Digestive Diseases and Sciences
https://doi.org/10.1007/s10620-024-08427-4

ORIGINAL ARTICLE

Development and Validation of a Colorectal Cancer Prediction Model: A Nationwide Cohort-Based Study

Ofer Isakov[1,2,3] · Dan Riesel[1] · Michael Leshchinsky[1] · Galit Shaham[1] · Ben Y. Reis[2,3,11,12] · Dan Keret[4] · Zohar Levi[5] · Baruch Brener[6,7] · Ran Balicer[1,3,8] · Noa Dagan[1,3,9] · Samah Hayek[1,10]

Received: 15 January 2024 / Accepted: 5 April 2024

Abstract

Background Early diagnosis of colorectal cancer (CRC) is critical to increasing survival rates. Computerized risk prediction models hold great promise for identifying individuals at high risk for CRC. In order to utilize such models effectively in a population-wide screening setting, development and validation should be based on cohorts that are similar to the target population.

Aim Establish a risk prediction model for CRC diagnosis based on electronic health records (EHR) from subjects eligible for CRC screening.

Methods A retrospective cohort study utilizing the EHR data of Clalit Health Services (CHS). The study includes CHS members aged 50–74 who were eligible for CRC screening from January 2013 to January 2019. The model was trained to predict receiving a CRC diagnosis within 2 years of the index date. Approximately 20,000 EHR demographic and clinical features were considered.

Results The study includes 2935 subjects with CRC diagnosis, and 1,133,457 subjects without CRC diagnosis. Incidence values of CRC among subjects in the top 1% risk scores were higher than baseline (2.3% vs 0.3%; lift 8.38; P value < 0.001). Cumulative event probabilities increased with higher model scores. Model-based risk stratification among subjects with a positive FOBT, identified subjects with more than twice the risk for CRC compared to FOBT alone.

Conclusions We developed an individualized risk prediction model for CRC that can be utilized as a complementary decision support tool for healthcare providers to precisely identify subjects at high risk for CRC and refer them for confirmatory testing.

Keywords Colorectal cancer screening · Colorectal cancer · Machine learning

✉ Samah Hayek
Samahhayek@tauex.tau.ac.il

[1] Innovation Division, Clalit Research Institute, Clalit Health Services, Tel Aviv, Israel

[2] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[3] The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute, Boston, MA, USA

[4] Gastroenterology and Hepatology Department, Clalit Health Services, Jerusalem, Israel

[5] Department of Gastroenterology, Beilinson Medical Center, Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

[6] Institute of Oncology, Davidoff Cancer Center, Rabin Medical Center, Beilinson Campus, Petah Tikva, Israel

[7] Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

[8] School of Public Health, Faculty of Health Sciences, Ben Gurion University of the Negev, Be'er Sheva, Israel

[9] Software and Information Systems Engineering, Ben Gurion University of the Negev, Be'er Sheva, Israel

[10] Department of Epidemiology and Preventive Medicine, School of Public Health, Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

[11] Predictive Medicine Group, Boston Children's Hospital, Boston, MA, USA

[12] Harvard Medical School, Boston, MA, USA

Published online: 25 April 2024

Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide [1] and one of the leading causes of cancer-related mortality in Europe and the United States [2]. Early diagnosis of CRC is crucial to enhancing the success of treatment approaches, increasing the survival rate and improving quality of life [3]. A number of randomized clinical trials have shown that screening for CRC is effective in reducing CRC incidence and related mortality [4, 5]. A recent multinational pragmatic randomized trial evaluated the 10-year impact of invited screening colonoscopy and found a significant reduction in CRC incidence [5]. Although there is evidence that CRC screening is beneficial, compliance and adherence to CRC screening remains low [6, 7]. During the past decade, several models have been developed for the purpose of CRC risk assessment and diagnosis. These models utilize various types of features, including patient demographics, anthropometric data, lifestyle characteristics, diagnoses, lab results, imaging, genetics and microbiome data [8–13]. However, these models have their own limitations. For example, some models were trained and tested on cohorts that include individuals without indication for CRC screening such as individuals older than 75 years, individuals with previous diagnosis of CRC, individuals undergoing workup for CRC diagnosis and individuals with recent positive screening tests [10, 14]. Furthermore, one model required the subjects to have a blood sample record within a few months prior to the diagnosis, and excluded subjects without such records [9]. Other models utilized features that are not readily available in the EHR such as microbiome, genetic, lifestyle and diet information, some of which required active patient participation by filling questionnaires, limiting their application in a population-wide screening setting [8, 11, 12]. Due to these limitations, it is difficult to predict the performance of these tools in a real-life screening setting.

As the purpose of such risk prediction models is to complement current screening methods, we hypothesize that training and evaluating such models on a cohort of subjects that closely resembles the population indicated for screening will enable a more accurate assessment of performance in a real-world setting and improve generalizability and utility. The aim of the current study was to establish an individualized CRC risk prediction model that complements current screening strategies by harnessing readily available comprehensive EHR data corresponding to a carefully selected population of subjects eligible for CRC screening.

Methods

Setting and Study Population

We conducted a retrospective cohort study utilizing the Clalit Health Services (CHS) EHR database to develop a CRC risk prediction model. CHS is the largest integrated payer-provider healthcare organization in Israel. CHS has a comprehensive health care data warehouse. Membership turnover within CHS is 1–2% annually, facilitating long-term follow-up and the ability to capture temporal trends within the data [15]. This study includes all CHS members who were 50–74 years old during the study period. The model was trained and validated on members who met the study eligibility criteria for CRC screening (did not undergo colonoscopy in the past 5 years or fecal occult blood test (FOBT) in the past 2 years) as of four index dates: January 1, 2013, January 1, 2015, January 1, 2017 and January 1, 2019. To be included in the cohort for each index date, members were also required to have at least 1 year of continuous CHS membership prior to that date. Individuals eligible during several index dates were included several times. Individuals matching any of the predefined exclusion criteria were excluded from model development and validation (Methods supplement 1).

Outcome Definition

Subjects were considered positive for the outcome if they had a diagnosis of CRC within 4–24 months after the index date (Methods supplement 1).

Model Inputs and Development

For each participant in the cohort, predictor features were extracted from the CHS EHR database up to 3 years prior to the index date. Extracted features included: demographic information, medical conditions, hospitalizations, medications, and labs. For each lab value we included the last recorded value and aggregated metrics.

The model was developed using a training set composed of the 2013, 2015 and 2017 cohorts and its performance was evaluated on a validation set composed of the 2019 cohort. The training set was further down-sampled in order to control for class imbalance. Feature selection was carried out to identify the top 50 features with the highest impact on the model during training (Methods supplement 1).

Model Performance Evaluation

Model discrimination was assessed using the area under the receiver operating curve (AUROC), with further metrics like sensitivity, specificity, positive predictive value (PPV) and lift evaluated at specified risk percentile thresholds. A cumulative incidence analysis over 4 years was carried out to appraise long-term risk identification for CRC. Performance measures were also examined for a subset of individuals who underwent FOBT for each combination of FOBT result and predicted risk (Methods supplement 1).

Feature Importance Analysis

Feature importance and feature risk contribution were evaluated using SHAP (SHapley Additive exPlanations) values [16]. To demonstrate the impact of the last values of specific lab features and the interaction with their trajectory on CRC predicted risk, partial dependence plots (PDP) were stratified by slope and plotted for the lab features deemed important by the model.

Statistical Analysis

Deviations between the incidence of CRC cases in each model, score risk percentile and baseline incidence along with corresponding confidence intervals were calculated using a two-sided binomial test. The cumulative incidence curves were compared using the log-rank test. 95% confidence intervals were calculated using the survminer R package. Death rates and CRC proportions were compared using the chi-squared test. Python (version 3.8.8) and the scikit-learn package were used for machine learning modeling [17]. 95% confidence intervals (CIs) for the AUROC were calculated using the DeLong method [18]. Statistical analyses were performed using R statistical software (version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria).

Results

Study Participants

The initial cohort included 3,571,164 individuals aged 50–74, out of which 2,157,192 (60.4%) had undergone an FOBT or a screening colonoscopy within 2 or 5 years prior to the index date, respectively and therefore were excluded from the model training and validation cohorts. Two lakhs sixty-two thousand hundred and seventy (7.3%) individuals were excluded as they were unlikely to benefit from screening or at a high risk for complications during an endoscopic evaluation (Supp. Figure 1). The model training and validation cohorts following index date-based sub-sampling included 867,588 subjects, out of which 2,196 (0.3%) were positive cases, and 268,804 subjects out of which 739 (0.3%) were positive cases, respectively. Exclusion criteria were applied to target a cohort most likely to benefit from CRC screening. Subjects who were included displayed lower all-cause mortality rates (6.6% vs. 11.7%; *P* value < 0.001) and a longer median time until death (6.2 years vs. 5.5 years) among those who did not survive throughout the study period compared to excluded subjects. Demographic features and risk factors for CRC did not demonstrate a significant difference between the training and validation sets (Supp Table 1). Among subjects diagnosed with CRC within 2 years of the index date, the median time until diagnosis was 413 and 408 days in the training and validation sets, respectively. The median [interquartile range: IQR] age was 64 [58, 69] and 59 [54, 65] years among those who experienced an event and those that did not, respectively. Slight differences were demonstrated in the lab values between the two groups. Labs corresponding to iron stores demonstrated a more notable difference, with median iron levels of 71 [52, 94] and 80 [62, 102] and median ferritin levels of 58 [24, 113] and 74 [39, 132] in the group that experienced the event compared to the rest of the cohort (Table 1).

Model Performance

The model demonstrated an AUROC of 0.672 on the validation set [95% CI 0.651–0.692]. Incidence values at the top score percentiles (corresponding to the PPV) were significantly higher compared to the baseline incidence, with an incidence of 2.3% (lift 8.38; *P* value = 9.3e−36), 1.09% (lift 3.95; *P* value = 4.75e−42), and 0.82% (lift 2.98; *P* value = 4.5e−43) at the top 1%, 5% and 10% risk percentile (Table 2). The incidence at the bottom score percentiles (i.e. bottom 10%) was significantly lower than the baseline incidence (0.01%; lift = 0.392; *P* value = 4.4e−9) (Supp. Figure 2A). Stratifying model performance by age demonstrated a consistently higher incidence of CRC in the top risk percentile across age groups. Sex-based stratification demonstrated consistent performance across both sex groups (Supp. Figure 2B).

Cumulative Incidence Analysis

Among subjects in the validation cohort, the median follow-up time was 48 months. Among subjects in the model's top risk percentiles, the cumulative incidence of CRC increased with higher model scores, supporting an association between the model's risk prediction and the time until CRC diagnosis (Fig. 1). Among subjects in the top 1% risk percentile the cumulative incidence

Table 1 Descriptive statistics of the study population

Characteristic	Missing (%)	CRC, *N* = 2935[a]	Control, *N* = 1,133,457[a]
Age (years)	< 0.1	64 (58, 69)	59 (54,65)
Gender	< 0.1		
Female		1372 (47%)	582,874 (51%)
Male		1563 (53%)	550,578 (49%)
Hemoglobin (g/dL)	18	13.60 (12.60, 14.60)	13.80 (12.90, 14.80)
Hematocrit (%)	18	41.5 (38.8, 44.3)	42.0 (39.5, 44.7)
RDW (%)	22	13.70 (13.20, 14.40)	13.50 (13.00, 14.10)
Platelets (k/μL)	18	246 (207, 294)	241 (204, 284)
Glucose (mg/dL)	18	100 (91, 117)	97 (89, 110)
Iron (μg/dL)	64	71 (52, 94)	80 (62, 102)
Ferritin (ng/mL)	71	58 (24, 113)	74 (39, 132)
GI bleed	0	165 (5.6%)	24,469 (2.2%)
Smoking (Pack years)	66	33 (18, 46)	33 (18, 45)
BMI (kg/m^2)	13	28.0 (25.1, 31.6)	27.3 (24.4, 30.9)
Prior malignancy	0	265 (9.0%)	62, 369 (5.5%)

[a] Median (25%, 75%); *n* (%)

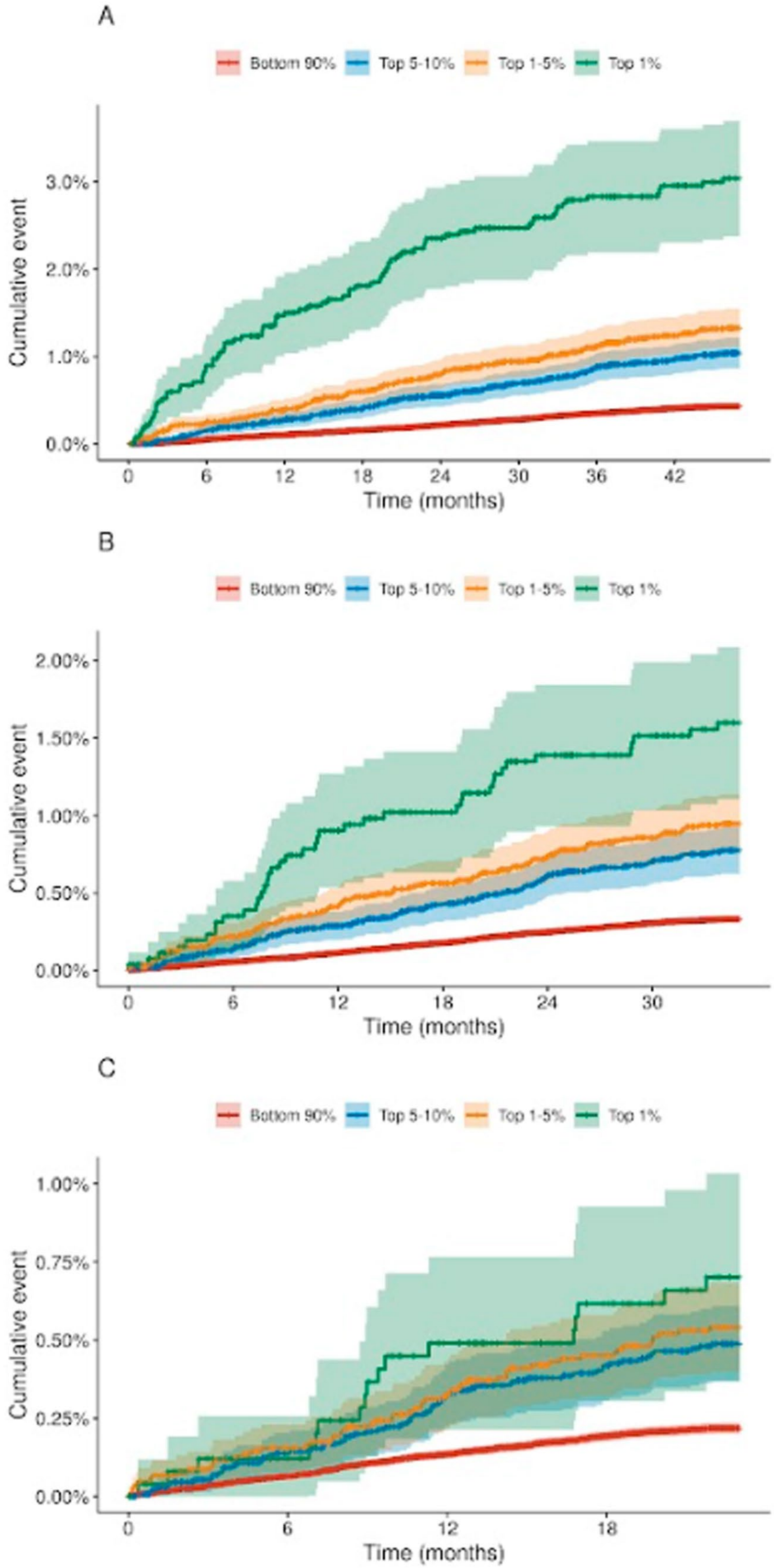
Table 2 Characteristics of participant by their risk score percentile

Characteristic	Risk percentile			
	Bottom 90%, *N* = 241,923[a]	Top 10%, *N* = 13,444[a]	Top 5%, *N* = 10,748[a]	Top 1%, *N* = 2689[a]
Age	59 (54, 64)	69 (66, 72)	70 (67, 73)	70 (66, 73)
MCH sd	0.44 (0.28, 0.64)	0.55 (0.38, 0.76)	0.61 (0.42, 0.86)	0.93 (0.65, 1.36)
Ferritin last value	78 (42, 138)	60 (29, 112)	44 (20, 91)	18 (10, 36)
ALT slope	0.000 (– 0.007, 0.006)	– 0.002 (– 0.008, 0.002)	– 0.003 (– 0.009, 0.001)	– 0.004 (– 0.009, 0.000)
MCH velocity	0.0001 (– 0.0007, 0.0010)	– 0.0003 (– 0.0013, 0.0008)	– 0.0008 (– 0.0019, 0.0004)	– 0.0023 (– 0.0038, – 0.0011)
HCT slope	0.001 (– 0.002, 0.003)	0.000 (– 0.003, 0.002)	– 0.001 (– 0.004, 0.001)	– 0.003 (– 0.006, 0.000)
Gender				
F	127,476 (53%)	5339 (40%)	4253 (40%)	1175 (44%)
M	114,447 (47%)	8105 (60%)	6495 (60%)	1514 (56%)
BMI	27.0 (24.1, 30.5)	28.6 (25.9, 32.2)	28.8 (26.2, 32.5)	29.3 (26.2, 33.0)
MCH slope	0.0001 (– 0.0006, 0.0009)	– 0.0002 (– 0.0012, 0.0008)	– 0.0007 (– 0.0018, 0.0004)	– 0.0023 (– 0.0038, – 0.0011)
Past malignancy	11,146 (4.6%)	2061 (15%)	2353 (22%)	621 (23%)
MCV slope	0.001 (– 0.001, 0.004)	0.000 (– 0.003, 0.003)	– 0.001 (– 0.004, 0.002)	– 0.005 (– 0.009, – 0.001)
Neutrophils min	3.50 (2.77, 4.32)	3.90 (3.20, 4.70)	3.93 (3.30, 4.72)	3.98 (3.30, 4.77)
Iron last value	83 (65, 104)	72 (56, 92)	66 (48, 85)	45 (31, 66)
RDW last value	13.40 (12.90, 13.90)	13.80 (13.30, 14.40)	14.10 (13.50, 14.70)	14.80 (14.00, 16.00)
RDW mean	13.40 (12.95, 13.90)	13.78 (13.32, 14.30)	13.97 (13.50, 14.55)	14.47 (13.87, 15.35)
PLT velocity	0.00 (– 0.03, 0.03)	0.01 (– 0.02, 0.04)	0.02 (– 0.01, 0.05)	0.03 (0.00, 0.08)
Past GI bleed	3936 (1.6%)	644 (4.8%)	921 (8.6%)	375 (14%)
ALT last value	19 (15, 26)	17 (13, 22)	16 (12, 21)	14 (11, 19)
HGB slope	0.0001 (– 0.0006, 0.0008)	– 0.0002 (– 0.0010, 0.0004)	– 0.0004 (– 0.0013, 0.0003)	– 0.0013 (– 0.0023, –0.0004)
CRC diagnosis	519 (0.2%)	74 (0.6%)	84 (0.8%)	62 (2.3%)

[a] Median (IQR); *n* (%)

of CRC was significantly higher compared to the bottom 90% (*P* value = 2.4e−87) and increased over time from 1.5% (95% CI 1.04–1.96%) in the first year to 3.04% (95% CI 2.38–3.7%) by the end of the fourth year. To assess whether the difference in risk remains significant throughout the follow-up period (i.e. the model has long-term predictive ability beyond the 2-year outcome period used for its training), cumulative incidence was compared for

Fig. 1 Cumulative incidence of CRC during the follow-up by risk percentiles. **A** Cumulative incidence of CRC throughout the follow-up period demonstrating a higher cumulative incidence in the top risk percentiles. Cumulative incidence was also calculated for all subjects who survived without a CRC diagnosis after 1 (**B**) and 2 (**C**) years



subjects that were not diagnosed with CRC and survived for 1 and 2 years following the index date (Fig. 1).

Risk Stratification Among Those That Performed Screening

Subjects who underwent FOBT within 2 years prior to the index date ($N = 1{,}524{,}738$) were excluded from the cohort regardless of the FOBT result. The incidence of CRC diagnosis within 2 years following FOBT in this cohort was slightly higher compared to the incidence in the validation cohort (0.310% vs 0.275%; *P* value = 0.003). Evaluation of the utility of the model as a decision support tool among those that did perform screening demonstrates predictive ability for both FOBT-positive and FOBT-negative individuals (Fig. 2a). Among FOBT-positive individuals the model can further stratify the risk, with the risk being more than two times higher after 2 years of follow-up among those who

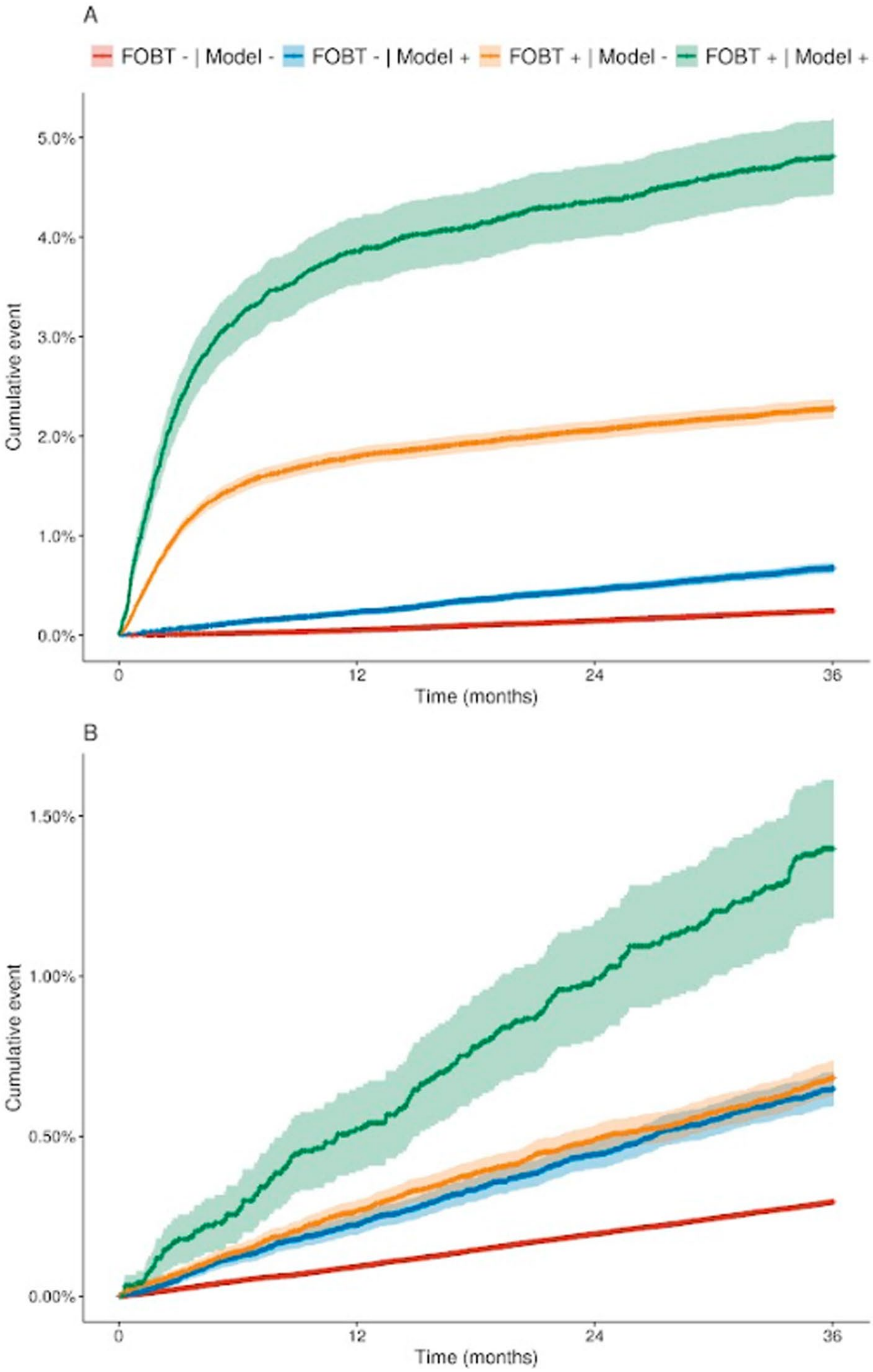


Fig. 2 Cumulative incidence of CRC by risk class and FOBT combination. **A** Three-year cumulative incidence of CRC stratified by FOBT result and model risk class (±). Risk class was defined by setting a risk score threshold resulting in the same rate of subjects predicted positive as the rate of positive FOBT in the cohort. **B** Cumulative incidence was also calculated for all subjects who survived without a CRC diagnosis after 1 year

were also tagged as high-risk by the model (4.32% [95% CI 4.0–4.7%] vs 2.1% [95% CI 2.0–2.15%]). Among subjects with a negative FOBT, CRC incidence was more than three times higher among those tagged as high-risk by the model (0.46% [95% CI 0.41–0.5%] vs 0.15% [95% CI 0.14–0.15%] for those not tagged as high-risk). Moreover, among subjects who were not diagnosed with CRC after 1 year of follow-up, the cumulative incidence during a 3 year follow-up period for those with a negative FOBT and a positive risk prediction was comparable to those with a positive FOBT and a negative risk prediction (0.65% vs 0.68%; *P* value = 0.37) (Fig. 2b).

Features Contributing to Model Performance

Features that were deemed important by the model showed clear trends when stratified according to risk percentiles (Supp. Table 2). The SHAP-based analysis identified the most important features to be age, gender and BMI, with a higher predicted risk for individuals with older age, male gender and higher BMI. These features were followed mostly by laboratory tests and previous malignancy-related diagnoses (Supp. Figure 3). Interestingly, numerous lab features that were found to be predictive of CRC reflected the dynamics of complete blood count and chemistry values over time (the slope and velocity of the collected lab values over time).

Examining the predicted risk of CRC as a function of lab values stratified by the slope over the follow-up period, revealed various interaction patterns between the last value of the lab test and its dynamics over time (Fig. 3). For lab results such as alanine transaminase (ALT) and platelets (PLT), both the last result and the dynamics over time demonstrated predictive ability, but no major interaction was noted between the two features. For HGB and HCT, lower last values were generally associated with increased risk, but this effect was much more pronounced with the presence of a negative slope over time, demonstrating an important interaction between the two features. Lastly, for MCH and MCV it seems that the last value was not predictive at all with the presence of a positive slope overtime, whereas a lower last value was highly predictive of the risk in the presence of a negative slope. In the validation cohort, the interaction between lab values and slopes was significantly associated

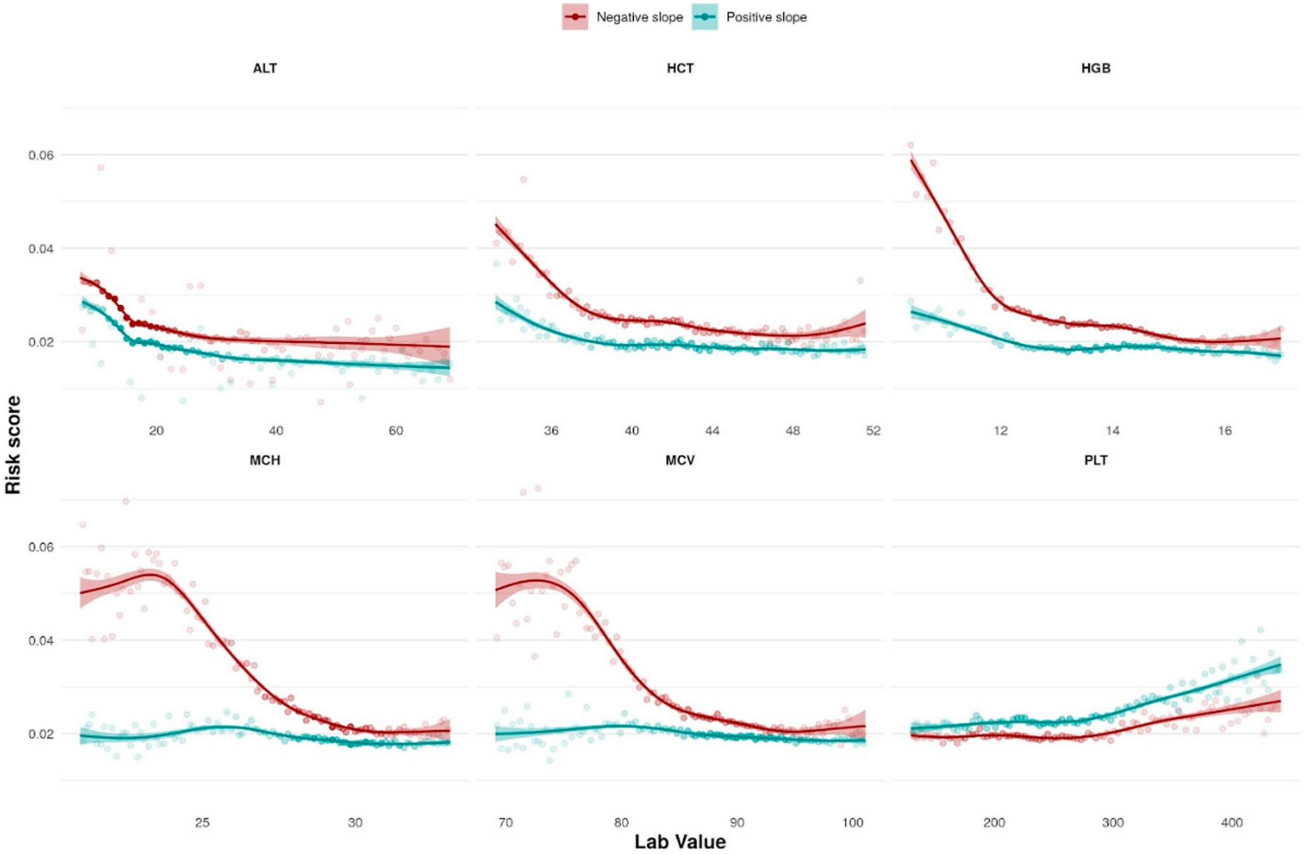


Fig. 3 Association between lab values and slopes and CRC risk. Risk scores are plotted vs. last lab values for the top six labs selected by the model as important discriminatory features. Scores are stratified by whether the calculated slope of lab values was negative (red) or positive (blue) during the 3 years prior to the index date

with the actual risk of CRC diagnosis for all the labs tested ($P < 0.001$).

Discussion

In this study, we describe the development and validation of a CRC risk prediction model based on EHR clinical and laboratory parameters. Our model, which was trained on one of the largest datasets to date, explored the predictive ability of thousands of features and utilized the data from over half a million subjects[14]. Performance was evaluated using two distinct validation cohorts: We demonstrated the model's high discrimination ability within a cohort of subjects that have not undergone CRC screening, noting the model's utility as a safety net for identifying high-risk individuals among those with low adherence to screening. We further demonstrated the discrimination ability of the model among subjects that underwent FOBT screening, noting the model's ability to further assist in decisions regarding those who underwent screening. Specifically, within the cohort of subjects with a negative FOBT, the model was able to pinpoint individuals whose CRC risk was comparable to those with a positive FOBT.

Despite increases in CRC screening rates over the past decade, the absolute rate remains suboptimal [7]. The ongoing lack of compliance can be attributed to patients' low awareness of screenings, fear of screening procedure—particularly colonoscopy, and general lack of communication with the physician [19]. Utilizing an EHR-based classification model for CRC identification could potentially improve awareness by providing physicians with a method for communicating risk to patients that need to undergo screening.

Across the entire validation cohort, in order to identify 10 CRC cases, 3,636 individuals would require a diagnostic colonoscopy. By stratifying the risk among these individuals and selecting the top 1% risk percentile, only 435 individuals would have to undergo a diagnostic colonoscopy, in order to identify the same amount of CRC cases. By screening the top 1.3% risk percentile, corresponding to 3,521 individuals, 10% of CRC cases within the validation cohort could be identified. This is markedly more efficient when compared with a non-stratified approach, which would require screening 26,881 individuals to achieve the same detection rate. A crucial consideration is that all of the patients in our cohort are already recommended for colonoscopy based on existing medical guidelines. Thus, our approach is designed to prioritize such patients without resulting in additional burden or potentially harmful practices.

Features that had the strongest impact on the model included characteristics such as age, gender and BMI. These identified features are consistent with the current literature regarding risk factors for CRC [20]. As expected, lab values characteristic of iron deficiency and anemia had a strong impact on the model. In addition to these, less obvious lab values—such as a decrease in ALT and aspartate transaminase (AST) values and higher glucose, alkaline phosphatase (ALKP) and triglyceride levels were also shown to increase the risk for CRC according to the model. Interestingly, while these associations are less commonly known, they have all been described in the medical literature [21–23].

Our model stands out because it was specifically designed to enhance existing screening approaches. Unlike many models developed over the past decade, which often relied on cohorts with varied indications or required data not commonly found in EHRs, our model was developed using a cohort of at-risk individuals eligible for screening colonoscopy. By focusing on this particular demographic, we believe our model offers enhanced accuracy and generalizability in real-world clinical settings, making it a valuable complement to existing screening strategies. Furthermore, CHS covers more than 50% of the Israeli population and therefore includes subjects from various ethnic backgrounds, providing a representative nation-wide cohort. It is therefore less likely that ethnic biases and healthcare inequalities would have a significant effect on model development [24].

A major strength of our model is the utilization of longitudinal follow-up data. Various features corresponding to the trajectory of laboratory value changes over time (e.g. slope and velocity), were selected by the model as impactful. Such features better reflect the evolving nature of the disease and the patient's health status compared to a single measurement in time. While a single data point could provide a snapshot of a patient's condition, it is incapable of capturing the inherent variability and changes over time, which are critical to understanding disease progression and risk prediction. A longitudinal follow-up approach, on the other hand, allows us to identify how such changes in certain lab values correspond to the onset or progression of CRC. In our study, we analyzed the interaction between the last recorded values and the slope of selected lab features. We showed that while both the last value and the slope contribute to the predictive capabilities of the model, for some features such as MCV and MCH, the interaction between the two uncovers discriminatory signals that would otherwise be missed.

This study has several potential limitations. First, a follow-up period of 48 months may not be sufficient for the purpose of CRC risk assessment, especially for slower-progressing forms of the disease which might have been present at the index date but were not identified throughout the follow-up period. Therefore our model's long-term accuracy beyond this period remains uncertain, and longer follow-up times are necessary to better assess its predictive capability over time. Furthermore, while the study accounts for a range of demographic and clinical features, it's reliance on electronic health records may be subject to information bias, including inaccuracies in coding, data entry errors, or missing data. There may also be unmeasured confounders or risk factors such as specific biomarkers and genetic factors not included in the model that could affect CRC risk. Finally, the applicability of the model in clinical practice also presents challenges as integration of predictive models into routine clinical workflows requires consideration of practical aspects such as healthcare provider training, patient acceptance, and system-level adaptations.

In conclusion, we developed a CRC risk stratification model that improves risk stratification both among subjects that did not undergo recommended screening and among those that underwent screening using an FOBT. This model leveraged information from one of the largest patient populations used for CRC risk evaluation to date and uses commonly available EHR-based features that allows for automatic risk evaluation on entire patient populations. Employing this model holds great potential to enhance the precision of CRC risk stratification, identify high-risk individuals who might be missed by conventional screening methods, and optimize the use of healthcare resources.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10620-024-08427-4.

Acknowledgments This study was supported by the Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute.

Author's contribution Ofer Isakov: Conceptualization; writing—original draft; formal analysis; methodology; visualization; Dan Riesel: Conceptualization, methodology, software; Michael Leshchinsky: Software, Galit Shaham: Conceptualization; software, Ben Y. Reis: Writing—review and editing; Dan Keret: Writing—review and editing, Zohar Levi: Conceptualization; writing—review and editing; Baruch Brener: Conceptualization; writing—review and editing, Ran Balicer: Writing—review and editing; Noa Dagan: Conceptualization; writing—original draft, methodology; Samah Hayek: Conceptualization; writing—original draft, formal analysis; methodology; visualization.

Funding Open access funding provided by Tel Aviv University.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Ethical approval The study protocol was approved by the CHS Institutional Review Board. As this was a retrospective study using de-identified patient data, the requirement for informed consent was waived.

References

1. Cancer (IARC) TIA for R on. Global Cancer Observatory [Internet]. [cited 2023 Mar 20]. Available from: https://gco.iarc.fr/.
2. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Translational Oncology [Internet]. Neoplasia Press; 2021 [cited 2023 Mar 20];14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8273208/.
3. US Preventive Services Task Force, Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB et al. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. *JAMA* 2021;325:1965–77.
4. Atkin WS, Edwards R, Kralj-Hans I, Wooldrage K, Hart AR, Northover JMA et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet.* 2010;375:1624–1633.
5. Bretthauer M, Løberg M, Wieszczy P, Kalager M, Emilsson L, Garborg K et al. Effect of colonoscopy screening on risks of colorectal cancer and related death. *N Engl J Med* 2022;387:1547–56.
6. Levin B, Lieberman DA, McFarland B, Smith RA, Brooks D, Andrews KS et al. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA Cancer J Clin.* 2008;58:130–60.
7. Fisher DA, Princic N, Miller-Wilson L-A, Wilson K, Fendrick AM, Limburg P. Utilization of a colorectal cancer screening test among individuals with average risk. *JAMA Network Open.* 2021;4:e2122269.
8. Aleksandrova K, Reichmann R, Kaaks R, Jenab M, Bueno-de-Mesquita HB, Dahm CC et al. Development and validation of a lifestyle-based model for colorectal cancer risk prediction: the LiFeCRC score. *BMC Med.* 2021;19:1.
9. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *J Am Med Inform Assoc.* 2016;23:879–890.
10. Lee E, Jung SY, Hwang HJ, Jung J. Patient-level cancer prediction models from a nationwide patient cohort: model development and validation. *JMIR Med Inform* 2021;9:e29807-08.
11. Xu W, Mesa-Eguiagaray I, Kirkpatrick T, Devlin J, Brogan S, Turner P et al. Development and validation of risk prediction models for colorectal cancer in patients with symptoms. *J Pers Med* 2023;13:1065.
12. Yang J, McDowell A, Kim EK, Seo H, Lee WH, Moon C-M et al. Development of a colorectal cancer diagnostic model and dietary risk assessment through gut microbiome analysis. *Exp Mol Med.* 2019;51:1–15.
13. Liang H, Yang L, Tao L, Shi L, Yang W, Bai J et al. Data mining-based model and risk prediction of colorectal cancer by using secondary health data: a systematic review. *Chin J Cancer Res.* 2020;32:242–251.

14. Burnett B, Zhou S-M, Brophy S, Davies P, Ellis P, Kennedy J et al. Machine learning in colorectal cancer risk prediction from routinely collected data: a review. *Diagnostics (Basel).* 2023;13:301.
15. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA et al. BNT162b2 mRNA covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine* 2021;384:1412–23.
16. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2:56–67.
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research.* 2011;12:2825–2830.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837–845.
19. Jones RM, Devers KJ, Kuzel AJ, Woolf SH. Patient-reported barriers to colorectal cancer screening: a mixed-methods analysis. *Am J Prev Med.* 2010;38:508–516.
20. Sawicki T, Ruszkowska M, Danielewicz A, Niedźwiedzka E, Arłukowicz T, Przybyłowicz KE. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers (Basel).* 2021;13:2025.
21. He M, Fang Z, Hang D, Wang F, Polychronidis G, Wang L et al. Circulating liver function markers and colorectal cancer risk: A prospective cohort study in the UK Biobank. *International Journal of Cancer* 2021;148:1867.
22. Vulcan A, Manjer J, Ohlsson B. High blood glucose levels are associated with higher risk of colon cancer in men: a cohort study. *BMC Cancer.* 2017;17:842.
23. Yang Z, Tang H, Lu S, Sun X, Rao B. Relationship between serum lipid level and colorectal cancer: a systemic review and meta-analysis. *BMJ Open* 2022;12:e052373.
24. Ameen S, Wong M-C, Yee K-C, Turner P. AI and clinical decision making: the limitations and risks of computational reductionism in bowel cancer screening. *Appl Sci.* 2022;12:3341–45.