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

Anna Zink[1], Ziad Obermeyer[2], and Emma Pierson[3]
March 30, 2023

Despite ethical and historical arguments for removing race from algorithms, the consequences of this practice for medical decision-making remain unclear. Eliminating race corrections can have unintended consequences: it can decrease both accuracy and equity, because structural racism distorts the measurement of many variables. We illustrate this using the example of family history of colorectal cancer commonly used in cancer screening. Using data from the Southern Community Cohort Study (SCCS), established to study cancer disparities, we analyze 77,836 adults with no history of colorectal cancer at baseline. First, we compare the prognostic value of self-reported family history for self-reported Black vs. White participants and find that family history is strongly predictive of cancer risk for White, but not Black participants. We then create two screening algorithms that model colorectal cancer risk. The baseline algorithm is race-blind, while the race-corrected algorithm adds Black race both as a main effect and as an interaction family history. The race-corrected algorithm improves upon the race-blind algorithm in a likelihood ratio test (p-value: <0.001). In addition, both race terms are significant: the main effect (p-value: <0.001) captures the fact that, among participants with no family history, Black participants have 1.34x higher odds than White participants of developing colorectal cancer. The interaction term between race and family history is also statistically significant (p-value 0.012), capturing the differential predictive value of family history across race groups. As a result, the race-corrected algorithm includes more Black participants among the predicted high-risk group. Our case 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.

[1]University of Chicago
[2]University of California, Berkeley
[3]Cornell University

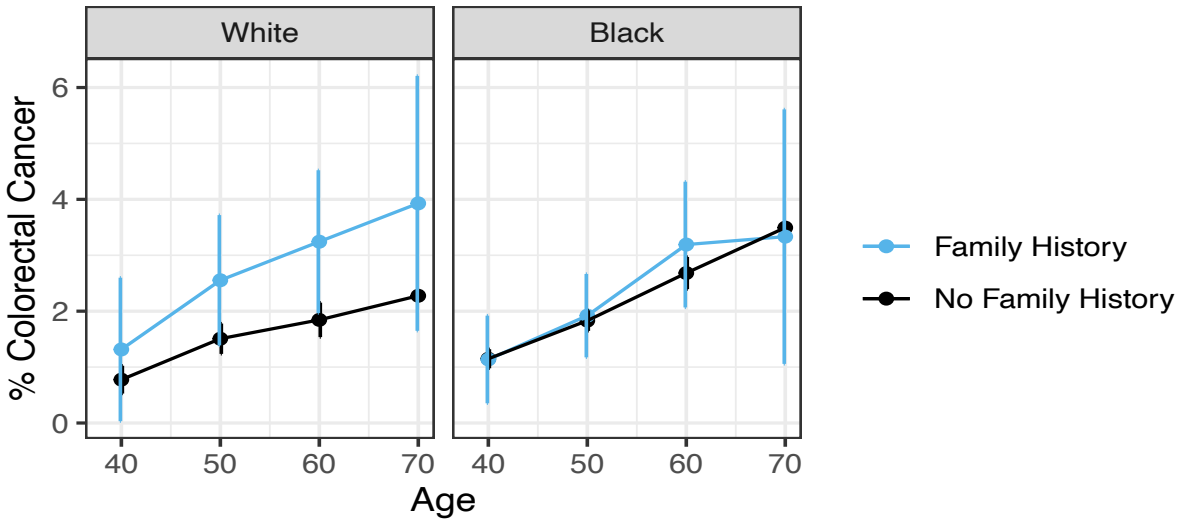
INTRODUCTION

Despite ethical and historical arguments for removing race from algorithms, the consequences of this practice for medical decision-making remain unclear.[1,2] Here we argue that, in some cases, race correction can increase both predictive performance and equity by allowing algorithms to correctly model differences in medical data quality across race groups. Medical data quality issues, and their impacts on equity, have been documented in diverse domains.[3] For concreteness, we focus on a specific case study: family history of colorectal cancer, commonly used in screening recommendations.

The recording of family history of colorectal cancer can be biased because of historic disparities in access to care. The absence of recorded family history, in particular, is less reassuring in Black patients who may be incorrectly recorded as having no family history either because the clinician does not ask or the patient does not know.[4,5] A race-blind risk prediction would fail to account for this, producing inappropriately low predicted risks for Black patients without recorded family history; in contrast, a race correction could capture how the prognostic value of recorded family history varies by race.

Using data from the Southern Community Cohort Study (SCCS)[6], established to study cancer disparities, we analyzed 77,836 adults with no history of colorectal cancer at baseline. We first compared the prognostic value of self-reported family history for self-reported Black vs. White participants (Figure 1). For White participants, family history was strongly predictive of cancer risk (OR: 1.75, 95% CI: 1.30-2.31, p-value <0.001), whereas for Black participants, it was not predictive (OR: 1.09, 95% CI: 0.84-1.38, p-value 0.504). We then created two screening algorithms that modeled colorectal cancer risk as a function of age, sex, family history, screening

Figure 1. % of Participants with Colorectal Cancer by Reported Family History and Race



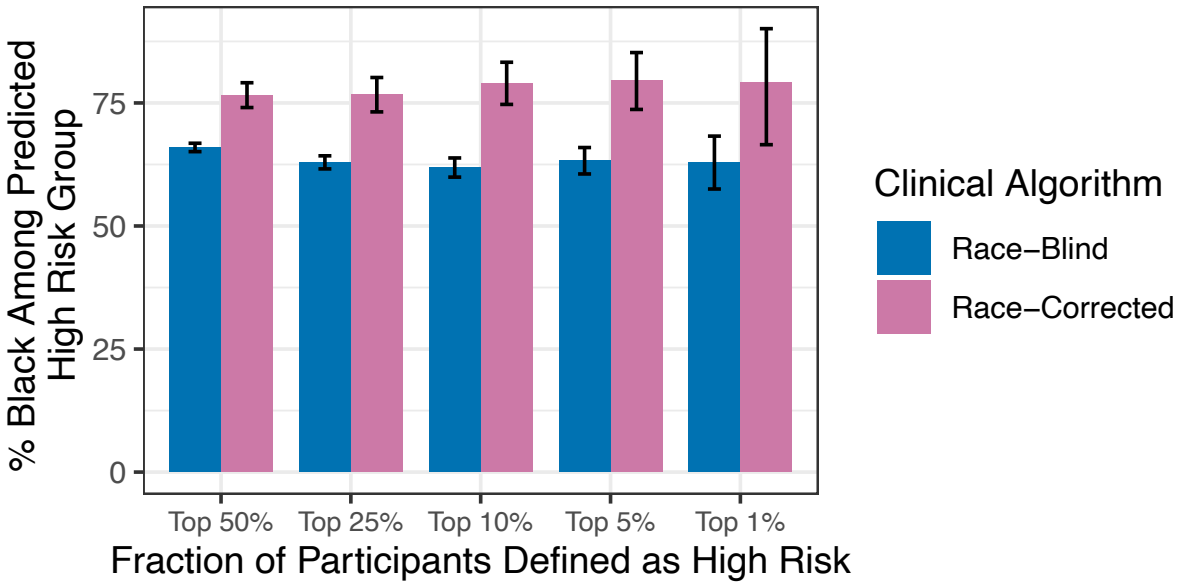
Note: Family history was predictive of cancer risk for White participants, but not Black participants.

history, and lifestyle habits.* The baseline algorithm was race-blind, while 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.

The race-corrected algorithm improved upon the race-blind algorithm in a likelihood ratio test (p-value: <0.001). In addition, 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 1.36x higher odds than White participants of developing colorectal cancer. The interaction term between race and family history was also statistically significant (p-value: 0.012), indicating a differential predictive value of self-reported family history across race groups. The race-corrected algorithm included more Black participants among the predicted high-risk group (Figure 2): 7.3% of participants in the top risk decile were

*We used the same set of controls used by the NIH Colorectal Cancer tool. Please refer to the supplement for the full list of variable definitions.

Figure 2. % Black Among Predicted High Risk Group



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

Black when using the race-corrected algorithm, as opposed to 63.6% when using the race-blind algorithm (p-value: <0.001).

Including race in a screening algorithm increased its predictive performance by allowing the algorithm to correctly capture the fact that family history was less accurately recorded for Black patients, potentially increasing equity by improving access to colorectal cancer screenings. Our case 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. While race correction can address shortcomings in existing medical datasets, it is also imperative to pursue non-algorithmic fixes to these systemic issues, ensuring that data is equitably and accurately collected for all patients.

DATA AND METHODS

Data Source

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.[7] 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 are collected from surveys administered at the time of enrollment and several follow-up periods. Data from the baseline survey is collected either through a self-administered survey or an in-person computer assisted interview. Follow-up surveys are done by telephone or self-administered: approximately 68% of participants completed the follow-up surveys. State cancer registry data is linked to participants when possible.

Study Outcomes and Covariates

The primary outcome is whether the participant developed colorectal cancer. This variable is measured using the follow-up survey, cancer registry data and National Death Index reports of malignant neoplasms of colon, rectum, and anus. In our main analyses, we include all recorded cases of colorectal cancer from any of these three sources.

All covariates are measured using data collected in the baseline survey. We define race groups based on the participants' description of their race or ethnicity at baseline. Participants have 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 can

mark all that apply. We define Black participants as any participants who describe themselves as Black/African-American. We define White participants as any participants who describe themselves as White only. More than 95% of the sample identifies as either Black or White only so we only include participants in these two groups for the analysis.

Family history of cancer is collected for participants' birth mother, birth father, full sisters, and full brothers. For each family member, respondents can select "yes", "no", or "don't know" for whether the person had cancer. Respondents who indicate that any of these family members have had cancer select which type of cancer they had. We define a participant as having a known family history of colorectal cancer if they indicate that one of their family members had colorectal cancer, consistent with previous work.[8] For the main analysis, we compare participants with known family history of cancer to participants who do 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 consider 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.

To fit the colorectal risk prediction algorithms, we use the same set of controls as the NIH Colorectal Cancer tool[9]: participant age at the time of enrollment, and an indicator for female, BMI greater than 30, ever had a sigmoidoscopy, ever had a colonoscopy, ever had polyps, 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 take NSAID or Aspirin regularly, whether they do any vigorous activity, and whether they eat vegetables each day.

Sample

The analysis sample is limited to Black and White participants with no history of colorectal cancer at baseline, consistent with previous work using the SCCS to study colorectal cancer.[10] Our final sample includes 77,836 participants between the ages of 40 and 79. More than two-thirds of participants identify as Black/African-American and the rest as White. Approximately 7% of the sample has a family history of colorectal cancer. A higher proportion of White participants (8.4%) than Black participants (5.9%) report a known family history of cancer. 7.2% of Black participants don't know if they have a family history of colorectal cancer (compared to 4.4% of White participants) corroborating our hypothesis that Black participants are more likely to have imperfectly recorded family history. We also note that colorectal cancer rates are higher in Black participants (2.0%) than in White participants (1.6%) despite lower rates of known family history among Black participants. Please refer to Table S1 for more information on the sample.

Main Analyses

First, we examine the prognostic value of family history by race group. In Figure 1a, we plot colorectal cancer rates by age group and family history status, stratified by race. We plot rates by age because it is an important risk factor for colorectal cancer and affects screening recommendations.[11,12] We run separate logistic regressions for Black versus White participants in which we predict colorectal cancer given age and family history. We report the odds ratio on the family history coefficient for each regression with 95% confidence intervals estimated using profile likelihood methods. The coefficient for family history is not statistically significant for Black participants, but it is significant for White participants (Table S2).

Next, we compare a race-corrected algorithm to a race-blind algorithm. The race-corrected algorithm includes a race main effect term and an interaction between race and family history, in addition to the full set of controls used by the NIH risk tool. The race-blind algorithm only has no access to race as a predictor and only includes the set of NIH controls. We test whether the prognostic value of family history differs significantly across race groups by adding an interaction term between family history and race in a race-corrected algorithm controlling for the full set of NIH controls. We find that the interaction term is statistically significant (p-value: 0.012). See Table S2 for details. We perform a likelihood ratio test to assess whether the goodness of fit is significantly improved using race correction and find that it is (p-value <0.001).

To assess the impact and predictive performance of the two algorithms, data are randomly split into a training set (70% of the data) and a holdout test set (30% of the data). We fit logistic regression models on the training data and assess the results on the holdout test data. To compute uncertainty on our estimates, we run 5,000 iterations in which we reshuffle the dataset: at each iteration we resplit the test/train data, refit the clinical algorithm, and repredict on the new test data. We use the distribution of performance measures across iterations to calculate confidence intervals and p-values.

First, we compare how the share of predicted high-risk participants who are Black differs between the race-blind and race-corrected algorithms. We define predicted high-risk participants as those in the top k% percentile of predicted risk (where k = 50, 25, 10, 5, and 1), and look at the share of Black participants among the predicted high-risk group. The race-corrected algorithm includes a larger share of Black participants among the predicted high-risk group. For example, in the race-corrected algorithm 76.2% of participants flagged in the top 50% of

predicted risk are Black compared to 66.4% in the race blind algorithm. This result holds across all cutoffs for defining high risk.

Then, we evaluate 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.[13] There is a positive but insignificant increase in AUC (0.610 versus 0.609, p-value: 0.241) in the race-corrected versus race-blind algorithm. All analyses are run in R version 4.2.1.

Sensitivity Analyses and Robustness Checks

Our main hypothesis is that the prognostic value of family history differs by race, resulting in a statistically significant interaction term between family history and race. In the main analysis, we find evidence to support this hypothesis. Here, we perform a set of checks to ensure that this result holds under different outcome definitions, model choices, and definition of family history.

First, we check that the interaction term between family history and race remains significant under three additional outcome definitions in the race-corrected algorithm: (1) a censored outcome that only includes colorectal cancer cases within a 10-year period from the start of enrollment, (2) colorectal cancer cases reported in the follow-up survey, and (3) colorectal cancer cases found in the state registry data. The censored outcome checks to make sure our results aren't confounded by different follow-up periods for participants. The latter two outcomes check that the results are not sensitive to the source of outcome data. We find that the interaction between family history and race remains significant across all outcome definitions although it is only significant at the 10% level for cancer registry outcomes (Table S3).

In the main analysis, we predict colorectal cancer outcomes using logistic regression, but other modeling choices could have been used. We therefore repeat 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.[14] For our analysis, the time to event is the diagnosis year minus the enrollment year for participants with a diagnosis of colorectal cancer and censoring year minus enrollment year for those without. The censoring year is the year of death (if applicable) or 2016, whichever occurs first. The results from the Cox model are consistent with the results using logistic regression: in interaction between family history and race remains significant (Table S4).

Finally, we confirm that our results are 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 group them with the "known family history" group. This might help address 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. The interaction between family history and race remains significant under this alternate grouping (Table S5). We also re-run the analysis with family history as a categorical variable with three different categories: No, Don't Know, and Yes. We find that the interaction between known family history and race remains significant (Table S6) indicating that reported family history has different prognostic value for White and Black participants even when coding it as a 3-level variable.

Table S1: Sample Summary

Variable	Black Participants	White Participants	All Participants
Female (%)	58.4	61.1	59.3
Enrollment Age (%)			
40-49	48.6	37.7	45.2
50-59	35.2	36.2	35.5
60-69	13.5	21.9	16.1
70-79	2.7	4.1	3.1
Race (%)			
Black	100.0	0	69.1
White	0	100.0	30.9
Family History of Colorectal Cancer (%)			
Yes	5.9	8.4	6.7
Don't Know	7.2	4.4	6.3
Colorectal Cancer (%)	2.0	1.6	1.9
Mortality (%)	25.3	27.1	25.9
Number of Participants	53,805	24,031	77,836

Table S2. Odds Ratio (95% Confidence Intervals) for Logistic Regression Predicting Colorectal Cancer

Variables	(1) Black Participants	(2) White Participants	(3) Race-Blind Algorithm	(4) Race-Corrected Algorithm
Family History	1.087 (0.844 to 1.377)	1.750*** (1.303 to 2.309)	1.331** (1.101 to 1.595)	1.819*** (1.352 to 2.402)
Family History: Black				0.616* (0.424 to 0.900)
Black				1.357*** (1.196 to 1.543)
Age Controls	Y	Y	Y	Y
Full NIH Controls			Y	Y

P Value: <0.001 '***' <0.01 '**' <0.05 '*' < 0.1 '.'

Table S3. Odds Ratio (95% Confidence Intervals) for Logistic Regression Predicting Colorectal Cancer under Different Outcome Definitions

Variables	(1) Black Participants	(2) White Participants	(3) Race-Blind Algorithm	(4) Race-Corrected Algorithm
Outcome: 10-Year Risk				
Family History	0.887 (0.724 to 1.293)	1.743** (1.246 to 2.383)	1.255* (1.006 to 1.548)	1.802*** (1.285 to 2.467)
Family History: Black				0.564* (0.366 to 0.873)
Black				1.376*** (1.192 to 1.593)
Outcome: Follow-Up Survey Data†				
Family History	0.909 (0.609 to 1.303)	1.877** (1.274 to 2.684)	1.262 . (0.960 to 1.628)	1.913** (1.297 to 2.740)
Family History: Black				0.483** (0.283 to 0.820)
Black				1.323** (1.110 to 1.584)
Outcome: Cancer Registry Data				
Family History	1.157 (0.885 to 1.530)	1.722** (1.198 to 2.407)	1.426** (1.133 to 1.771)	1.886*** (1.310 to 2.641)
Family History: Black				0.652 . (0.415 to 1.030)
Black				1.321*** (1.132 to 1.546)
Age Controls	Y	Y	Y	Y
Full NIH Controls			Y	Y

P Value: <0.001 '***' <0.01 '**' <0.05 '*' < 0.1 '.'

†Excludes 31.7% of the sample that didn't complete follow-up surveys.

Table S4. Hazard Ratios (95% Confidence Intervals) for Cox Proportional Hazard Model Predicting Colorectal Cancer

Variables	(1) Black Participants	(2) White Participants	(3) Race-Blind Algorithm	(4) Race-Corrected Algorithm
Family History	0.995 (0.765 to 1.296)	1.722*** (1.275 to 2.326)	1.254* (1.029 to 1.528)	1.766*** (1.306 to 2.387)
Family History: Black				0.583** (0.391 to 0.870)
Black				1.304*** (1.139 to 1.493)
Age Controls	Y	Y	Y	Y
Full NIH Controls			Y	Y

P Value: <0.001 '***' <0.01 '**' <0.05 '*' < 0.1 '.'

Table S5. Odds Ratio (95% Confidence Intervals) for Logistic Regression Predicting Colorectal Cancer using Alternative Family History Definition

Variables	(1) Black Participants	(2) White Participants	(4) Race-Blind Algorithm	(5) Race-Corrected Algorithm
Family History†	1.164 (0.980 to 1.374)	1.689*** (1.310 to 2.153)	1.310*** (1.137 to 1.503)	1.695*** (1.314 to 2.162)
Family History†: Black				0.689* (0.512 to 0.934)
Black				1.365*** (1.198 to 1.561)
Age Controls	Y	Y	Y	Y
Full NIH Controls			Y	Y

P Value: <0.001 '***' <0.01 '**' <0.05 '*' < 0.1 '.'

†The alternative definition of family history used here groups participants who don't know whether a family member has had colorectal cancer or not with participants who report a known family history.

Table S6. Odds Ratio (95% Confidence Intervals) for Logistic Regression Predicting Colorectal Cancer using 3-level Categorical Family History Definition

Variables	(1) Black Participants	(2) White Participants	(3) Race-Blind Algorithm	(4) Race-Corrected Algorithm
Family History:Don't Know	1.213 . (0.971 to 1.496)	1.485 . (0.951 to 2.212)	1.264* (1.038 to 1.526)	1.407 (0.900 to 2.098)
Family History:Yes	1.106 (0.858 to 1.402)	1.793*** (1.333 to 2.368)	1.355** (1.120 to 1.626)	1.856*** (1.378 to 2.455)
Family History Don't Know:Black				0.848 (0.537 to 1.388)
Family History Yes:Black				0.613* (0.421 to 0.897)
Black				1.364*** (1.197 to 1.560)
Age Controls	Y	Y	Y	Y
Full NIH Controls			Y	Y

P Value: <0.001 '***' <0.01 '**' <0.05 '*' < 0.1 '.'

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. Manski CF. Patient-centered appraisal of race-free clinical risk assessment. *Health Econ*. Published online July 5, 2022. doi:10.1002/hec.4569

3. 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

4. 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

5. 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

6. 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

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

8. 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

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

10. Warren Andersen S, Blot WJ, Lipworth L, Steinwandel M, Murff HJ, Zheng W. Association of Race and Socioeconomic Status With Colorectal Cancer Screening, Colorectal Cancer Risk, and Mortality in Southern US Adults. *JAMA Netw Open*. 2019;2(12):e1917995. doi:10.1001/jamanetworkopen.2019.17995

11. 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

12. Mehta SJ, Morris AM, Kupfer SS. Colorectal Cancer Screening Starting at Age 45 Years—Ensuring Benefits Are Realized by All. *JAMA Netw Open*. 2021;4(5):e2112593. doi:10.1001/jamanetworkopen.2021.12593

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

14. 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