

Article

Development and Internal Validation of a Machine Learning-Based Colorectal Cancer Risk Prediction Model

Deborah Jael Herrera ^{1,†} , Daiane Maria Seibert ^{2,†} , Karen Feyen ² , Marlon van Loo ³ , Guido Van Hal ^{1,*,‡}  and Wessel van de Veerdonk ^{1,3,‡} 

¹ Family Medicine and Population Health Department (FAMPOP), Faculty of Medicine and Health Sciences, University of Antwerp, 2610 Antwerp, Belgium; deborah.herrera@student.uantwerpen.be (D.J.H.); wessel.vandeveerdonk@thomasmore.be (W.v.d.V.)

² Centre of Expertise—Design and Technology, Campus De Nayer, Thomas More University of Applied Sciences, 2860 Sint-Katelijne-Waver, Belgium; daiane.seibert@thomasmore.be (D.M.S.)

³ Centre of Expertise—Care and Well-Being, Campus Zandpoortvest, Thomas More University of Applied Sciences, 2800 Mechelen, Belgium; marlon.vanloo@thomasmore.be

* Correspondence: guido.vanhal@uantwerpen.be

† These authors contributed equally to this work.

‡ These authors also contributed equally to this work.

Abstract: Background: Colorectal cancer (CRC) remains a leading cause of cancer-related mortality worldwide. While screening tools such as the fecal immunochemical test (FIT) aid in early detection, they do not provide insights into individual risk factors or strategies for primary prevention. This study aimed to develop and internally validate an interpretable machine learning-based model that estimates an individual's probability of developing CRC using readily available clinical and lifestyle factors. **Methods:** We analyzed data from 154,887 adults, aged 55–74 years, who participated in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. A risk prediction model was built using the Light Gradient Boosting Machine (LightGBM) algorithm. To translate these findings into clinical practice, we implemented the model into a risk estimator that categorizes individuals as average, increased, or high risk, highlighting modifiable risk factors to support patient–clinician discussions on lifestyle changes. **Results:** The LightGBM model incorporated 12 predictive variables, with age, weight, and smoking history identified as the strongest CRC risk factors, while heart medication use appeared to have a potentially protective effect. The model achieved an area under the receiver operating characteristic curve (AUROC) of 0.726 (95% confidence interval [CI]: 0.698–0.753), correctly distinguishing high-risk from average-risk individuals 73 out of 100 times. **Conclusions:** Our findings suggest that this model could support clinicians and individuals considering screening by guiding informed decision making and facilitating patient–clinician discussions on CRC prevention through personalized lifestyle modifications. However, before clinical implementation, external validation is needed to ensure its reliability across diverse populations and confirm its effectiveness in real-world healthcare settings.

Keywords: prediction model; colorectal cancer; screening; machine learning; risk threshold



Academic Editor: Tamara Čačev

Received: 24 January 2025

Revised: 12 March 2025

Accepted: 14 March 2025

Published: 24 March 2025

Citation: Herrera, D.J.; Seibert, D.M.; Feyen, K.; van Loo, M.; Van Hal, G.; van de Veerdonk, W. Development and Internal Validation of a Machine Learning-Based Colorectal Cancer Risk Prediction Model. *Gastrointest. Disord.* **2025**, *7*, 26. <https://doi.org/10.3390/gidisord7020026>

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the Creative Commons

Attribution (CC BY) license

(<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, advances in colorectal cancer (CRC) screening have contributed to a significant decrease in mortality, mainly through the early detection and removal of precancerous lesions [1,2]. However, CRC remains a global burden, standing as the third most

commonly diagnosed malignancy and the second leading cause of cancer-related deaths worldwide, accounting for roughly 10 of all cancer cases. Many well-established screening guidelines have played a role in improving outcomes, yet these recommendations typically rely on generalized, age-based recommendations, advising average-risk individuals to begin screening at age 50 [3,4]. This one-size-fits-all approach does not account for individual variability in risk factors such as genetics, lifestyle, and family history, potentially leading to suboptimal screening outcomes, nor does it offer a straightforward means for clinicians to discuss personalized risks with patients.

Risk prediction models have been used to address these limitations by facilitating personalized, risk-based screening [4,5]. This strategy can optimize resource allocation in screening by focusing efforts and resources on those who are most likely to benefit, ensuring that high-risk individuals are prioritized for early detection. At the same time, it reduces unnecessary procedures in average-risk individuals, minimizing the potential for overdiagnosis, reducing patient burden, and preventing unnecessary costs and complications associated with unnecessary screening procedures [6].

Recent systematic reviews have evaluated existing CRC risk prediction models, revealing moderate to high discriminatory accuracy. Yet, despite encouraging developments in model-based risk assessment, recent reviews have highlighted recurrent hurdles. First, several existing tool algorithms that produce a risk score with little explanation of the contributing factors, such as lack of interpretability, impede clinicians' ability to counsel patients effectively [7]. Second, many existing tools rely on difficult-to-gather data, such as precise dietary intake, detailed genomics, or extensive lifestyle inventories, making them challenging to implement in settings with limited consultation time or in under-resourced clinics [8,9]. Third, while some models have reported promising discriminatory accuracy, they fail to include potentially relevant populations and risk factors, use inadequate statistical methods, select arbitrary thresholds that do not adapt to the needs of different populations or healthcare environments, and provide insufficient reporting on model stability and clinical applicability [7].

Among widely used models, Q Cancer is a primary care-based risk algorithm designed for integrated use in electronic health records (EHRs) [10]. While it incorporates multiple clinical and demographic variables, including family history, BMI, smoking status, and comorbidities, it did not include lifestyle factors (e.g., processed meat intake and physical activity) and is primarily trained on UK-based datasets, limiting its generalizability to other populations. Additionally, its reliance on pre-existing medical records may not fully capture modifiable risk factors that patients can address through lifestyle changes. Similarly, the APCS risk score predicts advanced colorectal neoplasia (ACN) using age, gender, smoking, BMI, diabetes, and alcohol intake, but it does not integrate modifiable lifestyle factors beyond smoking and alcohol [11,12]. The Kaminski risk score [13] and NCI-CRC Risk Assessment Tool include additional variables, such as polyp history and NSAID use, but their reliance on detailed clinical data limits their feasibility in primary care settings, where rapid risk assessment is crucial [14].

Even among models with high discriminatory accuracy, challenges such as complex risk stratification approaches and unclear cutoffs hindered their practical implementation. For optimal clinical use, these models should be well-integrated into EHRs to automatically display patient-specific risk scores, making them easy to use during consultations [15,16]. Additionally, these models must be interpretable, allowing clinicians to understand the factors driving predictions and communicate them effectively to individuals [7,15]. To address these limitations, we utilized data from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, a large randomized controlled trial, focusing on clinical feasible variables that could be collected within a typical consultation. Using this robust dataset, we

aimed to develop and internally validate an interpretable ML-based CRC risk prediction model that estimates an individual's probability of developing CRC using readily available clinical and lifestyle factors. The model also provides transparent, feature-level insights enabling clinicians to refine risk stratification, personalize screening recommendations, and support informed decision making in healthcare settings.

2. Methods

2.1. Study Design and Population

This study utilized secondary data from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, a large randomized controlled trial that enrolled approximately 155,000 participants, aged from 55 to 74 years, between November 1993 and July 2001. Participants were recruited in 10 PLCO Screening Centers across the United States. They were randomly assigned to either the intervention arm, receiving specific cancer screenings, or the control arm, receiving standard medical care.

Participants were eligible if they met the original inclusion criteria established by the PLCO trial as follows: aged 55–74 years at enrollment; no personal history of CRC; not currently undergoing treatment for any cancer (except basal or squamous cell skin cancer); and no prior complete removal of the entire colon. Additionally, from April 1995 onwards, individuals who had undergone a colonoscopy, sigmoidoscopy, or barium enema within the last three years were excluded. Participants involved in other cancer screening or prevention trials were also excluded.

For CRC screening, intervention arm participants underwent two flexible sigmoidoscopies (FSGs) as follows: one at baseline and another three or five years later, depending on protocol adjustments. Cancer diagnoses were collected until 31 December 2009 and mortality data through 2015, providing a median follow-up of 11.3 years. The dataset was obtained through the Cancer Data Access System (CDAS) by submitting a project proposal and adhering to data use policies (approval reference: <https://cdas.cancer.gov/approved-projects/3478/>) (accessed on 15 August 2022).

2.2. Outcome Measure

CRC was assessed using detailed records from various sub-datasets of the PLCO Trial [17]. CRC diagnoses included all stages from I to IV and cases where the carcinoid could not be assessed. FSG exam results were categorized as negative, abnormal suspicious, abnormal non-suspicious, or inadequate, with the adequacy of exams recorded. Lesion details included the largest lesion's location (rectum, sigmoid colon, descending colon, and splenic flexure) and size (diminutive, small, and large).

Advanced adenomas were identified if polyps were villous, had dysplasia, or were large (≥ 1 cm), including in situ carcinomas. Polyps were further classified based on whether they were completely villous (not applicable, not completely villous, completely villous, and unknown), the level of dysplasia (none, severe, moderate, mild, and unknown), and histology (adenoma, hyperplastic, benign polyp, colonic mucosa, other, and not available).

2.3. Data Processing

We followed a structured process of data handling, from factor preselection to model validation. A schematic overview of the entire data processing workflow and the definition of key terms and definitions to machine learning are shown in Figure 1 and Table S1 (Supplementary Materials), respectively.

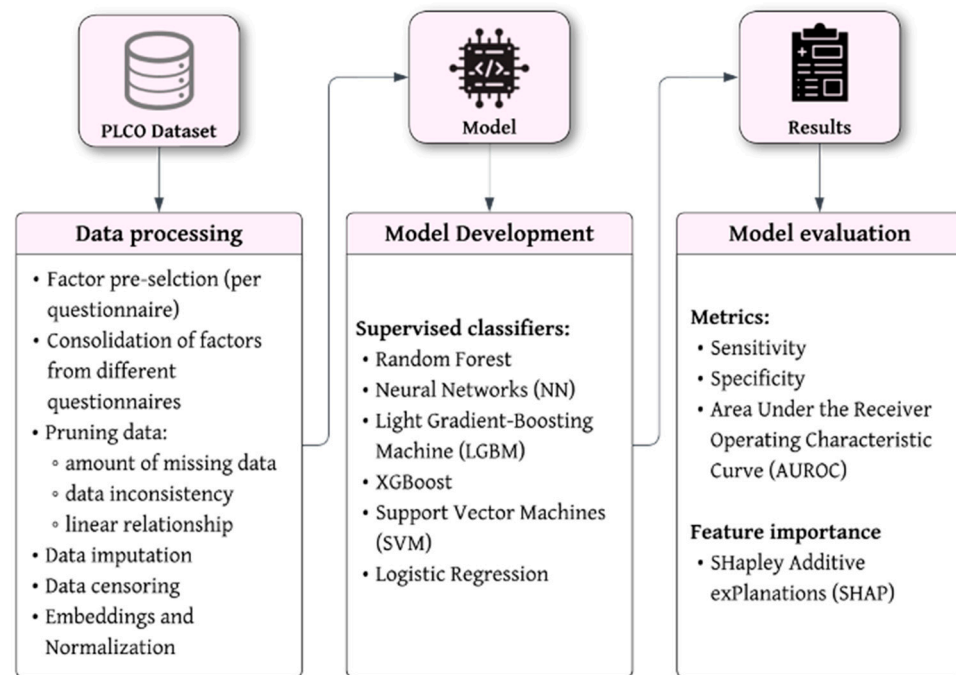


Figure 1. Schematic diagram for dataset processing, model development, and validation. The arrows represent the flow of data and processes in model development using the PLCO dataset. Each step transitions from data processing → model development → model evaluation, ensuring the model is trained, validated, and assessed for accuracy.

2.3.1. Factor Preselection

We identified the relevant risk factors based on our systematic review of existing models, which synthesized evidence on advanced neoplasia in CRC screening [9]. The review highlighted key risk factors commonly used across studies, and we selected those available in the PLCO dataset that are known to have a strong association with CRC risk. While we focused on factors well-established in the literature, we also tested different factors available in the dataset to assess their potential relevance to our model.

2.3.2. Pruning of Data

The initial data pruning was performed with patients that did not handle any questionnaire to the study (4106 cases). The first feature selection process was conducted manually, excluding factors that were geographically related, unrelated, or following screening for CRC (e.g., fractures and chemotherapy). Additionally, factors exhibiting high collinearity (Pearson's $r \geq 0.85$) and those with negligible correlation with the target variable (Pearson's ≤ 0.001) were excluded.

Variables that patients may find difficult to answer accurately were also excluded. For instance, asking patients to report the exact grams of red meat or fiber consumed daily can be challenging and may lead to unreliable data due to recall bias or estimation errors. Instead, we favored more practical variables such as general drinking habits (e.g., consumption of alcoholic drinks per week) that are easier for patients to report and for clinicians to assess.

Further feature selection was conducted based on the existing literature on CRC screening, ensuring that the selected variables align with the constraints of a standard medical consultation, which typically lasts 15 min. Additionally, variables that are impractical to obtain within a clinical setting were excluded to enhance the model's applicability in real-world medical practice. With a final set of 12 factors, 267 cases were found with more than 30% missing data, which were excluded to ensure data integrity and reliability for

model training. Furthermore, 796 cases missing age were also removed, as age is a critical factor in cancer risk prediction.

2.3.3. Handling Missing Data

To address missing values in the dataset, we applied mode imputation based on similarity for categorical factors and k-nearest neighbors for numerical features. These approaches are grounded according to the principle that imputing missing data using records similar to the one with missing values yields more accurate results than methods relying on the entire dataset [18].

Correlated factors and similar patients were used to predict and fill in missing data points. In practice, for each participant with missing data, we identified a subset of participants with similar demographic and clinical profiles. Correlated factors were used to define the similarity among participants. For example, if the hypertension factor was missing for a participant, we imputed this value using the mode for hypertension of participants who shared at least 4 similar characteristics. The characteristics used for the imputation of the categorical factors—diabetes, hypertension, heart problems, and smoking status—are age, sex, BMI, number of cigarettes per day, and alcohol-related factors (Figure 2). On the other hand, for the numerical factors, the number of neighbors used for imputation was 10, meaning that the 10 most similar cases were used to perform the imputation of the missing value.

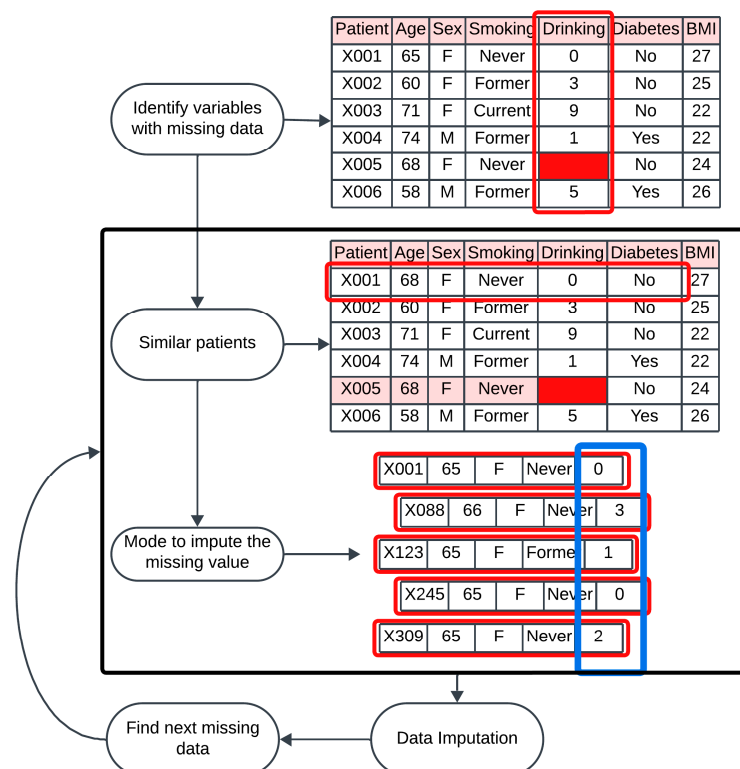


Figure 2. Step-by-step process of data imputation for categorical factors. The red squares highlight missing values and similar patients used for imputation, while the blue squares show the mode-imputed values. The arrows indicate the process of identifying, selecting, and filling missing data using similar patient records.

2.3.4. Data Restriction by Age

For patients diagnosed with CRC, we restricted data based on the age at which CRC was detected or the age at colonoscopy to ensure consistent data handling. Additionally, a second form of censoring was applied to patients initially categorized as negative for

CRC but who later died from CRC (missed diagnoses or post-colonoscopy CRCs (PC-CRCs)). These cases often arise from missed lesions, incomplete polyp removal, or rapidly developing cancers that were undetectable during colonoscopy. Between these cases, inconsistencies were found where deaths initially attributed to CRC were later identified to be due to other causes. To reduce uncertainty and potential biases, we decided to exclude these patients from our analysis.

2.4. Model Development

Supervised Classifiers

Several ML algorithms were tested to develop the CRC risk prediction model. Following the sensitivity analysis, we decided to use LightGBM due to its demonstrated efficiency and accuracy in handling large-scale datasets (see Table S1 for the definition of LightGBM). This method followed a gradient boosting framework, combining weak models (decision trees), where each new model (tree) aims to correct the errors of the previous ones.

2.5. Model Evaluation

2.5.1. AUROC of the Model

In evaluating the performance of our machine learning model, we focused on three key metrics as follows: [1] the Area Under the Receiver Operating Characteristic curve (AUROC) [2], sensitivity, and [3] specificity. The AUROC provides an aggregate measure of the model's performance across all classification thresholds. The AUROC value ranges from 0 to 1, with higher values indicating better overall performance. An AUROC of 0.5 suggests no discriminative power, while an AUROC of 1.0 signifies perfect discrimination.

2.5.2. Sensitivity and Specificity of the Model

Sensitivity, or the true positive rate, measures the model's ability to correctly identify positive instances. It is calculated as the ratio of true positives (*TP*) to the sum of true positives and false negatives (*FN*), expressed as

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity, or the true negative rate, assesses the model's ability to correctly identify negative instances. It is defined as the ratio of true negatives (*TN*) to the sum of true negatives and false positives (*FP*), given by

$$\text{Specificity} = \frac{TN}{TN + FP}$$

These metrics provide a clear understanding of the model's performance and a balance in the detection of true positives, the exclusion of false positives, and the overall classification ability across various thresholds.

2.6. Regressor Model and Threshold Selection

The model used is a regressor that outputs a percentage representing the probability of a patient having CRC. Unlike a classifier that would simply return a binary outcome—1 for cancer cases and 0 for non-cancer cases—a regressor analyzes historical data to identify patterns and provides a probability that quantifies the individual's likelihood of CRC. This probabilistic output allows for more nuanced decision making in healthcare settings.

To determine whether patients are classified as positive or negative for CRC, a threshold percentage was selected. To obtain this threshold, we conducted a validation process

utilizing the test dataset that comprised 15% of the total dataset, separate from the training dataset (70%) and the evaluation dataset (15%).

2.7. Feature Importance

To enable clinicians to give personalized information to patients about their CRC risk, SHapley Additive exPlanations (SHAP) analysis was performed to identify the most important factors that predict the risk of CRC. SHAP values are calculated by inputting patient data to the trained model to explain each prediction's feature-level importance. This results in a ranked list of factors showing their contributions to the risk prediction.

3. Results

3.1. Participant Selection

The flowchart in Figure 2 outlines the participant selection and dataset preparation process for the development of the ML-based CRC model. Out of a total of 154,887 observations, 5169 (3.34%) were excluded due to excessive missing clinical data.

Following data imputation, an additional 241 cases were removed, as they were negative cases who later died from post-colonoscopy colorectal cancer (PCCRC). The final eligible participants consisted of 149,718 individuals, who were divided into the following four groups: 74,077 (49.5%) in the control arm, 8798 (5.9%) in the intervention arm with no flexible sigmoidoscopy (FSG), 27,276 (18.2%) in the intervention arm with one FSG, and 39,326 (26.3%) in the intervention arm with two FSGs. These groups were further split into training, evaluation, and test sets (Figure 3).

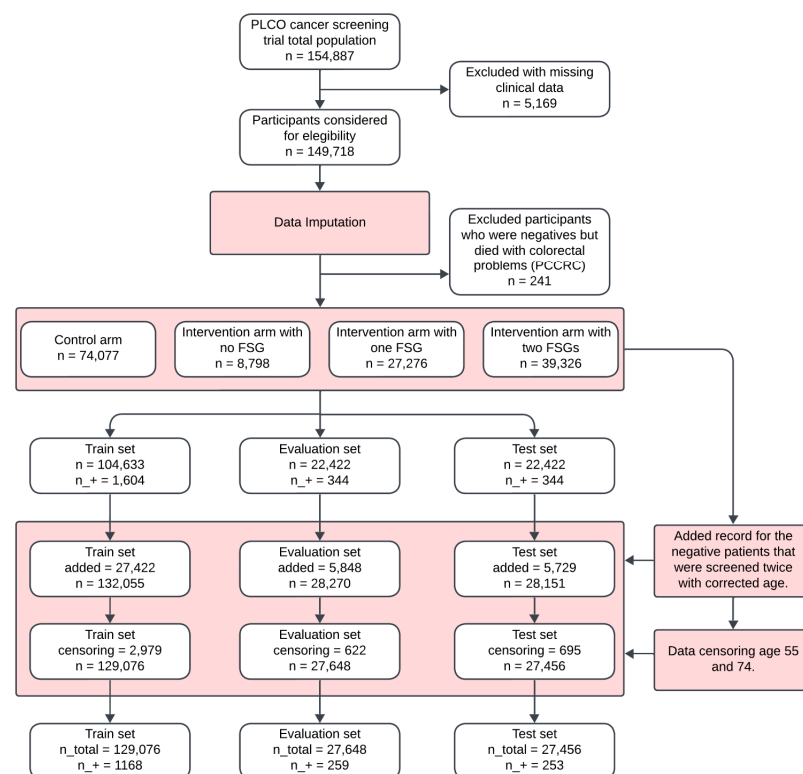


Figure 3. Inclusion criteria flowchart. PCCRC: post-colonoscopy colorectal cancers; $n_{+/-}$: number of positive cases, where advanced adenomas were detected. The arrows represent the step-by-step filtering, grouping, and processing of participants, from eligibility screening to dataset splitting, with exclusions, imputation, and age-based censoring to refine the final study population.

3.2. Participants' Demographic and Clinical Characteristics

The demographic and clinical characteristics of the study population were comparable across the training, evaluation, and test sets, as summarized in Table 1. The average age of participants was 62.64 years (SD $\pm$ 5.36), with a nearly equal gender distribution of 50.76% females and 49.24% males.

Table 1. Characteristics of the study population in the PLCO cancer screening trials, specifically for the CRC screening dataset and the distribution of positive cases across different CRC risk thresholds and categories.

Factor	Missing Quantity (n)	Value	Total (n = 149,718)
Age (years)*	0	Mean $\pm$ SD	62.64 $\pm$ 5.36
Sex	0	Female Male	50.76% 49.24%
Weight (kg)	868	Mean $\pm$ SD	79.44 $\pm$ 16.90
Height (cm)	345	Mean $\pm$ SD	170.32 $\pm$ 9.98
BMI	771	Mean $\pm$ SD	27.28 $\pm$ 4.92
Hypertension	178	Yes No	34.26% 65.74%
Heart problems	245	Yes No	13.63% 86.37%
Diabetes	215	Yes No	7.77% 92.23%
Smoking history	16	Current Former Never smoked	10.70% 43.09% 46.21%
Smoke quantity (daily)	142	0 1–10 11–20 21–30 31–40 41–60 61–80 81+	42.28% 13.76% 19.59% 10.70% 5.90% 3.05% 0.56% 0.14%
Alcohol (drinks/day)	28131	Mean $\pm$ SD	0.71 $\pm$ 1.88
Alcohol 40 * (drinks/day)	38232	Mean $\pm$ SD	0.70 $\pm$ 1.44

* Quantity of alcoholic drinks consumed: Includes alcohol from beer, wine, and liquor for individuals age 40–54 (measures in drinks per day). * Age at which the cancer was detected during FSG for positive cases and age at which the last screening test was performed for negative cases.

Health conditions were prevalent in the dataset, with 34.26% of participants diagnosed with hypertension, 13.63% reporting heart problems, and 7.77% diagnosed with diabetes. In terms of smoking history, 10.70% were current smokers, 43.09% were former smokers, and 46.21% had never smoked. Alcohol consumption was generally low across the population, with an average of 0.71 drinks per day (SD $\pm$ 1.88).

The demographic and clinical characteristics of participants were comparable across the training, evaluation, and test sets, with consistent mean age, sex distribution, BMI, and prevalence of health conditions (Table 2).

Table 2. Demographic and clinical characteristics of the training, evaluation, and test sets.

Factor	Categories	Training (n = 129,076)	Evaluation (n = 27,648)	Test (n = 27,456)
Sex	Female	50.27%	50.50%	50.70%
	Male	49.73%	49.50%	49.30%
Weight (kg)	Mean $\pm$ SD	79.61 $\pm$ 16.84	79.67 $\pm$ 16.68	79.49 $\pm$ 16.72
Height (cm)	Mean $\pm$ SD	170.51 $\pm$ 9.99	170.58 $\pm$ 10	170.49 $\pm$ 9.94
BMI	Mean $\pm$ SD	27.28 $\pm$ 4.88	27.28 $\pm$ 4.84	27.25 $\pm$ 4.82
Hypertension	Yes	33.58%	33.50%	33.64%
	No	66.42%	66.50%	66.36%
Heart pathology	Yes	13.54%	13.91%	13.71%
	No	86.46%	86.09%	86.29%
Diabetes	Yes	7.51%	7.56%	7.42%
	No	92.49%	92.44%	92.58%
Smoking history	Current	10.28%	10.35%	10.28%
	Former	42.95%	43.42%	43.27%
	Never smoked	46.77%	46.23%	46.45%
Smoke quantity (daily)	0	46.80%	46.27%	46.49%
	1–10	13.72%	13.70%	13.46%
	11–20	19.27%	19.72%	20.26%
	21–30	10.61%	10.85%	10.55%
	31 or more	9.59%	9.46%	9.25%
Alcohol (drinks/week)	0	22.14%	22.39%	22.02%
	1 to 7	76.84%	76.52%	76.93%
	More than 7	1.03%	1.09%	1.05%
Alcohol 40 (drinks/week)	0	18.18%	18.40%	18.06%
	1 to 7	80.74%	80.61%	80.71%
	More than 7	1.08%	0.99%	1.23%

3.3. Model Development and Performance

To develop a reliable machine learning model for predicting colorectal cancer (CRC) risk, we incorporated key clinical and lifestyle factors, including sex, age, weight, height, body mass index (BMI), hypertension, heart pathology, diabetes, smoking history, smoking quantity, and alcohol consumption. These factors were selected based on their established association with CRC risk.

3.3.1. Model Selection and Evaluation

To determine the most effective predictive model, we tested multiple machine learning algorithms on the PLCO dataset, comparing their ability to differentiate between individuals at risk of CRC and those not at risk. Table 3 presents a comparison of the most effective machine learning models, excluding those with poor performance. The LightGBM (LGBM) model demonstrated the best balance between sensitivity (correctly identifying individuals with CRC) and specificity (correctly identifying those without CRC). The Neural Network (NN) model exhibited a slightly higher accuracy (0.616) but a lower sensitivity (0.711) than LGBM. The Random Forest (RF) model performed similarly to LGBM, achieving

0.692 sensitivity and 0.601 specificity. Meanwhile, XGBoost showed the highest accuracy (0.649), with a more balanced sensitivity (0.68) and specificity (0.649).

Table 3. Performance of the models on the PLCO dataset. Models with poor performance were excluded from the table for clarity.

Model	TN	FP	FN	TP	Accuracy	Sensitivity	Specificity
NN	16,735	10,468	73	180	0.616	0.711	0.615
LGBM	16,519	10,684	64	189	0.609	0.747	0.607
RF	16,545	10,658	78	175	0.609	0.692	0.601
XGBoost	17,642	9561	81	172	0.649	0.68	0.649

TN: True Negatives; FP: False Positives; FN: False Negatives; TP: True Positives; NN: Neural Network; LGBM: LightGBM; RF: Random Forest; XGBoost: Extreme Gradient Boosting.

3.3.2. Model Performance Metrics

The model included the following factors: sex, age, weight, height, BMI, hypertension, heart pathology, diabetes, family history of CRC smoking history, smoking quantity, and alcohol consumption. Following hyperparameter tuning, the LightGBM model achieved an AUROC of 0.726 (Figure 4A and Table 3), reflecting moderate discriminative power for differentiating between positive and negative CRC cases. Furthermore, the model exhibits a sensitivity of 0.747, correctly identifying 74.7% of true positive cases, and a specificity of 0.6072, correctly identifying 60.72% of true negative cases.

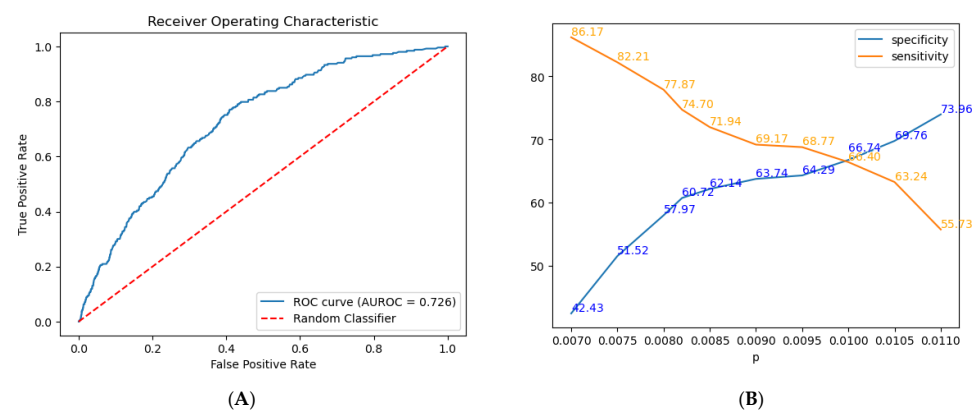


Figure 4. Model performance (A) after hyperparameter tuning using LightGBM and (B) the specificity vs. sensitivity of the model (p = threshold).

Additional commonly used metrics include the Positive Predictive Value (PPV) and Negative Predictive Value (NPV). The LGBM model achieved an NPV of 0.996 and a PPV of 0.017. However, due to the highly imbalanced nature of our dataset, with one positive case for every 89 negative cases, these metrics are significantly affected by class distribution.

3.4. Regressor Model

Figure 4B shows the trade-off between specificity and sensitivity at various probability thresholds (p) for our model. Specifically, we evaluated predictions across a range of thresholds, selecting a specific threshold for analysis. As p increases from 0.007 to 0.011, specificity improves from 42.43% to 73.96%, while sensitivity decreases from 86.17% to 55.73%.

3.5. Threshold Selection

To support screening, we prioritized sensitivity, selecting a lower threshold of 0.0082 to better identify at-risk individuals, balancing the risk of false positives with the need for early CRC detection. We considered the optimal threshold (0.0082) to maximize sensitivity and specificity at 0.7470 and 0.6072, respectively.

While the F2 score (0.007796), which weighs sensitivity twice as much as specificity, could have been used, with sensitivity at 0.798 and specificity at 0.560, it may overemphasize sensitivity at the cost of specificity, making our chosen threshold a more practical choice (Figure 4B). The probability threshold for categorizing CRC risk was based on the incidence of positive cases with higher risk scores (Table 4).

Table 4. Distribution of positive cases across different CRC risk thresholds and categories.

Threshold Interval	Risk Categories	Positive Cases	Total Cases per Interval	Percentage of Positive Cases
0–0.0082	Average	64	16,583	0.385937
0.0082–0.013	Increased	88	6676	1.318155
0.013–1	High	101	4197	2.406481

3.6. Feature Importance

Figure 5 displays the contribution of each factor to the CRC risk prediction model, utilizing SHAP. Factors are ordered along the y-axis based on their significance in the model's predictions. SHAP quantifies the impact of each factor on an individual's prediction, with positive values indicating an increased risk of CRC and negative values indicating a decreased risk.

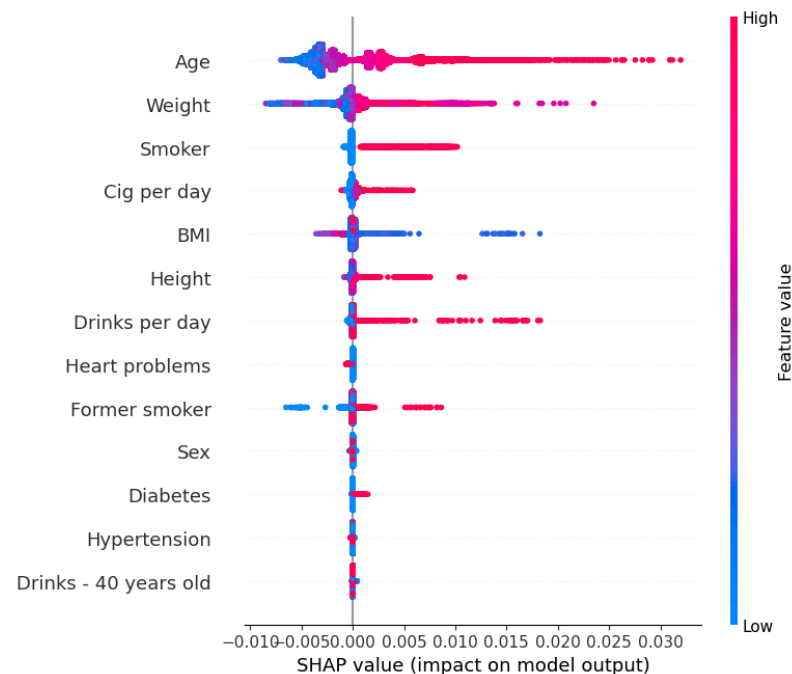


Figure 5. SHAP analysis on feature importance using SHAP analysis.

Our analysis identified age, weight, and smoking status as significant factors of increased CRC risk, implying that higher values in these factors are associated with higher predicted risk. Conversely, cases of regular use of medications for heart conditions are slightly correlated with a reduced risk of CRC.

3.7. Clinical Applicability of the Model

To facilitate real-world implementation, we integrated our CRC risk prediction model into an interactive risk estimator tool designed for use in primary care consultations. This tool dynamically adapts its assessment based on patient responses and provides personalized screening recommendations to support clinical decision making (<https://bibopp-acc.vito.be/orient/deelnemen>) (accessed on 6 March 2025) (Figure 6).

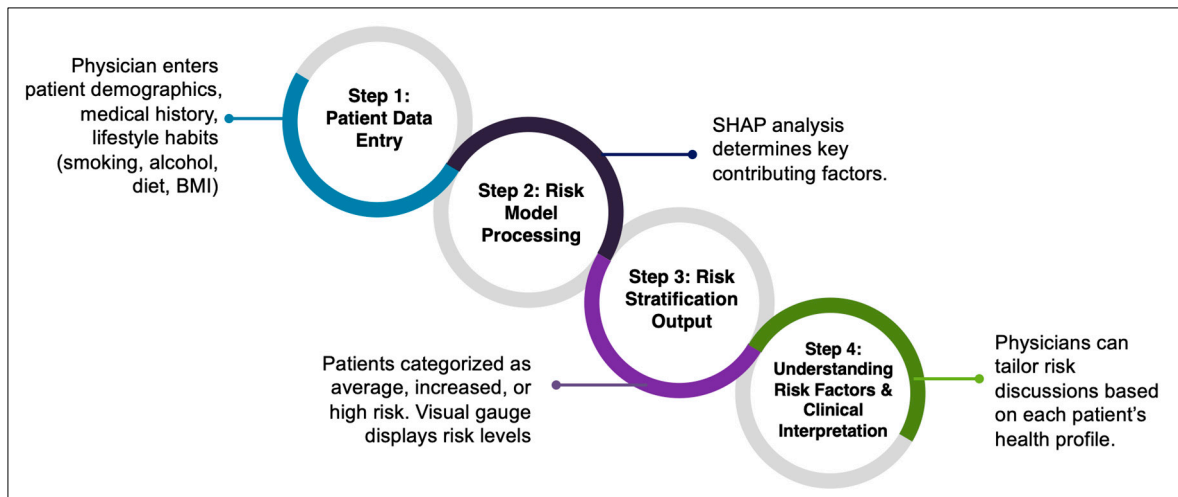


Figure 6. Workflow for using the CRC risk estimator.

The model generates a personalized risk score, categorizing patients into average-, increased-, or high-risk groups. Results are presented in an easy-to-understand, color-coded gauge, allowing both physicians and patients to quickly interpret their risk level and make informed decisions about screening and lifestyle modifications.

For high-risk patients with a family history of CRC, the tool automatically recommends a colonoscopy, ensuring they receive timely screening. For individuals without a family history, the tool assesses modifiable lifestyle risk factors such as diet, smoking, alcohol use, BMI, hypertension, and diabetes, offering personalized advice on reducing risk.

4. Discussion

4.1. Summary of Findings

While FIT is a widely used non-invasive screening tool with high sensitivity and specificity for detecting CRC [19], it does not inform patients about modifiable behaviors or conditions contributing to their risk. As a result, patients with positive FIT results may not understand the factors behind their outcome or ways to reduce future risk. Our primary goal was to develop and internally validate a machine learning-based risk prediction model that not only estimates CRC risk with reasonable performance but also addresses the key limitations of existing models, including interpretability, real-world applicability, and adaptability to different clinical settings.

Our model demonstrated moderate discriminatory power, achieving an AUROC of 0.726. This indicates that the model can moderately distinguish between individuals with and without CRC. This falls within the range reported for existing logistic regression-based models (0.60–0.75), demonstrating comparable predictive capability while incorporating machine learning advantages. We incorporated the following 12 easily obtainable clinical factors into the model: sex, age, weight, height, body mass index (BMI), hypertension status, heart conditions, diabetes status, smoking history, smoking quantity, and alcohol consumption. Unlike many existing CRC risk models that require genetic markers, bio-

chemical tests, or complex dietary assessments, our model was designed for use in routine clinical consultations, where data collection time and resources may be limited.

4.2. Comparison with Conventional Risk Models

To assess the added value of our model over traditional logistic regression-based approaches, we compared its performance to existing CRC risk models. While many logistic regression models have been developed to estimate CRC risk [20,21], our model incorporates advanced ML techniques, improving predictive performance by capturing non-linear relationships and interactions among risk factors.

Previous logistic regression models report AUC values ranging from 0.60 to 0.75 [9], with many relying on family history, genetic risk scores, or laboratory values that are not always practical for real-world implementation. In contrast, our model does not require genetic markers or invasive testing, making it more applicable in routine clinical settings. Additionally, our model demonstrated comparable or superior performance (AUROC 0.726) to traditional logistic regression models, without requiring additional laboratory assessments. Moreover, the use of SHAP interpretability techniques enhances clinical transparency, allowing clinicians to see which factors drive individual risk scores, an advantage over conventional regression models that often provide limited interpretability.

4.3. Risk Stratification and Threshold Selection

A key feature of our predictive model is the generation of a risk score that categorizes patients into average-, increased-, or high-risk groups for CRC. This stratification is based on the probability threshold we selected and enables clinicians to tailor their communication and recommendations according to the patient's risk level. For instance, patients identified as increased or high risk can be counseled on specific lifestyle modifications, the importance of regular screening, or referred for further diagnostic evaluations.

In selecting the probability threshold for categorizing CRC risk, we prioritized sensitivity (0.747) to minimize the likelihood of overlooking individuals at risk. A highly sensitive threshold ensures that fewer at-risk individuals are missed, which is particularly important in early cancer detection. However, an overly sensitive model may increase false positives, leading to unnecessary follow-ups, patient anxiety, and additional healthcare costs. Conversely, a stricter threshold that improves specificity would reduce false positives but might miss individuals who could benefit from early intervention.

To balance these factors, we evaluated multiple threshold levels and selected 0.0082 as the most appropriate cut-off for risk classification. At a threshold of 0.0082, the model achieves a sensitivity of 74.7%, meaning that nearly 75% of patients with CRC risk are correctly identified. While the specificity at this threshold was around 61%, we considered this acceptable for a screening-oriented tool where the emphasis is on maximizing the detection of potential CRC cases. Compared to a lower threshold (e.g., 0.0078), which would have increased sensitivity to 79.8% but reduced specificity to 56%, the selected threshold provides a practical balance between detecting high-risk cases and minimizing false positives. This selection aligns with clinical priorities, where early detection is emphasized over avoiding false positives. In a screening setting, it is generally preferable to flag more at-risk individuals and conduct further testing rather than miss potential CRC cases. Supporting this approach, Osborne et al. found that both patients and healthcare professionals consider gains in true-positive diagnoses for CRC worth the trade-off of increased false positives, with participants willing to accept up to a 45% decrease in specificity to achieve a 10% gain in sensitivity [22].

While 0.0082 is an optimal general-use threshold, it can be adjusted based on clinical priorities. In high-risk populations (e.g., individuals with strong family history, chronic

inflammation, or multiple risk factors), a lower threshold (e.g., 0.0078) could be used to increase sensitivity and identify more patients needing earlier intervention. In settings where resources are limited, a slightly higher threshold could be applied to reduce unnecessary follow-ups, ensuring that only the highest-risk individuals are referred for further testing.

4.4. Interpretability and Feature Importance

To interpret the contributions of individual risk factors, we employed SHAP analysis. This method allowed us to understand how each factor influences CRC risk on a per-patient basis. Our findings indicated that age, weight, and smoking history are the most significant contributors to increased CRC risk, corroborating the existing literature that identifies these factors as important risk determinants [9,23]. Conversely, regular use of medications for heart conditions appeared to have a protective effect, which may be attributed to the potential anti-inflammatory properties of certain cardiovascular drugs or the associated healthier lifestyle behaviors among these patients.

However, evidence on this association remains mixed. Some studies have reported no association between CRC risk or recurrence and the use of common cardiovascular medications like statins and antihypertensives [24,25]. These discrepancies may be due to limitations such as small sample sizes, inadequate power to detect small to moderate associations, and the potential misclassification of medication adherence. Given these conflicting findings, further research with larger cohorts and careful control of confounding factors is warranted to clarify the potential protective effects of heart medications on CRC risk.

Comparing our model to existing predictive tools for CRC risk, we note that many require extensive clinical data, invasive procedures, or specialized laboratory tests, limiting their practicality for widespread use [9,26–30]. For instance, some models incorporate genetic markers or detailed biochemical profiles that are not routinely available in primary care settings [30,31]. In contrast, our model relies on information typically collected during standard healthcare visits, enhancing its feasibility for integration into everyday clinical practice without imposing additional burdens on patients or clinicians.

4.5. How Primary Care Physicians Can Use the Model

To make our CRC risk prediction model practical for routine clinical use, we have integrated it into an interactive risk estimator tool that primary care physicians can use during patient consultations. This tool adapts in real time based on patient responses and provides personalized screening recommendations, making it easier to discuss risk and prevention with patients.

4.5.1. Immediate Referral for High-Risk Patients

If a patient reports a family history of CRC, the tool automatically recommends a colonoscopy, following the established screening guidelines. This ensures that those at the highest genetic risk are fast-tracked for further evaluation.

4.5.2. Assessing Lifestyle Risk Factors

For patients without a family history, the tool continues by evaluating modifiable risk factors, things they can control to lower their CRC risk. These include diet, smoking, alcohol consumption, BMI, and existing conditions like hypertension and diabetes.

4.5.3. Personalized Risk Score and Easy-to-Understand Results

After gathering patient data, the model calculates a personalized risk score and classifies individuals as average, increased, or high risk. To make results easier to understand,

the risk level is visualized using a color-coded gauge, helping both clinicians and patients quickly grasp what the score means.

4.5.4. Providing Practical Steps for Prevention

The model automatically calculates a personalized CRC risk score, with certain factors weighing more heavily than others. For instance, age is a dominant factor in CRC risk assessment, which means that even if a patient engages in unhealthy lifestyle habits, their overall risk score might still appear moderate due to their age-related risk. However, not all risk factors, such as smoking or alcohol consumption, are always reflected directly in the risk gauge. Instead, these behaviors are visually flagged as potential risks, allowing clinicians to assess their significance in a broader clinical context.

4.6. Temporal Limitations

A key limitation to our study is the reliance on an older dataset, the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial (1993–2001), which may not fully capture the evolving risk factors and epidemiological patterns of CRC in contemporary populations. Over the past two decades, CRC incidence has shifted, with an increasing prevalence of early-onset CRC (E-O CRC) in individuals under 50 years old, in particular [32].

Lifestyle changes, including rising obesity rates, altered dietary patterns, increased consumption of ultra-processed foods, sedentary behavior, and changes in gut microbiome composition, have significantly contributed to the growing burden of CRC in younger populations [33,34]. These newer risk factors may not be adequately reflected in older datasets, limiting the generalizability of models trained on historical cohorts.

4.6.1. Changes in Risk Factor Profiles over Time

Several studies have documented changes in modifiable CRC risk factors over time, leading to evolving risk profiles compared to those observed in earlier cohorts. Dietary patterns, obesity rates, and sedentary behavior have increased significantly in recent decades, altering exposure to key lifestyle-based risk factors.

For example, Deng et al. (2023) emphasized the growing role of ultra-processed food consumption, sweetened beverages, and sedentary behavior in CRC risk among younger populations in China (2015–2021) [35], whereas earlier models focused more on total caloric intake and general dietary habits [36,37]. Similarly, the rising prevalence of obesity and metabolic syndrome has introduced new risk pathways that were less prominent in older datasets [38].

Despite these changes, traditional lifestyle factors, such as smoking, diet, and family history, remain the cornerstone of CRC risk prediction, as they are both strong predictors and actionable intervention points. While recent studies have explored polygenic risk scores (PRS) for CRC stratification [7], PRS remains impractical for routine clinical use due to cost, accessibility, and the complexity of genetic counseling [8].

Furthermore, individuals with a strong family history of CRC are already referred for colonoscopy [39,40], making additional genetic risk models redundant in many cases. Thus, while modifiable risk factors have evolved, shifting toward concerns about obesity, processed food consumption, and metabolic changes, CRC prediction models must continue prioritizing clinically accessible lifestyle factors to ensure widespread clinical applicability, facilitate real-time risk assessment during consultations, and provide actionable guidance for preventive interventions.

4.6.2. Evolving Screening Practices and Their Impact on Risk Prediction

The screening landscape has also changed significantly since the PLCO trial, affecting how risk models should be calibrated. When the PLCO trial was conducted, flexible sigmoidoscopy (FSG) was a commonly used screening tool, whereas modern screening practices now emphasize non-invasive tests like FIT and FIT-DNA, along with risk-adapted strategies [41,42]. Models trained on PLCO data may not account for changes in CRC detection rates due to improved screening uptake and more sensitive diagnostic tools.

New guidelines recommend CRC screening initiation at age 45 instead of 50, a crucial adjustment that is not reflected in older datasets [40]. Risk prediction models must be re-calibrated to account for this shift, as applying older risk thresholds could misclassify younger high-risk individuals who would benefit from earlier screening interventions. Given these concerns, external validation using more recent population-based datasets is necessary to confirm the applicability of our model to contemporary clinical settings.

4.7. Strengths and Practical Implications

Our study offers several key strengths that enhance its clinical applicability, interpretability, and usability in real-world settings. First, we employed a rigorous, evidence-based approach to feature selection, using a systematic review to identify clinically relevant CRC risk factors. Unlike models that rely on hard-to-obtain data such as polygenic risk scores or precise dietary intake, our model prioritizes routinely collected variables, improving feasibility for standard consultations. Second, our model is trained on a large, prospectively collected dataset (PLCO Cancer Screening Trial), ensuring longitudinal follow-up and reducing recall bias. The dataset's demographic diversity also enhances generalizability across populations.

Furthermore, a major advantage of our study is the development of a user-friendly risk estimator that can be used by clinicians during consultations. This estimator automatically interprets the model's probabilistic output and categorizes individuals into average-, increased-, or high-risk groups. Unlike conventional risk models that only provide a risk score, our tool generates personalized recommendations based on the patient's specific risk profile. These recommendations include lifestyle modifications, preventive measures, and guidance on future screening strategies, enabling clinicians to offer tailored advice in real time. This functionality supports shared decision making by simplifying complex risk predictions into actionable insights for both clinicians and patients.

5. Conclusions

Our findings show that age, weight, and smoking history are the strongest predictors of colorectal cancer (CRC) risk, while heart medication use appeared to have a potentially protective effect. The model correctly identified high- and average-risk individuals in 73 out of 100 cases, demonstrating its potential to complement existing screening tools like FIT by providing personalized risk assessments and additional insights for CRC prevention. Additionally, we provided multiple threshold options to balance sensitivity and specificity, allowing physicians to adjust the threshold based on clinical priorities. This flexibility helps to detect more high-risk individuals while minimizing unnecessary follow-ups, making the model adaptable to different healthcare settings.

To translate these findings into clinical practice, we developed and internally validated a CRC risk prediction model based on health and lifestyle factors, which has been integrated into an interactive risk estimator for clinical use. This tool categorizes individuals as average, increased, or high risk while also identifying modifiable risk factors to support informed decision making and personalized lifestyle modifications. After clinicians input patient responses, the tool provides risk stratification and relevant health insights,

helping guide prevention strategies. However, before clinical implementation, external validation is necessary to confirm the model's reliability across diverse populations and assess its effectiveness in real-world healthcare settings. Additionally, usability testing in clinical practice is needed to evaluate whether the risk estimator enhances clinician-patient discussions and facilitates shared decision making for CRC prevention.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/gidisord7020026/s1>, Table S1: Key terms and definitions to machine learning in clinical research. Reference [43] are cited in the Supplementary Materials.

Author Contributions: D.J.H.: Conceptualization; Investigation; Methodology; Visualization; Writing—Original Draft; and Writing—Review and Editing. D.M.S.: Conceptualization; Data Curation; Investigation; Formal Analysis; Methodology; Visualization; Writing—Original Draft; and Writing—Review and Editing. K.F.: Conceptualization; Data Curation; Formal Analysis; Methodology; and Writing—Review and Editing. M.v.L.: Conceptualization and Writing—Review and Editing. W.v.d.V.: Conceptualization; Methodology; Funding Acquisition; Supervision; and Writing—Review and Editing. G.V.H.: Conceptualization; Funding Acquisition; Supervision; and Writing—Review and Editing. D.J.H. and D.M.S. contributed equally as co-first authors. W.v.d.V. and G.V.H. contributed equally as co-last authors. All authors have read and agreed to the published version of the manuscript.

Funding: This study is part of the ORIENT (tOwaRds Informed dEcisions iN colorecTal cancer screening) Project, funded by Kom op tegen Kanker (Stand up against Cancer) (ID: 12751). The project focuses on developing and pilot-testing the SDM tool for CRC screening. The funders had no role in the study design, data collection, analysis, interpretation, or writing. All authors had full access to the data and the final decision to submit for publication.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of a de-identified secondary dataset (PLCO), which was accessed through the National Cancer Institute's (NCI) Cancer Data Access System (CDAS) under an approved project (Reference: 3478). Since the dataset does not contain personally identifiable information and is made available for secondary research following ethical and regulatory guidelines, additional Institutional Review Board approval beyond the CDAS/NCI review was not required, in accordance with standard ethical guidelines for anonymized data analysis.

Informed Consent Statement: Patient consent was waived due to the secondary nature of the data.

Data Availability Statement: No new data were created in this study. The data analyzed in this study were obtained from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial dataset, which is publicly available through the National Cancer Institute (NCI) Cancer Data Access System (CDAS). Researchers can request access to the PLCO dataset by submitting a formal application to CDAS. Details on the application process and data access requirements can be found at <https://cdas.cancer.gov/>. Please note that access to the dataset is subject to approval to ensure compliance with data privacy and ethical guidelines.

Acknowledgments: We are immensely thankful to the ORIENT Academic Advisory Committee Members for their guidance and unwavering support throughout this study.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic curve
CRC	Colorectal Cancer
DTA	Data Transfer Agreement
FSG	Flexible Sigmoidoscopy
ML	Machine Learning
PLCO	Prostate, Lung, Colorectal, and Ovarian cancer screening trial
SHAP	SHapley Additive exPlanations

References

1. World Health Organization. Colorectal Cancer. 2023. Available online: <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer#:~:text=Key%20facts,of%20cancer-related%20deaths%20worldwide> (accessed on 21 November 2023).
2. Morgan, E.; Arnold, M.; Gini, A.; Lorenzoni, V.; Cabasag, C.J.; Laversanne, M.; Vignat, J.; Ferlay, J.; Murphy, N.; Bray, F. Global burden of colorectal cancer in 2020 and 2040: Incidence and mortality estimates from GLOBOCAN. *Gut* **2023**, *72*, 338–344. [PubMed]
3. De Keersmaeckere, S. *European Health Union: New Approach on Cancer Screening*; European Commission: Brussels, Belgium, 2022.
4. Cairns, J.M.; Greenley, S.; Bamidele, O.; Weller, D. A scoping review of risk-stratified bowel screening: Current evidence, future directions. *Cancer Causes Control* **2022**, *33*, 653–685. [PubMed]
5. Toumazis, I.; Bastani, M.; Han, S.S.; Plevritis, S.K. Risk-Based lung cancer screening: A systematic review. *Lung Cancer* **2020**, *147*, 154–186. [PubMed]
6. Hull, M.A.; Rees, C.J.; Sharp, L.; Koo, S. A risk-stratified approach to colorectal cancer prevention and diagnosis. Vol. 17, *Nature Reviews Gastroenterology and Hepatology*. *Nat. Res.* **2020**, *17*, 773–780.
7. Briggs, S.E.; Law, P.; East, J.E.; Wordsworth, S.; Dunlop, M.; Houlston, R.; Hippisley-Cox, J.; Tomlinson, I. Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: Population based cohort study. *BMJ* **2022**, *379*, e071707. [CrossRef]
8. Arnau-Collell, C.; Díez-Villanueva, A.; Bellosillo, B.; Augé, J.M.; Muñoz, J.; Guinó, E.; Moreira, L.; Serradesanferm, A.; Pozo, À.; Torà-Rocamora, I.; et al. Evaluating the Potential of Polygenic Risk Score to Improve Colorectal Cancer Screening. *Cancer Epidemiol. Biomark. Prev.* **2022**, *31*, 1305–1312.
9. Herrera, D.J.; van de Veerdonk, W.; Seibert, D.M.; Boke, M.M.; Gutiérrez-Ortiz, C.; Yimer, N.B.; Feyen, K.; Ferrari, A.; Van Hal, G. From Algorithms to Clinical Utility: A Systematic Review of Individualized Risk Prediction Models for Colorectal Cancer. *Gastrointest. Disord.* **2023**, *5*, 549–579. [CrossRef]
10. Hippisley-Cox, J.; Coupland, C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: Prospective cohort study. *BMJ Open* **2015**, *5*, e007825.
11. Sung, J.J.Y.; Wong, M.C.S.; Lam, T.Y.T.; Tsoi, K.K.F.; Chan, V.C.W.; Cheung, W.; Ching, J.Y.L. A modified colorectal screening score for prediction of advanced neoplasia: A prospective study of 5744 subjects. *J. Gastroenterol. Hepatol.* **2018**, *33*, 187–194.
12. Luu, X.Q.; Lee, K.; Kim, J.; Sohn, D.K.; Shin, A.; Choi, K.S. The classification capability of the Asia Pacific Colorectal Screening score in Korea: An analysis of the Cancer Screenee Cohort. *Epidemiol. Health* **2021**, *43*, e2021069.
13. Kaminski, M.F.; Polkowski, M.; Kraszewska, E.; Rupinski, M.; Butruk, E.; Regula, J. A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. *Gut* **2014**, *63*, 1112–1119. [CrossRef] [PubMed]
14. Freedman, A.N.; Slattery, M.L.; Ballard-Barbash, R.; Willis, G.; Cann, B.J.; Pee, D.; Gail, M.H.; Pfeiffer, R.M. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *J. Clin. Oncol.* **2009**, *27*, 686–693. [CrossRef] [PubMed]
15. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1379. [CrossRef]
16. Abdul Rahman, H.; Ottom, M.A.; Dinov, I.D. Machine learning-based colorectal cancer prediction using global dietary data. *BMC Cancer* **2023**, *23*, 144. [CrossRef]
17. National Cancer Institute; National Institute of Health. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. 2015. Available online: <https://cdas.cancer.gov/plco/> (accessed on 12 February 2024).
18. Fouad, K.M.; Ismail, M.M.; Azar, A.T.; Arafa, M.M. Advanced methods for missing values imputation based on similarity learning. *PeerJ Comput. Sci.* **2021**, *7*, e619. [CrossRef]

19. Song, L.L.; Li, Y.M. Current noninvasive tests for colorectal cancer screening: An overview of colorectal cancer screening tests. *World J. Gastrointest. Oncol.* **2016**, *8*, 793–800. [[CrossRef](#)]
20. Cai, Q.-C.; Yu, E.-D.; Xiao, Y.; Bai, W.-Y.; Chen, X.; He, L.-P.; Yang, Y.-X.; Zhou, P.-H.; Jiang, X.-L.; Xu, H.-M.; et al. Derivation and validation of a prediction rule for estimating advanced colorectal neoplasm risk in average-risk chinese. *Am. J. Epidemiol.* **2012**, *175*, 584–593. [[CrossRef](#)]
21. Imperiale, T.F.; Monahan, P.O.; Stump, T.E.; Glowinski, E.A.; Ransohoff, D.F. Derivation and validation of a scoring system to stratify risk for advanced colorectal neoplasia in asymptomatic adults a cross-sectional study. *Ann. Intern. Med.* **2015**, *163*, 339–346. [[CrossRef](#)]
22. Boone, D.; Mallett, S.; Zhu, S.; Yao, G.L.; Bell, N.; Ghanouni, A.; von Wagner, C.; Taylor, S.A.; Altman, D.G.; Lilford, R.; et al. Patients' & healthcare professionals' values regarding true- & false-positive diagnosis when colorectal cancer screening by CT colonography: Discrete choice experiment. *PLoS ONE* **2013**, *8*, e80767.
23. Peng, L.; Weigl, K.; Boakye, D.; Brenner, H. Risk scores for predicting advanced colorectal neoplasia in the average-risk population: A systematic review and meta-analysis. *Am. J. Gastroenterol.* **2018**, *113*, 1788–1800. [[CrossRef](#)]
24. Bowles, E.J.A.; Yu, O.; Ziebell, R.; Chen, L.; Boudreau, D.M.; Ritzwoller, D.P.; Hubbard, R.A.; Boggs, J.M.; Burnett-Hartman, A.N.; Sterrett, A.; et al. Cardiovascular medication use and risks of colon cancer recurrences and additional cancer events: A cohort study. *BMC Cancer* **2019**, *19*, 270.
25. Boudreau, D.M.; Koehler, E.; Rulyak, S.J.; Haneuse, S.; Harrison, R.; Mandelson, M.T. Cardiovascular medication use and risk for colorectal cancer. *Cancer Epidemiol. Biomark. Prev.* **2008**, *17*, 3076–3080. [[CrossRef](#)] [[PubMed](#)]
26. Cooper, J.A.; Ryan, R.; Parsons, N.; Stinton, C.; Marshall, T.; Taylor-Phillips, S. The use of electronic healthcare records for colorectal cancer screening referral decisions and risk prediction model development. *BMC Gastroenterol.* **2020**, *20*, 78.
27. Stegeman, I.; de Wijkerslooth, T.R.; Stoop, E.M.; van Leerdam, M.E.; Dekker, E.; van Ballegooijen, M.; Kuipers, E.J.; Fockens, P.; Kraaijenhagen, R.A.; Bossuyt, P.M. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. *Gut* **2014**, *63*, 466–471. [[CrossRef](#)]
28. Tao, S.; Hoffmeister, M.; Brenner, H. Development and validation of a scoring system to identify individuals at high risk for advanced colorectal neoplasms who should undergo colonoscopy screening. *Clin. Gastroenterol. Hepatol.* **2014**, *12*, 478–485. [[CrossRef](#)]
29. Musselwhite, L.W.; Redding, T.S.; Sims, K.J.; O'leary, M.C.; Hauser, E.R.; Hyslop, T.; Gellad, Z.F.; Sullivan, B.A.; Lieberman, D.; Provenzale, D. Advanced neoplasia in Veterans at screening colonoscopy using the National Cancer Institute Risk Assessment Tool. *BMC Cancer* **2019**, *19*, 1097.
30. Yang, H.; Choi, S.; Park, S.; Jung, Y.S.; Choi, K.Y.; Park, T.; Kim, J.Y.; Park, D.I. Derivation and validation of a risk scoring model to predict advanced colorectal neoplasm in adults of all ages. *J. Gastroenterol. Hepatol.* **2017**, *32*, 1328–1335.
31. van't Klooster, C.C.; Ridker, P.M.; Cook, N.R.; Aerts, J.G.; Westerink, J.; Asselbergs, F.W.; van der Graaf, Y.; Visseren, F.L. Prediction of Lifetime and 10-Year Risk of Cancer in Individual Patients With Established Cardiovascular Disease. *JACC Cardio Oncol.* **2020**, *2*, 400–410.
32. Siegel, R.L.; Giaquinto, A.N.; Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **2024**, *74*, 12–49.
33. Moon, J.Y.; Kye, B.H.; Ko, S.H.; Yoo, R.N. Sulfur Metabolism of the Gut Microbiome and Colorectal Cancer: The Threat to the Younger Generation. *Nutrients* **2023**, *15*, 1966. [[CrossRef](#)]
34. Jardim, S.R.; de Souza, L.M.P.; de Souza, H.S.P. The Rise of Gastrointestinal Cancers as a Global Phenomenon: Unhealthy Behavior or Progress? *Int. J. Environ. Res. Public Health* **2023**, *20*, 3640. [[CrossRef](#)]
35. Deng, J.W.; Zhou, Y.L.; Dai, W.X.; Chen, H.M.; Zhou, C.B.; Zhu, C.Q.; Ma, X.; Pan, S.; Cui, Y.; Xu, J.; et al. Noninvasive predictive models based on lifestyle analysis and risk factors for early-onset colorectal cancer. *J. Gastroenterol. Hepatol.* **2023**, *38*, 1768–1777. [[PubMed](#)]
36. Kim, J.Y.; Choi, S.; Park, T.; Kim, S.K.; Jung, Y.S.; Park, J.H.; Kim, H.J.; Cho, Y.K.; Sohn, C.I.; Jeon, W.K.; et al. Development and validation of a scoring system for advanced colorectal neoplasm in young Korean subjects less than age 50 years. *Intest. Res.* **2019**, *17*, 253–264. [[PubMed](#)]
37. Ma, E.; Sasazuki, S.; Iwasaki, M.; Sawada, N.; Inoue, M. 10-Year risk of colorectal cancer: Development and validation of a prediction model in middle-aged Japanese men. *Cancer Epidemiol.* **2010**, *34*, 534–541. [[PubMed](#)]
38. Zhang, K.; Ma, Y.; Luo, Y.; Song, Y.; Xiong, G.; Ma, Y.; Sun, X.; Kan, C. Metabolic diseases and healthy aging: Identifying environmental and behavioral risk factors and promoting public health. *Front. Public Health* **2023**, *11*, 1253506.
39. Imperiale, T.F.; Monahan, P.O.; Stump, T.E.; Ransohoff, D.F. Derivation and validation of a predictive model for advanced colorectal neoplasia in asymptomatic adults. *Gut* **2021**, *70*, 1155–1161. [[CrossRef](#)]
40. Meester, R.G.S.; van de Schootbrugge-Vandermeer, H.J.; Breekveldt, E.C.H.; de Jonge, L.; Toes-Zoutendijk, E.; Kooyker, A.; Nieboer, D.; Ramakers, C.R.; Spaander, M.C.W.; van Vuuren, A.J.; et al. Faecal occult blood loss accurately predicts future detection of colorectal cancer. A prognostic model. *Gut* **2023**, *72*, 101–108.

41. Thomsen, M.K.; Pedersen, L.; Erichsen, R.; Lash, T.L.; Sørensen, H.T.; Mikkelsen, E.M. Risk-stratified selection to colonoscopy in FIT colorectal cancer screening: Development and temporal validation of a prediction model. *Br. J. Cancer* **2022**, *126*, 1229–1235.
42. Ferrari, A.; Neefs, I.; Hoeck, S.; Peeters, M.; Van Hal, G. Towards Novel Non-Invasive Colorectal Cancer Screening Methods: A Comprehensive Review. *Cancers* **2021**, *13*, 1820. [[CrossRef](#)]
43. Weissler, E.H.; Naumann, T.; Andersson, T.; Ranganath, R.; Elemento, O.; Luo, Y.; Freitag, D.F.; Benoit, J.; Hughes, M.C.; Khan, F.; et al. The role of machine learning in clinical research: Transforming the future of evidence generation. *Trials* **2021**, *22*, 537. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.