

Predicting Early-Onset Colorectal Cancer in Individuals Below Screening Age Using Machine Learning and Real-World Data

Chengkun Sun¹, Erin M. Mobley, Ph.D.^{2,3}, Michael B. Quillen⁴, Max Parker⁴, Meghan Daly, M.D.², Rui Wang, Ph.D.⁵, Isabela Visintin, M.D.², Ziad Awad, M.D.², Jennifer Fishe, M.D.⁶, Alexander Parker, Ph.D.^{7,3}, Thomas J. George, M.D.^{3,4}, Jiang Bian, Ph.D.^{1,3} and Jie Xu, Ph.D.^{1,*}

¹ Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL 32611;

² Department of Surgery, College of Medicine, University of Florida, Jacksonville, FL 32209;

³ University of Florida Health Cancer Center, University of Florida, Gainesville, FL 32610

⁴ Department of Medicine, College of Medicine, University of Florida, Gainesville, FL 32611;

⁵ Center for Data Solutions, College of Medicine, University of Florida, Jacksonville, FL 32209;

⁶ Department of Emergency Medicine, College of Medicine, University of Florida, Jacksonville, FL 32209;

⁷ College of Medicine, University of Florida, Jacksonville, FL 32209;

* Correspondence: xujie@ufl.edu;

Abstract

Background: Colorectal cancer (CRC) is now the leading cause of cancer-related deaths among young Americans. Our study aims to predict early-onset CRC (EOCRC) using machine learning (ML) and structured electronic health record (EHR) data for individuals under the screening age of 45.

Methods: We identified a cohort of patients under 45 from the OneFlorida+ Clinical Research Consortium. Given the distinct pathology of colon cancer (CC) and rectal cancer (RC), we created separate prediction models for each cancer type with various ML algorithms. We assessed multiple prediction time windows (0, 1, 3, and 5 years) and ensured robustness through propensity score matching (PSM) to account for confounding variables. Model performance was assessed using established metrics. Additionally, we employed the Shapley Additive exPlanations (SHAP) to identify risk factors for EOCRC.

Results: Our study yielded results, with Area Under the Curve (AUC) scores of 0.811, 0.748, 0.689, and 0.686 for CC prediction, and 0.829, 0.771, 0.727, and 0.721 for RC prediction at 0, 1, 3, and 5 years, respectively. Notably, predictors included immune and digestive system disorders, along with secondary cancers and underweight, prevalent in both CC and RC groups. Blood diseases emerged as prominent indicators of CC.

Conclusion: This study highlights the potential of ML techniques in leveraging EHR data to predict EOCRC, offering valuable insights for potential early diagnosis in patients who are below the recommended screening age.

Keywords: Colorectal cancer (CRC), Prediction, Machine learning (ML), Electronic health records (EHRs)

1. Introduction

Colorectal cancer (CRC) is a significant public health challenge, ranking as the third leading cause of cancer-related mortality among both males and females in the United States.¹ It is estimated that in 2023, approximately 153,020 individuals will be diagnosed with CRC, and 52,550 will succumb to the disease.¹ While cancer is typically a disease of older age, a concerning trend has emerged – the increasing incidence of early-onset colorectal cancer (EOCRC) in individuals younger than the age of 50 years.^{1,2} This increased incidence has led the US Preventive Services Task Force to modify its recommendations, lowering the age to start CRC screening to age 45.³ Patients diagnosed with EOCRC tend to present at later stages and face lower disease-specific survival rates, underscoring the need for early detection and treatment initiation.⁴ Nevertheless, challenges in addressing EOCRC are compounded by poorly defined risk factors and the role of diagnostic delays. As a result, early prediction and comprehensive understanding of the risk factors of EOCRC are essential for prevention and treatment, particularly for patients who fall below the recommended screening age.

The rapid integration of artificial intelligence (AI) and big data analytics has significantly expanded the horizons of medical research and clinical care.⁵ Diverse data sources, including imaging and genomic data, have been harnessed for CRC detection through the application of statistical and machine learning (ML) algorithms. Some approaches have included the analysis of tumor DNA and circulating RNA expression profiling data to identify potential pathogenic factors.^{6,7} Additionally, computer tomography (CT)—based radionics, combined with ML algorithms, have been employed to predict the KRAS mutation in CRC patients, demonstrating the potential of ML in clinical decision support.⁸ Further, a random forest (RF) model trained with standard clinical and pathological prognostic variables, coupled with MRI images, achieved an impressive Area Under the Curve (AUC) score of 0.94 when predicting survival in CRC patients, highlighting the importance of MRI-based texture features patient survival prediction.⁹ However, imaging data produces a small number of unexplainable predictors (~100), and does not consistently improve

diagnostic accuracy and disease prediction, especially when only using imaging data.¹⁰ Furthermore, advanced imaging modalities and genomic data can be costly, with limited accessibility, and lack diversity and representativeness in samples, which could impact timely and accurate diagnosis for all individuals affected by EOCRC or widen already present disparities in patient outcomes.

In contrast to imaging and genomic data, structured data from the electronic health record (EHR) offer a more accessible and cost-effective data source for initial research. Originally designed for administrative and billing purposes, structured EHR data have evolved into valuable tools for healthcare research, capturing a wealth of patient information, including clinical diagnoses, procedures, medications, and laboratory results, among others.¹¹ The integration of ML and deep learning with EHR data has demonstrated substantial potential for disease prediction, including Alzheimer's disease, gestational diabetes mellitus (GDM), and coronary heart disease (CHD).¹²⁻

¹⁴ In the context of CRC, several ML approaches have been employed to predict the risk of the disease. For example, Shanbehzadeh et al. used structured EHR data and four data mining algorithms to predict CRC risk, identifying critical attributes for the prediction model using the weight statistical Chi-square test.¹⁵ However, the weight statistical Chi-square test assumes independence among variables, which may not hold true in complex datasets where variables are likely correlated. Another study leveraged convolutional neural networks to predict CRC risk based on the structured EHR data from the Taiwan National Health Insurance database.¹⁶ Hisham et al. explored multiple ML methods to construct predictive models for CRC among patients aged between 35 and 50.¹⁷ However, these studies faced challenges in effectively matching cases and control groups, leading to increased bias and concerns regarding confounding. Furthermore, another limitation across studies is the failure to distinguish between colon and rectal cancers, despite the differences in clinical presentation, molecular carcinogenesis, pathology, surgical topography and procedures, and multimodal treatment strategies between these two cancers.¹⁸ Additionally, the lack of model explanations regarding clinical diagnosis of CRC undermined the

interpretability and reliability of their strategies. As a result, there is a pressing need for improved methodologies to enhance the reliability and understanding of ML models in EO CRC prediction. In light of these gaps in existing literature, our primary objective is to build separate ML models for the prediction of colon and rectal cancers in patients prior to reaching the screening age of 45 years, leveraging EHR data to identify potential unique risk factors for each cancer type. To achieve this goal, we employed a range of traditional ML models to predict these cancers at various time intervals before their onset (following the setting of Li et al¹²). To mitigate potential data bias and confounding issues, we implemented propensity score matching (PSM) to establish a comparable matched control group.¹⁹ Additionally, we utilized the Shapley Additive exPlanations (SHAP)²⁰ for model interpretation, thereby enhancing our ability to discern the contribution of individual features. By improving interpretability, our aim is to pinpoint the risk factors that pre-date the development of EO CRC.²¹

2. Materials and Methods

2.1. Data source and study population

This study used de-identified EHR data from the OneFlorida+ Clinical Research Consortium funded by the Patient-Centered Outcomes Research Institute (PCORI), as one of the 8 clinical data research network contributing to the National Patient-Centered Clinical Research Network (PCORnet).²² The OneFlorida+ data encompasses a wide range of patient characteristics from health systems across the southeast, including EHR data collected using the PCORnet Common Data Model²² regarding demographics, diagnoses, medications, procedures, vital signs, lab tests, and more.

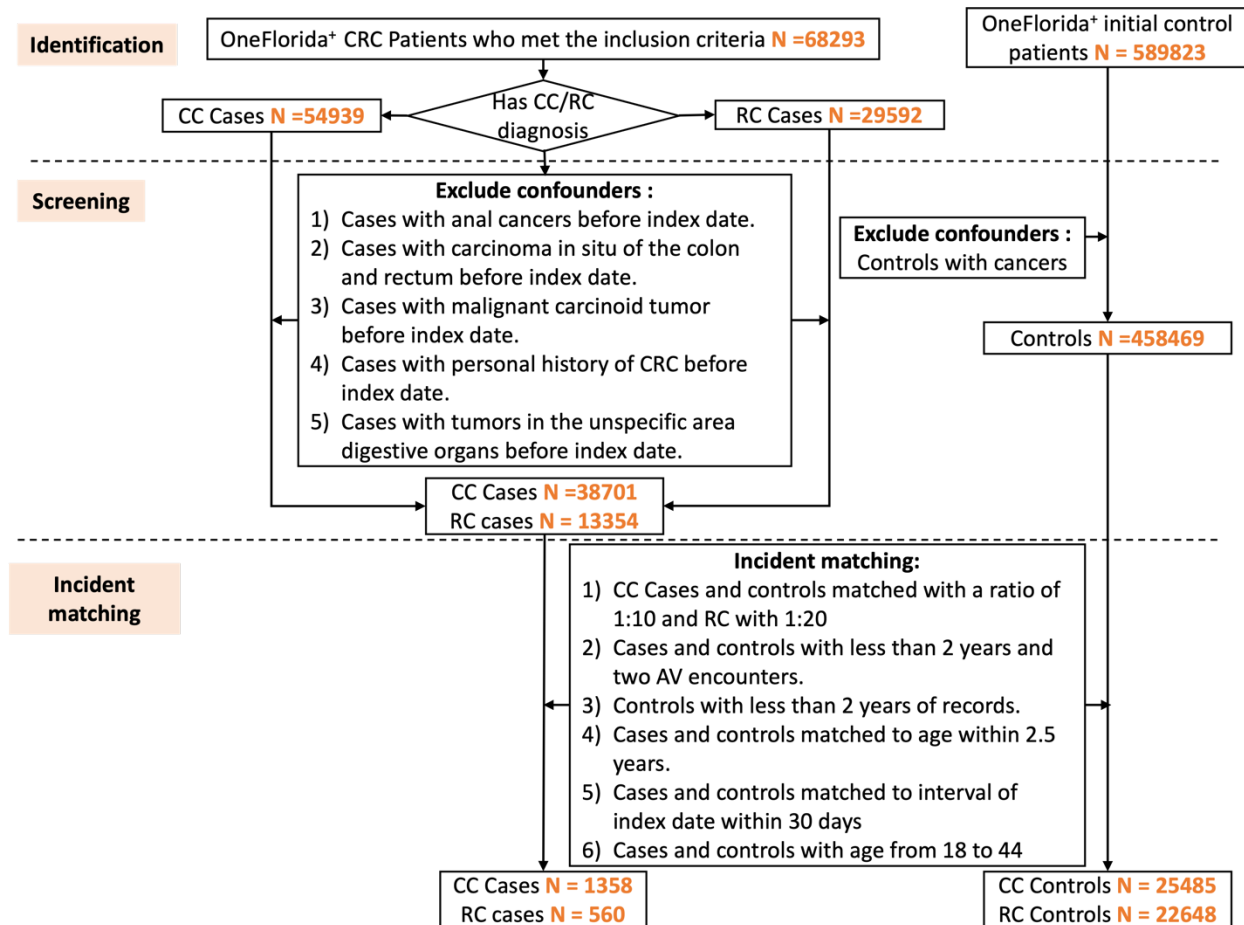


Figure 1. Flowchart of patient selection from OneFlorida+.

The construction of our study cohort using OneFlorida+ is outlined in Figure 1. OneFlorida+ identified individuals from the OneFlorida+ network, with encounters from January 2012 to January 2023 who met our inclusion criteria as either a case or control. We identified cases of colon cancer (CC) using the International Classification of Diseases (ICD)-9 code of C18* or C49A4 or ICD-10 code of 153*, or rectal cancer (RC) cases with the ICD-9 code of C19*, C20*, C21.0, C21.1, and ICD-10 code of 154.0 and 154.1. The initial cohort consisted of 68,293 CRC cases (54,939 CC cases, 29,592 RC cases), and 589,823 controls. From those, we excluded patients diagnosed with both CC and RC, other prior cancers, or who were diagnosed ≥ 45 years of age. Our final study cohort comprised 1,358 CC cases with 25,485 controls and 560 RC cases with 22,648 controls.

We used an incident matching process to match cases and controls to ensure a fair comparison across these groups. Initially, we retained cases and controls with more than two years of records and at least two encounters before the first onset date of either colon or rectal cancer and ensured that the age gap between matched cases and controls was within 2.5 years. By calculating propensity scores based on race, ethnicity, sex, and birth year (within 2.5 years), we employed a narrow caliper of 0.05 with a nearest neighbor approach to achieve a 1:5 case-to-control ratio for each prediction window group.²³ This rigorous methodology ensures a balanced study population for reliable analysis and EOCRC prediction.

2.2. Study setting

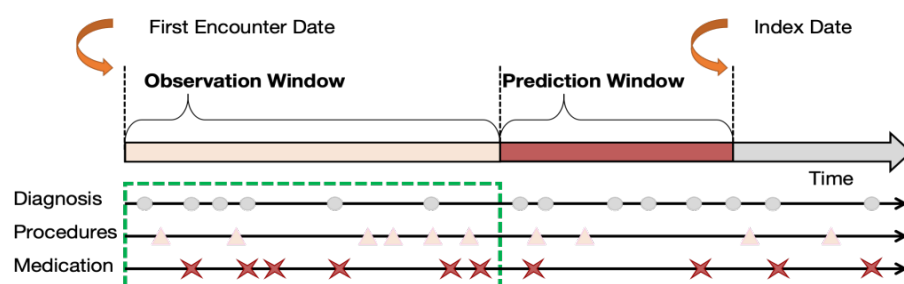


Figure 2. Visualization of the observation and prediction windows for the prediction task. The index date for CRC cases is the date of diagnosis. For the control group, the index date is defined as the closest encounter date to the diagnosis date of the matched case group. The prediction window is the time period before the index date during which CRC cases are predicted. The observation window refers to the specific period during which data is collected or observed for analysis.

Further, we incorporated a range of different observation periods and prediction windows to test our prediction algorithms, considering the different use cases. We considered four different prediction windows: 0-year, 1-year, 3-year, and 5-year before CRC diagnosis.

2.3. Data preprocessing

The predictors we extracted include demographics, vitals, diagnoses, medications, and procedures documented throughout the observation periods. Age at index date was calculated and categorized into three groups (e.g., 18-29, 30-39, 40-44). One-hot encoding²⁴ was used to represent age groups, race, and sex variables. For missing data, we imputed the missing values

with the mean of the numerical data derived from the entire sample within each prediction window group. Furthermore, BMI data was categorized into clinically relevant groups, including underweight (≤ 18.5), normal (18.5-23), overweight (23-30), and obese (≥ 30). Diastolic and systolic measurements were categorized into distinct hypertension stages.

Diagnoses, which were initially represented using ICD-9 and ICD-10 codes, were subjected to a data dimensionality reduction process that mapped them into Phecodes.^{25,26} Revenue codes and Current Procedural Terminology (CPT) codes²⁷ were leveraged to capture billed medical procedures. To integrate these data, we also employed the Clinical Classifications Software (CCS) code.²⁸ For drug information, National Drug Code (NDC)²⁹ and RxNorm codes were used for encoding. NDC codes were mapped into RxNorm codes, and further consolidated into Anatomical Therapeutic Chemical (ATC) classes.³⁰ To ensure completeness, all features that could not be mapped were retained to prevent any missing information. These steps to transform the data enhanced interpretability and relevance of our predictive models.

2.4. Experiments and validation

We explored several widely used ML models, including linear models such as logistic regression (LR) and the support vector machine (SVM), as well as nonlinear models like XGBoost and RF. We adopted two modeling strategies, including (1) prediction without CRC-related features; and (2) prediction without cancer-related features, covering the CRC-related features. For the first strategy, features that may be indicative of CRC differential diagnoses (e.g., neoplasm of unspecified nature of digestive system) or treatments for CRC (e.g., chemotherapy, radiotherapy) were removed from the models and not used as predictors. For the second strategy, we took a more stringent approach by eliminating all diagnoses, drugs, and procedures that could be associated with any cancer from the extracted predictors. This step aimed to identify risk factors while eliminating the influence of other types of cancers, enabling us to focus exclusively on non-

cancer-related predictors. Regardless of the feature engineering strategy, we maintained a consistent experimental setup.

The entire dataset was randomly split into a training dataset and a testing dataset with a ratio of 4:1. Model optimization was conducted on the training set through 5-fold cross-validation, and we fine-tuned hyperparameters using Bayesian optimization. To ensure the reproducibility of our experiments, we fixed the random state seed across all model runs.

To assess the effectiveness of our models comprehensively, we employed a battery of evaluation metrics, including AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. To mitigate the risk of overfitting and to derive robust confidence intervals (CIs), we implemented a bootstrapping strategy. This involved conducting 100 experiments by randomly resampling the training and testing datasets. In addition to traditional performance metrics, we delved into the interpretability of the XGBoost models. Specifically, we computed SHAP values²⁰ to gain insights into the inner workings of the ML algorithms and to identify the core contribution predictors. This approach aimed to unveil the high-risk factors associated with EO CRC, shedding light on the most influential features in our prediction model.

3. Results

Table 1 provides an overview of the identified study cohorts after PSM for both CC and RC across various prediction windows. Notably, CC cases outnumber RC cases, with approximately twice as many CC cases. Patients in the RC groups were slightly older compared to those in the CC group. Sex distribution in the RC groups was closer to parity (2:3 male to female) than in the CC group (2:5 male to female). Both RC and CC groups exhibited diverse racial and ethnic representation. In addition, as the prediction window lengthened, the number of cases decreased. Specifically, there were 560 (0-year), 560 (1-year), 383 (3-year), and 225 (5-year) RC cases, and 1358 (0-year), 1358 (1-year), 884 (3-year), and 532 (5-year) CC cases in each prediction window.

Table 1. Descriptive statistics in case and control groups.

Baseline variables	CC cases (n = 1358)	CC controls (n = 6790)	RC cases (n= 560)	RC controls (n= 2800)
Age, mean (std)	36.54 (5.88)	36.69 (5.73)	37.70 (5.70)	36.80 (5.53)
Sex, N (%)				
Female	938 (69.07)	4461 (65.70)	323 (57.68)	1617 (57.75)
Male	420 (30.93)	2329 (34.30)	237 (42.32)	1183 (42.25)
Race and Ethnicity, N (%)				
Hispanic	338 (24.89)	1527 (22.49)	101 (18.04)	514 (18.36)
Non-Hispanic White	554 (40.80)	2893 (42.61)	239(42.68)	1212 (43.29)
Non-Hispanic Black	353 (25.99)	1857 (27.35)	178 (31.79)	887 (31.68)
Other	14 (1.03)	66 (0.97)	4 (0.71)	9 (0.32)
Unknown	99 (7.29)	447 (6.58)	38 (6.79)	178 (6.36)

CC: colon cancer; RC: rectal cancer

Table 2. AUC comparison for CC prediction using ML models across different prediction windows (0, 1, 3, and 5 years).

Feature Strategy	Model	0-year AUC (95% CI)	1-year AUC (95% CI)	3-year AUC (95% CI)	5-year AUC (95% CI)
Excluding CRC-related features	LR	0.809 (0.806,0.812)	0.733 (0.73,0.736)	0.683 (0.679,0.688)	0.674 (0.668,0.679)
	SVM	0.748 (0.745,0.751)	0.689 (0.685,0.692)	0.614 (0.61,0.618)	0.616 (0.61,0.621)
	RF	0.811 (0.808,0.814)	0.748 (0.745,0.751)	0.689 (0.684,0.694)	0.686 (0.68,0.692)
	XGBoost	0.802 (0.799,0.806)	0.745 (0.741,0.748)	0.689 (0.684,0.694)	0.657 (0.651,0.663)
Excluding cancer-related features	LR	0.788 (0.786,0.791)	0.713 (0.71,0.716)	0.669 (0.665,0.674)	0.661 (0.656,0.667)
	SVM	0.725 (0.722,0.729)	0.646 (0.643,0.65)	0.604 (0.6,0.608)	0.611 (0.606,0.617)
	RF	0.77 (0.767,0.773)	0.716 (0.713,0.719)	0.684 (0.679,0.688)	0.663 (0.658,0.668)
	XGBoost	0.76 (0.757,0.764)	0.714 (0.711,0.717)	0.662 (0.657,0.666)	0.643 (0.638,0.648)

LR: logistic regression; SVM: Support vector machines; RF: Random Forest

Table 2 presents the results of CC prediction using two feature engineering strategies: one excluding CRC-related features and the other excluding cancer-related features. Additional evaluation metrics for CC prediction across all settings can be found in the Supplementary Material (refer to Tables S1-S2). In most cases, tree-based models (XGBoost and RF)

outperformed linear models (SVM and LR), yielding higher AUC values. Specifically, after removing CRC-related features, the RF model achieved the highest AUC [95% CI] for the 0-year prediction (0.811 [0.808, 0.814]), while RF performed best for the 1-year (0.748 [0.745, 0.751]), 3-year (0.689 [0.684, 694]), and 5-year (0.686 [0.68, 0.692]) predictions for CC. However, after removing features associated with prior cancers, the model performance decreased: LR achieved AUC [95% CI] values of 0.788 [0.786, 0.791] for 0-year prediction; RF achieved AUC [95% CI] values of 0.716 [0.713, 0.719] for 1-year, 0.684 [0.679, 0.688] for 3-year, and 0.663 [0.658, 0.668] for 5-year prediction. Performance metrics, including specificity, sensitivity, PPV, NPV, and F1 score, exhibited similar trends.

Table 3. AUC comparison for RC prediction using ML models across different prediction windows (0, 1, 3, and 5 years).

Feature Strategy	Model	0-year AUC (95% CI)	1-year AUC (95% CI)	3-year AUC (95% CI)	5-year AUC (95% CI)
Excluding CRC-related features	LR	0.819 (0.815,0.824)	0.763 (0.758,0.767)	0.722 (0.716,0.728)	0.693 (0.686,0.7)
	SVM	0.78 (0.774,0.785)	0.694 (0.689,0.699)	0.656 (0.649,0.662)	0.658 (0.65,0.665)
	RF	0.826 (0.822,0.83)	0.771 (0.766,0.777)	0.719 (0.713,0.726)	0.72 (0.712,0.727)
	XGBoost	0.829 (0.825,0.834)	0.766 (0.762,0.771)	0.727 (0.721,0.732)	0.721 (0.713,0.729)
Excluding cancer-related features	LR	0.807 (0.803,0.812)	0.748 (0.743,0.752)	0.709 (0.703,0.715)	0.69 (0.683,0.697)
	SVM	0.767 (0.761,0.772)	0.686 (0.68,0.691)	0.653 (0.646,0.659)	0.656 (0.648,0.663)
	RF	0.806 (0.802,0.81)	0.756 (0.751,0.76)	0.724 (0.718,0.73)	0.711 (0.704,0.719)
	XGBoost	0.811 (0.806,0.815)	0.749 (0.744,0.753)	0.724 (0.718,0.729)	0.679 (0.672,0.687)

LR: logistic regression; SVM: Support vector machines; RF: Random Forest

Table 3 provides RC prediction results using the same feature engineering strategies and four prediction windows. Additional evaluation metrics for RC prediction across all settings can be found in the Supplementary Material (refer to Tables S3-S4). Again, after removing CRC-related features, the XGBoost model achieved the highest AUC [95% CI] for the 0-year prediction (0.829

[0.825, 0.834]), while RF performed best for the 1-year (0.771 [0.766,0.777]), and XGBoost did best for 3-year (0.727 [0.721, 0.732]), and 5-year (0.721 [0.713,0.729]) predictions for RC. Eliminating cancer-related features resulted in a performance decrease: XGBoost achieved AUC [95% CI] values of 0.811 [0.806, 0.815] for 0-year prediction. RF achieved AUC [95% CI] values of 0.756 [0.751, 0.76] for 1-year, 0.724 [0.718, 0.73] for 3-year, and 0.711 [0.704, 0.719] for 5-year prediction. Performance metrics exhibited consistent trends.

In both the CC and RC prediction tasks, we observed a decline in model performance as the prediction window length increased. Notably, when we removed cancer-related features, the AUC declined. This highlights the pivotal role these features play in enhancing prediction performance. To gain deeper insights into the risk factors associated with these findings, we present SHAP summary plots for CC and RC predictions using two feature engineering strategies and for 0-year and 3-year prediction windows in Figures 3 and 4. Supplementary SHAP summary plots for all other models can be found in the Supplementary Material (Figures S1–S2). Within the CC group, several predictors emerged as positively associated with the risk of CC. Notably, several diagnoses involving various tumors, such as suspected cancer, secondary malignant neoplasm, benign neoplasm of uterus, benign neoplasm of skin, neoplasm of uncertain behavior, neoplasm of uncertain behavior of skin, cancer of other female genital organs and myeloproliferative diseases were identified as influential factors. Gastrointestinal symptoms, encompassing conditions like gastrointestinal hemorrhage, other disorders of intestine, other symptoms involving the abdomen and pelvis, noninfectious gastroenteritis, appendiceal conditions, diverticulosis and diverticulitis, intestinal obstruction without hernia, and disorders of the intestine also exhibited a positive association with CC risk. Additionally, medical procedures related to gastrointestinal diseases and symptoms, including upper gastrointestinal endoscopy, were significantly associated with the development of CC. In the RC group, similar positive predictors were identified, mirroring the trends observed in the CC group, including gastrointestinal symptoms (e.g., gastrointestinal hemorrhage, anal and rectal conditions) and the presence of other cancers

or tumors (e.g., secondary malignant neoplasms, benign neoplasms of the uterus or skin). Additionally, the presence of autoimmunity, diseases associated with a potentially weakened immune system (e.g HIV, viral warts and HPV), and conditions like hemorrhoids were linked to a heightened long-term risk of RC. Being underweight was a significant symptom associated with both CC and RC. Conversely, obesity, overweight and normal weight appeared to be negatively associated with RC development. Importantly, after removing cancer-related features from consideration, the significance of anemias surged to the forefront in both the CC and RC groups. These included indicators such as iron deficiency anemias and other anemias. Nevertheless, gastrointestinal diseases and immunodeficiency pathological changes remained substantial factors contributing to CC risk, while factors such as HPV and weight retained their significance as primary determinants of RC. The use of anti-inflammatory or antirheumatic medications were associated with decreased risk of RC.

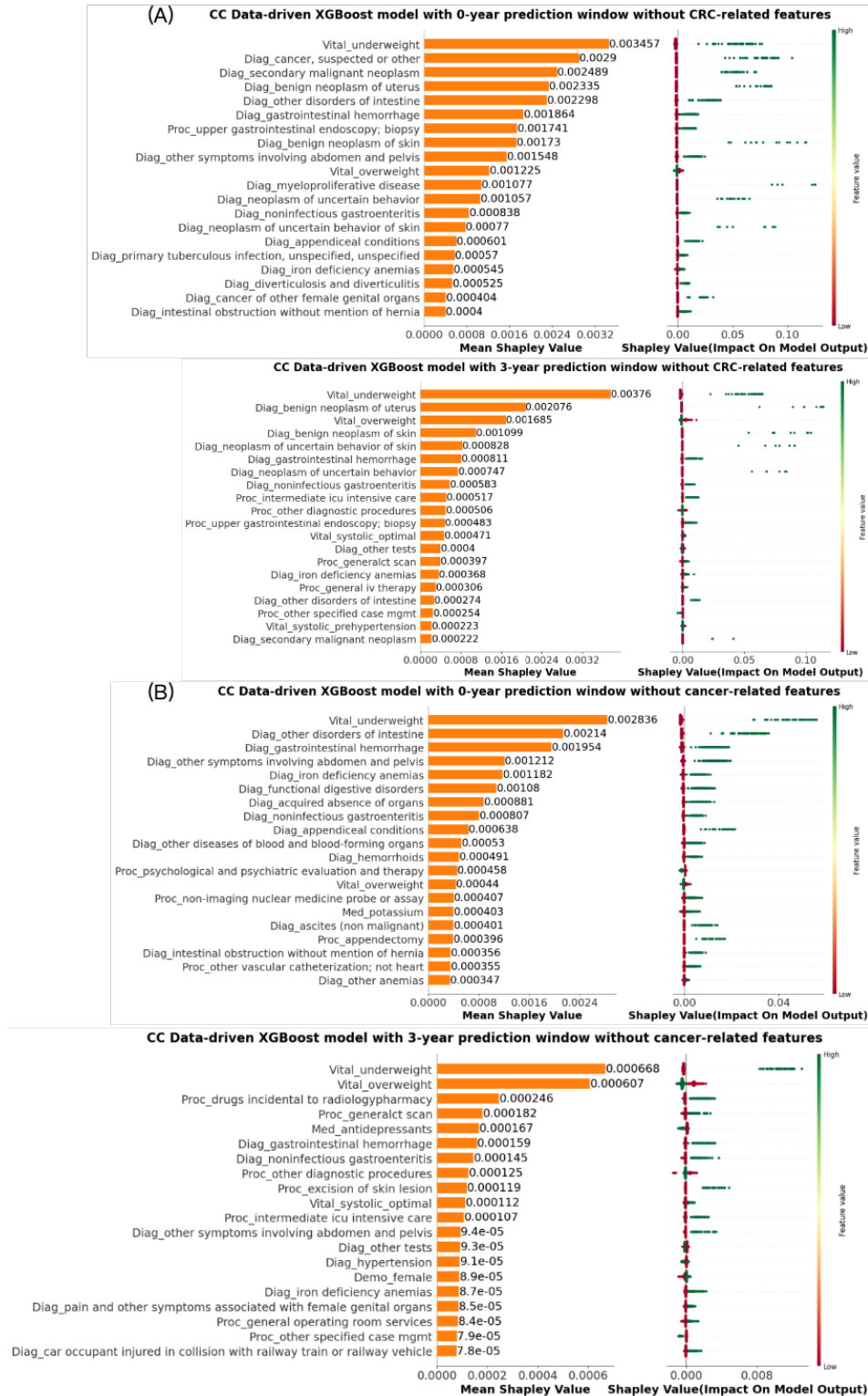


Figure 3. SHAP summary plot of the top 20 features in CC prediction using XGBoost models with 0-year and 3-year prediction windows: (A) excluding CRC-related features; (B) excluding cancer-related features. The prefix before the “_” in the y-axis labels of plots indicates the source of the corresponding features in the PCORnet data model. Specifically, these sources are: Diagnosis (Diag), Procedure (Proc), Medication (Med), Vital Signs (Vital), and Demographics (Demo).



Figure 4. SHAP summary plot of the top 20 features in RC prediction using XGBoost models with 0-year and 3-year prediction windows: (A) excluding CRC-related features; (B) excluding cancer-related features.

3. Discussion

In this study, we employed four traditional ML algorithms (i.e., XGBoost, RF, SVM, and LR) and obtained informative results predicting EOCRC using structured EHR data. In most cases, the tree-based models, (XGBoost and RF) outperformed linear models, achieving the best AUC scores for various prediction windows. Additionally, even after excluding cancer diagnosis variables (e.g., pancreatic, skin, thyroid cancer), undergoing cancer-related procedures (e.g., liver biopsy, bone marrow biopsy), cancer treatments (e.g., cisplatin, doxycycline), our models continued to achieve acceptable AUC scores. Immune and digestive system disorders, blood diseases, and secondary cancers were identified as significant predictors.

Most of our experimental findings were consistent with existing published research. Cancer-related diseases and diagnoses emerged as risk factors leading to the diagnosis of EOCRC, both for colon and rectal cancers. For example, uterine cancer was identified as a driver of EOCRC, suggesting a potential genetic association between these malignancies in younger patients.³¹ Research also demonstrates that the incidence rate of second primary cancers among survivors is significantly higher than cancer in the general population, and survivors experience notable morbidity and mortality from their cancer treatment.³² Additionally, the use of CT scans for other medical reasons could contribute to the incidental identification of EOCRC cases.³³ Notably, we know that some forms of cancer treatment (e.g., radiation) predisposes one to an increased risk for secondary malignancies, including EOCRC, particularly in patients surviving a childhood cancer.³⁴

Inflammatory bowel diseases (IBD) are well established risk factors for colorectal cancers, particularly during young adulthood. The chronic inflammation associated with IBD leads to the release of growth cytokines, excess blood flow, and metabolic free radicals, all of which contribute to the heightened risk of developing colorectal cancer.³⁵ Therapies for IBD sometimes involve immune suppression, another known risk factor for cancers. Furthermore, many gastrointestinal diseases can cause malabsorption or malnutrition,³⁶ resulting in patients being underweight which

can also contribute to immune dysfunction or suppression.³⁷ However, overweight patients were at low risk of EO CRC as our analysis demonstrated despite emerging evidence that being overweight may be associated with an increased risk of tumor recurrence and colorectal carcinogenesis.^{38,39} The temporal use of antibiotics in relation to subsequent development of EO CRC is an interesting finding as it supports several previously reported roles that the gut microbiome may play in colorectal cancer protection and development.⁴⁰ Our analysis highlighted that the diagnosis of iron deficiency anemia pre-dated colon cancer, but had less association with rectal cancers. It is logical, given that colon cancers are situated more proximal in the gastrointestinal tract, causing occult chronic blood loss and subsequent anemia rather than overt gross bleeding as is typically evident from rectal cancers.

Additionally, our study observed a significantly higher incidence of colorectal cancer cases among HIV-infected patients compared to HIV-uninfected individuals.⁴¹ The heightened risk can be attributed to disruptions in immune function caused by immunodeficiency, which exposes individuals to a higher susceptibility against cancer-causing viruses, including HPV, EBV, KSHV, etc., as evidenced in our analysis.⁴² Another notable finding was the association between colon cancer and diseases of myeloproliferative disease. Similar to other cancers, the potential link could be related to genetics, treatments that induce DNA damage that could predispose to EO CRC, and chronic immune dysregulation. Overall, our study sheds light on the complex interplay between inflammatory bowel diseases, malnutrition, immune function, and specific blood-related diseases in the development of CRC. Understanding these relationships is crucial in advancing our knowledge of EO CRC risk factors and devising targeted interventions for at-risk populations.

Our study does have several limitations. First, the mechanism through which identified medical factors are associated with EO CRC is speculative. For example, CT scans contributed significantly to the model's performance, but the specific reasons are unclear. EHRs didn't record the reason why patients underwent CT scans. Perhaps some patients obtained CT scans

because of symptoms related to undiagnosed CRC while others received CT scans for other reasons with the incidental finding of CRC. It is less likely that CT scans could be associated with causing CRC due to radiation exposure. For that to occur, the cumulative lifetime exposure would need to be very high with exposure over a number of decades for that to occur. Perhaps CT imaging itself is just a surrogate for access to care whereby EO CRC is more likely to be eventually diagnosed as opposed to patients who might expire for other reasons with CRC, but prior to a diagnosis. Second, the exclusion of confounder samples and features posed difficulties, given the lack of universally accepted standards for phenotype definitions and ambiguous descriptions. These challenges hindered the design of the most optimal experiment.⁴³ Third, our experiments are carried out based on the EHR data, which inherently contains flaws, including missing values and potential mistakes in records. Efforts were made to fill missing values, but comprehensive amendments remained challenging. The characteristics of the EHR data, such as temporality, irregularity, sparsity, and data imbalance, can result in abnormal outcomes when applying machine learning models.^{44,45} Moreover, systematic bias, such as erroneous use of ICD codes due to strategic billing, may impact data-driven predictions.⁴⁶ The EHR data utilized in the OneFlorida+ dataset are overwhelmingly hospital-based data, which may further introduce selection bias in that ambulatory practices (where most relatively healthy patients receive their routine care) is inconsistently represented. Despite these limitations, we believe our model provides interesting insights into medical variables that pre-date and are associated with EO CRC.

4. Conclusion

In conclusion, our study demonstrated the potential of traditional ML algorithms in predicting EO CRC using real-world data for individuals below the screening age guideline. The identification of significant predictors and their consistency with academic research findings provide valuable insights for pursuing additional hypotheses or targeting potential patients at risk for EO CRC. However, addressing the challenges and limitations related to data quality,

experimental design, and ML models' development is essential for improving the accuracy and reliability of EOCRC prediction models. Future research should focus on refining the experimental design, exploring alternative feature selection techniques, incorporating LLM based on both ambulatory and inpatient data, and integrating domain knowledge to enhance the performance of the prediction models. Ultimately, these efforts will contribute to early detection and better management of CRC, with the goal to improve patient outcomes.

References

1. Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A. & Jemal, A. Colorectal cancer statistics, 2023. *CA Cancer J. Clin.* (2023) doi:10.3322/caac.21772.
2. Hofseth, L. J. et al. Early-onset colorectal cancer: initial clues and current views. *Nat. Rev. Gastroenterol. Hepatol.* 17, 352–364 (2020).
3. US Preventive Services Task Force et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* 325, 1965–1977 (2021).
4. McClelland, P. H.-T., Liu, T. & Ozuner, G. Early-Onset Colorectal Cancer in Patients under 50 Years of Age: Demographics, Disease Characteristics, and Survival. *Clin. Colorectal Cancer* 21, e135–e144 (2022).
5. Analysis of healthcare big data. *Future Gener. Comput. Syst.* 109, 103–110 (2020).
6. Luo, H. et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Sci. Transl. Med.* 12, (2020).
7. Song, W. & Fu, T. Circular RNA-Associated Competing Endogenous RNA Network and Prognostic Nomogram for Patients With Colorectal Cancer. *Front. Oncol.* 9, 1181 (2019).
8. Porto-Álvarez, J. et al. CT-Based Radiomics to Predict KRAS Mutation in CRC Patients Using a Machine Learning Algorithm: A Retrospective Study. *Biomedicines* 11, 2144 (2023).
9. Daye, D. et al. Quantitative tumor heterogeneity MRI profiling improves machine learning–based prognostication in patients with metastatic colon cancer. *Eur. Radiol.* 31, 5759–5767 (2021).
10. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med* 5, 48 (2022).
11. The Evolving Use of Electronic Health Records (EHR) for Research. *Semin. Radiat. Oncol.* 29, 354–361 (2019).
12. Li, Q. et al. Early prediction of Alzheimer’s disease and related dementias using real-world electronic health records. *Alzheimers. Dement.* (2023) doi:10.1002/alz.12967.

13. Artzi, N. S. et al. Prediction of gestational diabetes based on nationwide electronic health records. *Nat. Med.* 26, 71–76 (2020).
14. Du, Z. et al. Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation. *JMIR Med Inform* 8, e17257 (2020).
15. Shanbehzadeh, M. et al. Comparison of four data mining algorithms for predicting colorectal cancer risk. *J Adv Med Biomed Res* 29, 100–108 (2021).
16. Wang, Y.-H., Nguyen, P.-A., Islam, M. M., Li, Y.-C. & Yang, H.-C. Development of Deep Learning Algorithm for Detection of Colorectal Cancer in EHR Data. *Stud. Health Technol. Inform.* 264, 438–441 (2019).
17. Hussan, H. et al. Utility of machine learning in developing a predictive model for early-age-onset colorectal neoplasia using electronic health records. *PLoS One* 17, e0265209 (2022).
18. Paschke, S. et al. Are Colon and Rectal Cancer Two Different Tumor Entities? A Proposal to Abandon the Term Colorectal Cancer. *Int. J. Mol. Sci.* 19, (2018).
19. Caliendo, M., Kopeinig, S. & Forschungsinstitut zur Zukunft der Arbeit. Some Practical Guidance for the Implementation of Propensity Score Matching. *JOURNAL ECONOMIC SURVEYS* (2005).
20. Shapley, L. S. A Value for N-Person Games. (1952).
21. Smith, M. & Alvarez, F. Identifying mortality factors from Machine Learning using Shapley values - a case of COVID19. *Expert Syst. Appl.* 176, 114832 (2021).
22. Shenkman, E. et al. OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute With a Community-Based Distributive Medical Education Model. *Acad. Med.* 93, 451–455 (2018).
23. Jessica N. Jacovidis, Kelly J. Foelber, and S. Jeanne Horst. The Effect of Propensity Score Matching Method on the Quantity and Quality of Matches. *THE JOURNAL OF EXPERIMENTAL EDUCATION* 85, 535–558.
24. Hinton, G. E. Learning distributed representations of concepts. *Proceedings of the eighth annual conference of the cognitive science society* 1, (1986.).
25. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to

- PheRS. *Annu Rev Biomed Data Sci* 4, 1–19 (2021).
26. PheWAS - Phenome Wide Association Studies. <https://phewascatalog.org/>.
 27. American Medical Association. *Introduction to CPT Coding: Basic Principles to Learning, Understanding, and Applying the CPT Code Set*. (American Medical Association Press, 2011).
 28. Elixhauser, A. *Clinical Classifications for Health Policy Research, Version 2: Software and User's Guide*. (1996).
 29. CreateSpace Independent Publishing Platform & Office of the Investigator General. *The Food and Drug Administration's National Drug Code Directory*. (Createspace Independent Publishing Platform, 2018).
 30. Anatomical Therapeutic Chemical (ATC) Classification Index with Defined Daily Doses (DDDs): List A: Sorted According to ATC Code Including Defined Daily Doses (DDDs) for Plain Substances. (1997).
 31. Denlinger, C. S. & Weinberg, D. S. Sporadic breast, ovarian, or uterine cancers as risk factors for colorectal cancer. *Curr. Colorectal Cancer Rep.* 5, 24–29 (2009).
 32. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Survivorship. NCCN https://www.nccn.org/professionals/physician_gls/pdf/survivorship.pdf.
 33. Kilian G M Brown, Michael J Solomon, Kate Mahon, Sarah O'Shannassy. Management of colorectal cancer. *the BMJ* (2019).
 34. Bhatia, S. & Sklar, C. Second cancers in survivors of childhood cancer. *Nat. Rev. Cancer* 2, 124–132 (2002).
 35. Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Gastroenterology Rev* 14, 89–103 (2019).
 36. Beirer, A. Malnutrition and cancer, diagnosis and treatment. *memo - Magazine of European Medical Oncology* 14, 168–173 (2021).
 37. Risk Factors for Early-Onset Colorectal Cancer. *Gastroenterology* 159, 492–501.e7 (2020).
 38. Ye, P., Xi, Y., Huang, Z. & Xu, P. Linking Obesity with Colorectal Cancer: Epidemiology and Mechanistic Insights. *Cancers* 12, (2020).
 39. A paradox between preoperative overweight/obesity and change in weight during postoperative

- chemotherapy and its relationship to survival in stage II and III colorectal cancer patients. *Clin. Nutr.* 40, 2410–2419 (2021).
40. Agus, A. et al. Western diet induces a shift in microbiota composition enhancing susceptibility to Adherent-Invasive *E. coli* infection and intestinal inflammation. *Sci. Rep.* 6, 19032 (2016).
 41. Coghill, A. E., Shiels, M. S., Suneja, G. & Engels, E. A. Elevated Cancer-Specific Mortality Among HIV-Infected Patients in the United States. *J. Clin. Oncol.* 33, 2376–2383 (2015).
 42. Yarchoan, S.-Y. P. K. L. How immunodeficiency can lead to malignancy. *Immunology* 2021, 287–295 (2021).
 43. Brandt, P. S. et al. Characterizing variability of electronic health record-driven phenotype definitions. *J. Am. Med. Inform. Assoc.* 30, 427–437 (2023).
 44. Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Computing* 55, 1–40 (2022).
 45. Kohane, I. S. et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. *J. Med. Internet Res.* 23, e22219 (2021).
 46. Zima., J. B. E. A. Machine Learning, Natural Language Processing, and the Electronic Health Record: Innovations in Mental Health Services Research. *Technology in Mental Health* 70, 346–349 (2019).