to altering the definition of family history. First, rather than grouping the participants with unknown family history with the “no family history” group, we grouped them with the “known family history” group (Table S4). This might help address the mismeasurement of family history for Black participants if many of those with unknown family history did in fact have a family member with colorectal cancer. We also re-ran the analysis with family history as a categorical variable with three different categories: No, Don’t Know, and Yes (Table S5).

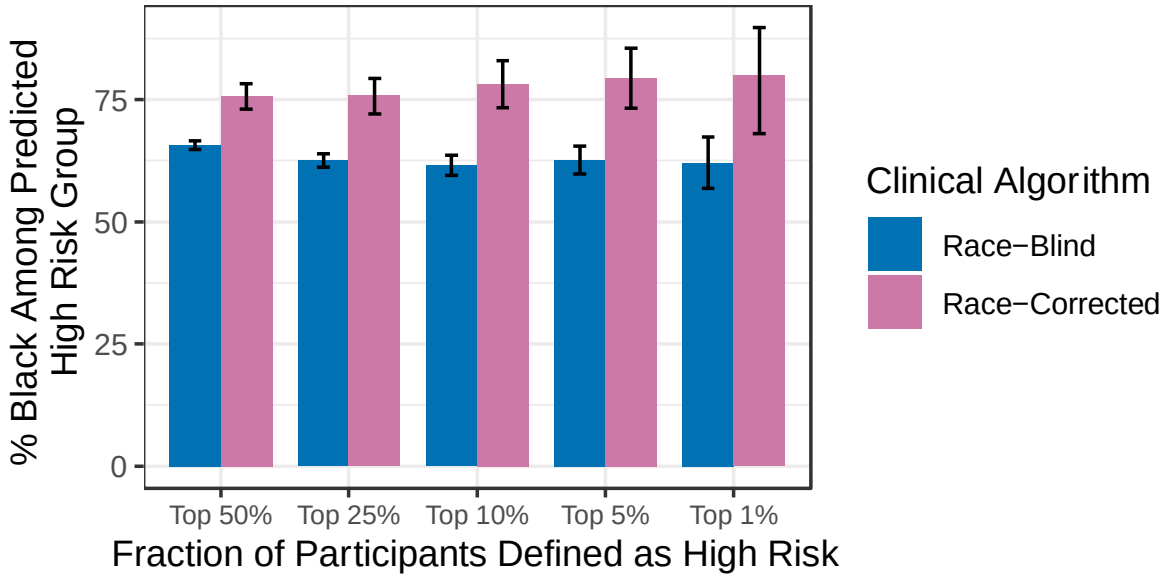
All analyses were run in R version 4.2.1.

Figure 1. 10-Year Colorectal Cancer Rates by Age, Family History, and Race



Note: Family history was predictive of cancer risk for White participants, but not Black participants.

Figure 2. % Black Among Predicted High Risk Group, by High Risk Percentiles



Note: The race-corrected algorithm included more Black participants among the predicted high risk group than the race-blind algorithm.

References

1. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med*. 2020;383(9):874-882. doi:10.1056/NEJMms2004740

2. Removing the Racial Coefficient From Calculations of eGFR. Accessed June 3, 2023. https://reports.mountsinai.org/article/neph2021-07-egfr-and-race

3. Diao JA, Wu GJ, Taylor HA, et al. Clinical Implications of Removing Race From Estimates of Kidney Function. *JAMA*. 2021;325(2):184-186. doi:10.1001/jama.2020.22124

4. Embracing Genetic Diversity to Improve Black Health | NEJM. Accessed June 3, 2023. https://www.nejm.org/doi/full/10.1056/NEJMms2031080

5. Khor S, Haupt EC, Hahn EE, Lyons LJL, Shankaran V, Bansal A. Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors. *JAMA Netw Open*. 2023;6(6):e2318495. doi:10.1001/jamanetworkopen.2023.18495

6. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. *Annu Rev Biomed Data Sci*. 2021;4(1):null. doi:10.1146/annurev-biodatasci-092820-114757

7. Family Health History and Cancer. Centers for Disease Control and Prevention. Published January 9, 2023. Accessed May 6, 2023. https://www.cdc.gov/cancer/family-health-history/index.htm

8. Murff HJ, Byrne D, Haas JS, Puopolo AL, Brennan TA. Race and Family History Assessment for Breast Cancer. *J Gen Intern Med*. 2005;20(1):75-80. doi:10.1111/j.1525-1497.2004.40112.x

9. Kupfer SS, McCaffrey S, Kim KE. Racial and gender disparities in hereditary colorectal cancer risk assessment: the role of family history. *J Cancer Educ Off J Am Assoc Cancer Educ*. 2006;21(1 Suppl):S32-36. doi:10.1207/s15430154jce2101s_7

10. Chavez-Yenter D, Goodman MS, Chen Y, et al. Association of Disparities in Family History and Family Cancer History in the Electronic Health Record With Sex, Race, Hispanic or Latino Ethnicity, and Language Preference in 2 Large US Health Care Systems. *JAMA Netw Open*. 2022;5(10):e2234574. doi:10.1001/jamanetworkopen.2022.34574

11. Andoh JE. The Stories We Don’t Know. *JAMA*. Published online April 13, 2023. doi:10.1001/jama.2023.5891

12. Colorectal Cancer Guideline | How Often to Have Screening Tests. Accessed May 11, 2022. https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/acs-recommendations.html

13. Signorello LB, Hargreaves MK, Blot WJ. The Southern Community Cohort Study: investigating health disparities. *J Health Care Poor Underserved*. 2010;21(1 Suppl):26-37. doi:10.1353/hpu.0.0245

14. Colorectal Cancer Risk Assessment Tool. Colorectal Cancer Risk Assessment Tool. Published 2022. Accessed May 13, 2022. https://www.cancer.gov/ccrisktool

15. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Centers for Disease Control and Prevention (US); 2014. Accessed June 7, 2023. http://www.ncbi.nlm.nih.gov/books/NBK179276/

16. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935. doi:10.1161/CIRCULATIONAHA.106.672402

17. American Cancer Society. Colorectal Cancer Statistics | How Common Is Colorectal Cancer? Accessed May 11, 2022. https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html

18. Mauri G, Sartore-Bianchi A, Russo AG, Marsoni S, Bardelli A, Siena S. Early-onset colorectal cancer in young individuals. *Mol Oncol*. 2019;13(2):109-131. doi:10.1002/1878-0261.12417

19. MPH AC MD. Racial disparities and early-onset colorectal cancer: A call to action. Harvard Health. Published March 17, 2021. Accessed May 11, 2022. https://www.health.harvard.edu/blog/racial-disparities-and-early-onset-colorectal-cancer-a-call-to-action-202103172411

20. Southern Community Cohort Study. Southern Community Cohort Study. Published 2022. Accessed November 1, 2022. https://www.southerncommunitystudy.org/

21. Win AK, MacInnis RJ, Hopper JL, Jenkins MA. Risk Prediction Models for Colorectal Cancer: A Review. *Cancer Epidemiol Biomarkers Prev*. 2012;21(3):398-410. doi:10.1158/1055-9965.EPI-11-0771

22. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

23. Amersi F, Agustin M, Ko CY. Colorectal Cancer: Epidemiology, Risk Factors, and Health Services. *Clin Colon Rectal Surg*. 2005;18(3):133-140. doi:10.1055/s-2005-916274

24. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp J Intern Med*. 2013;4(2):627-635.

25. Abd ElHafeez S, D'Arrigo G, Leonardis D, Fusaro M, Tripepi G, Roumeliotis S. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxid Med Cell Longev*. 2021;2021:e1302811. doi:10.1155/2021/1302811