RESEARCH LETTER

Risk Prediction Modeling for Colorectal Adenomas: An Avenue Toward Prevention of Early Onset Colorectal Cancer

In response to a rise in early-onset colorectal cancer (EOCRC), many societies controversially changed the age of screening initiation from 50 to 45 years. A major limitation of this strategy is that the median age of diagnosis for EOCRC is 44 years, so many people would develop EOCRC even before being eligible for screening, although they may harbor colorectal adenomas (CRAs).[1] The optimal approach is to risk-stratify persons and recommend screening accordingly. We derive and internally validate a prediction model for CRAs in persons aged less than 50 years to facilitate EOCRC prevention.

The prediction model was created within a retrospective cohort study of persons between ages 18 and 49 who underwent a colonoscopy at the University of Miami's ambulatory care center between January 1, 2020 and January 1, 2023 (Figure A1). We identified whether persons had any resected CRA at colonoscopy, defined as at least 1 tubular, villous, or tubulovillous adenoma, and collected relevant covariates (seen in Table 1).

Risk prediction modeling (70/30 train/test dataset) included 4 techniques: multivariable logistic regression (backward selection), random forest, gradient boosting, and artificial neural network, with 5-fold cross-validation. Testing included calibration and discrimination by visually inspecting models, plotting receiver operator characteristic curves, and computing area under the curve (AUC).

A priori sample size calculations indicated power to achieve an AUC of 0.60 with a confidence interval (CI) width of 0.14. R (version 4.3.2) was used for statistical analyses. Missingness was not informative; therefore, imputation was not pursued. This study was approved by the Institutional Review Board at the University of Miami.

We identified 1417 individuals who met inclusion criteria, of which a total of 275 (19.4%) had at least 1 CRA (Table 1). Risk prediction modeling demonstrated AUCs of 0.71 (95% CI: 0.65–0.77, logistic

Table 1. Cohort Characteristics

	No adenoma	Adenoma	
Variable	N = 1142	N = 275	*P* value
Age, y			
Median (Q1, Q3)	44.0 (33.0, 47.0)	47.0 (45.0, 48.0)	<.001
Missing	6 (0.5%)	0 (0.0%)	
Male, n (%)	441 (38.6%)	135 (49.1%)	.002
Race, n (%)			
Asian/Pacific Islander	37 (3.2%)	7 (2.5%)	.068
Black or African American	139 (12.2%)	26 (9.5%)	
White	902 (79.0%)	236 (85.8%)	
Other	11 (1.0%)	0 (0.0%)	
Unknown/Refused	53 (4.6%)	6 (2.2%)	
Ethnicity, n (%)			.016
Hispanic/Latino	665 (58.2%)	184 (66.9%)	
Non-Hispanic/Latino	419 (36.7%)	84 (30.5%)	
Unknown/Refused	58 (5.1%)	7 (2.5%)	
Country of origin-birth, n (%)			
US	512 (44.8%)	72 (26.2%)	
Foreign	460 (74.7)	156 (25.3%)	
Unknown/Refused	170 (78.3%)	47 (21.7%)	<.001
Body Mass Index, kg/m^2			<.001
Median (Q1, Q3)	26.3 (23.2, 30.0)	28.3 (25.6, 31.9)	
Missing	19 (1.7%)	3 (1.1%)	
Tobacco use, n (%)			.288
Current	66 (5.8%)	20 (7.3%)	
Quit	149 (13.0%)	43 (15.6%)	
Never	927 (81.2%)	212 (77.1%)	
Diabetes, n (%)	110 (9.6%)	32 (11.6%)	.315
Aspirin, n (%)	50 (4.4%)	6 (2.2%)	.119

Categorical and continuous characteristics were compared across those who did and did not have the outcome of interest (CRAs) using Fisher's exact test and Wilcoxon rank sum test, respectively.

Table 2. Logistic Regression Model Equation

Variable	Estimate	Standard error	Odds ratio (95% CI)	*P* value
(Intercept)	−8.8030	1.0136	0.00 (0.00–0.00)	<.001
Age	0.1414	0.0203	1.15 (1.11–1.20)	<.001
Male	0.4350	0.1740	1.54 (1.10–2.17)	.012
US-born	−0.6374	0.1898	0.53 (0.36–0.77)	.001
BMI	0.0455	0.0156	1.05 (1.02–1.08)	.004
Diabetes	0.0679	0.2799	1.07 (0.62–1.85)	.808
Aspirin use	−1.2801	0.5079	0.28 (0.10–0.75)	.012

Ethnicity, race, and smoking were not significant during model building, and not included in the final model.
BMI, body mass index; CI, confidence interval.

regression), 0.64 (95% CI: 0.57–0.71, random forest), 0.71 (95% CI: 0.65–0.77, gradient boosting), and 0.70 (95% CI: 0.63–0.76, artificial neural network). Given AUC and simplicity, the logistic regression model was selected as the final model (Table 2). At the optimal cutpoint of 0.16, the model had a sensitivity and specificity of 95% and 12%, respectively, in persons aged 45 years and more. In persons aged less than 45 years, sensitivity and specificity were 28% and 85%, respectively.

In this study, we derive and validate a model with good performance statistics for the presence of CRAs in average-risk individuals aged less than 50 years undergoing colonoscopy, which could be used for decisions about early screening. As EOCRC rises, there is a clear dilemma: healthcare burden, suboptimal screening uptake, and age of diagnosis being less than that of screening initiation suggest that we need innovative solutions.[1] Our model serves as a first step toward replacing the present “one-size-fits-all” approach.

The promising specificity of our model in persons aged less than 45 years is of note. While individuals may harbor CRAs at an early age, these lesions will not be detected within current guidelines. Implementing risk-stratified modeling can facilitate cancer prevention—not only early detection—and balance prevention with the practicalities of limited healthcare resources. That the sensitivity and specificity vary so widely in those on either side of age 45 reflects the relatively lower rate of adenoma detection in persons aged less than 45 years. This also makes risk-stratified screening an excellent consideration, if validated in external models. Our model is also notable as it contains readily available predictors and was created in a diverse population. A 2022 risk prediction model for EOCRC (not CRAs) used genetic and environmental risk scores, with AUC estimates of 0.54–0.63.[2] That our model has superior performance statistics while containing more accessible predictors is a strength. Another EOCRC model had approximate AUC 0.75, yet this was focused on male veterans.[3] We also use a cohort that approximates a true “average-risk” population. Other risk prediction models incorporate persons with inflammatory bowel disease or family history, groups that should be undergoing guideline-recommended screening.[4,5] Finally studies on risk factors for CRAs in younger persons are limited, and while it is reasonable to assume that they parallel those of EOCRC, this is an understudied area. Prior literature has demonstrated risk factors for EOCRC include being male and increasing age, while aspirin use and healthy lifestyle and weight are considered to be protective.[6–8] We demonstrate these are all predictors of CRAs, as well, reinforcing the need for healthy lifestyles to lessen the risk of cancer and precancerous lesions.

Importantly, our model demonstrates that multiple risk factors together can predict risk of CRA, but that individually, the effect size of each risk factor alone is low. This highlights the need for models such as this one, and underlines that future studies should use prospective cohorts with questionnaires that capture data not typically available in electronic medical records to create a comprehensive and accurate risk prediction model. Our study has limitations. In our convenience sample, CRA prevalence of 19.4% is relatively low, but given our broader age range and that not all examinations were screening, this is not unexpected.[9] Our study reflects a diverse population in South Florida, and future studies should confirm our findings to ensure broad generalizability. There may be unmeasured confounders, such as diet quality, alcohol, and lifestyle. Similarly, we do not have access to nuanced family history, although clearly identified high-risk individuals were not included. Our finding that being of US origin is associated with reduced risk of CRAs highlights the need for risk models derived from diverse cohorts, as this has not previously been described, although this too should be corroborated in larger studies as it may be due to factors that we are unable to capture (eg, access to and engagement with healthcare, socioeconomic differences, etc.) or a function of sample size/power (where we cannot adjust for all other relevant factors). Country of origin/birth is asked in our electronic medical record, and as a self-reported measure, may be subject to misclassification. Finally, this prediction model is not yet externally validated. The immediate next steps should be continued investigation within a prospective cohort with sufficient granularity and diversity to refine a risk prediction model that can help stratify individuals aged less than 50 years for colorectal cancer screening.

RYAN HOOD[1,*]
DIVYA DASANI[2,*]
CATHERINE BLANDON[3]
SHRIA KUMAR[3,4]

[1] Miller School of Medicine at the University of Miami, Miami, Florida
[2] Department of Medicine, Miller School of Medicine at the University of Miami, Miami, Florida
[3] Division of Digestive Health and Liver Diseases, Department of Medicine, Miller School of Medicine at the University of Miami, Miami, Florida
[4] Sylvester Comprehensive Cancer Center, Miller School of Medicine at the University of Miami, Miami, Florida

Correspondence:
Address correspondence to: Shria Kumar, MD, MSCE, 1120 NW 14th St, Locator Code C-240, Miami, Florida 33136. e-mail: shriakumar@med.miami.edu.

Supplementary Materials

Material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gastha.2024.04.009.

References

1. Kastrinos F, et al. Gastroenterology 2023;164(5):812–827.
2. Archambault AN, et al. J Natl Cancer Inst 2022;114(4):528–539.
3. Imperiale TF, et al. Cancer Prev Res (Phila) 2023;16(9):513–522.
4. Deng JW, et al. J Gastroenterol Hepatol 2023;38(10):1768–1777.
5. Gu J, et al. BMC Cancer 2022;22(1):122.
6. Low EE, et al. Gastroenterology 2020;159(2):492–501.e7.
7. Nguyen LH, et al. JNCI Cancer Spectr 2018;2(4):pky073.
8. O'Sullivan DE, et al. Clin Gastroenterol Hepatol 2022;20(6):1229–1240.e5.
9. Bilal M, et al. Am J Gastroenterol 2022;117(5):806–808.

***Denotes co-first authorship.**

Abbreviations used in this paper: **AUC, area under the curve; BMI, body mass index; CRA, colorectal adenoma; CRC, colorectal cancer; EOCRC, early-onset colorectal cancer; IBD, inflammatory bowel disease; OR, odds ratio; ROC, receiver operator characteristic**

Most current article

2772-5723
https://doi.org/10.1016/j.gastha.2024.04.009

Received February 28, 2024. Accepted April 17, 2024.

Conflicts of Interest:
The authors disclose no conflicts.

Funding:
This work is supported by an ASCO Conquer Cancer Career Development Award (2023CDA-9764863415) and the Sylvester Comprehensive Cancer Center (5P30CA240139).

Ethical Statement:
This study was approved by the Institutional Review Board at the University of Miami (20230850).

Data Transparency Statement:
The data analyzed in this study were obtained from the University of Miami Health System. Restrictions apply to the availability of these data, and they are not publicly available. Deidentified data are however available from the authors upon reasonable request, with permission of the University of Miami Institutional Review Board.

Reporting Guidelines:
STROBE.