BMJ Open Gastroenterology

Derivation and validation of lifestyle-based and microbiota-based models for colorectal adenoma risk evaluation and self-prediction

Yi-Lu Zhou, Jia-Wen Deng, Zhu-Hui Liu, Xin-Yue Ma, Chun-Qi Zhu, Yuan-Hong Xie, Cheng-Bei Zhou, Jing-Yuan Fang

To cite: Zhou Y-L, Deng J-W, Liu Z-H, *et al*. Derivation and validation of lifestyle-based and microbiota-based models for colorectal adenoma risk evaluation and self-prediction. *BMJ Open Gastroenterol* 2025;**12**:e001597. doi:10.1136/bmjgast-2024-001597

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjgast-2024-001597).

Y-LZ and J-WD contributed equally.

Y-LZ are joint first authors. J-WD are joint first authors.

Received 26 September 2024
Accepted 11 March 2025

Division of Gastroenterology and Hepatology, NHC Key Laboratory of Digestive Diseases, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

Correspondence to
Dr Jing-Yuan Fang;
jingyuanfang@sjtu.edu.cn

Dr Cheng-Bei Zhou;
helenairezhou@126.com

ABSTRACT

Objective Early warning and screening of colorectal adenoma (CRA) is important for colorectal cancer (CRC) prevention. This study aimed to construct a non-invasive prediction model to improve CRA screening efficacy.

Methods This study incorporated three cohorts, comprising 9747 participants who underwent colonoscopy. In cohort 1, 683 participants were prospectively recruited with comprehensive lifestyle information and faecal samples. CRA-associated bacteria were identified through 16S rRNA sequencing and quantitative real-time PCR. CRA prediction models were established using lifestyle and gut microbiota information. Cohort 2 prospectively enrolled 1529 participants to validate the lifestyle-based model, while cohort 3 retrospectively analysed 7535 individuals to determine the recommended initial colonoscopy screening ages for different risk groups based on age-specific CRA incidence rates.

Results Multivariable logistic regression yielded a prediction model incorporating 14 variables, demonstrating robust discrimination (c-statistic=0.79, 95% CI 0.75, 0.82). Other machine learning approaches showed comparable performance (random forest: 0.78, 95% CI 0.73, 0.81; gradient boosting: 0.78, 95% CI 0.76, 0.83). The ages for starting colonoscopy screening were established at 42 years for the high-risk group vs 53 years for the low-risk group. The inclusion of *Fusobacterium nucleatum* and *pks*$^{+}$ *Escherichia coli* enhanced the model's performance (c-statistic=0.84–0.86).

Conclusion Integrated mathematical modelling incorporating lifestyle parameters and gut microbial signatures provides an effective non-invasive strategy for CRA risk stratification, while the accompanying machine learning-assisted prediction application enables cost-effective, population-level screening implementation to optimise CRC prevention protocols.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Lifestyle habits and gut microbiota can be closely related to the occurrence of colorectal adenomas.

WHAT THIS STUDY ADDS

⇒ This study has established a non-invasive prediction model for colorectal adenoma based on lifestyle habits and gut microbiota information with good predictive performance.

⇒ This study has developed a 10 megabyte app for predicting the risk of colorectal adenoma and proposed different colonoscopy screening strategies for distinct risk groups.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study combines lifestyle habits and gut microbiota information to predict the occurrence of adenoma, developing a non-invasive, convenient, and effective method for colorectal adenoma prediction for better application in a large population.

INTRODUCTION

As the most prevalent precancerous lesion of colorectal cancer (CRC), the detection of colorectal adenoma (CRA) represents a critical opportunity for prevention and early warning of this pervasive malignancy worldwide.[1] However, in China, the implementation of large-scale colonoscopy screening among asymptomatic individuals faces significant challenges due to its high cost and limited popularity nationwide. Despite Chinese expert consensus advocating for colonoscopy screening to commence at the age of 50 for asymptomatic individuals,[2] significant heterogeneity in colorectal neoplasm risk profiles across demographic groups remains unaddressed.

Lifestyle habits and gut microbial composition are established as pivotal determinants in the CRC pathogenesis. Robust evidence confirms that unhealthy diet,[3] smoking,[4] drinking[5] and chronic diseases[6] may increase the risk of colorectal neoplasms, whereas protective effects are associated with high-fibre diets,[7] regular physical activity[5] and appropriate chemopreventive intervention.[8]

Therefore, integrating lifestyle parameters as predictors may optimise colonoscopy screening efficacy.[9]

Previous risk stratification approaches have been constrained by oversimplified lifestyle scoring systems lacking individualised predictive value. Consequently, developing a precision prediction model incorporating detailed lifestyle–microbiota profiles becomes imperative. Our methodology addresses this need through risk-adapted screening protocols designed to enhance high-risk CRA detection efficiency in population cohorts, ultimately advancing CRC prevention strategies.

METHODS

Study design

This observational study included three independent cohorts (figure 1A and online supplemental figure 1), meticulously designed to advance CRA risk prediction and screening optimisation. Cohort 1 prospectively enrolled 683 participants undergoing colonoscopy tests at the East Campus of Renji Hospital, Shanghai Jiao Tong University School of Medicine, from 2022 to 2023. Comprehensive data on lifestyle habits and stool samples were gathered from cohort 1 for derivation of the CRA risk model. For external validation, cohort 2 prospectively recruited 1529 participants from the West Campus of Renji Hospital, also undergoing colonoscopy tests from 2022 to 2023, with their lifestyle habits documented via standardised questionnaires. Additionally, cohort 3 retrospectively assembled 7535 participants without CRA history who underwent colonoscopy tests for routine physical examinations at the Physical Examination Center of Renji Hospital between 2013 and 2022, establishing real-world CRA incidence patterns in screening-naïve populations. Subsequently, individualised CRA screening strategies for different risk groups were developed through the CRA risk model.

This observational study is reported according to the TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis+Artificial Intelligence) statement[10] (online supplemental materialresearch checklist)).

Mathematical model establishment

A mathematical model for CRA risk stratification was developed via multivariable logistic regression analysis. The modelling process comprised two sequential phases: variable prescreening and stepwise selection. Initial univariable logistic regression analyses screened all candidate predictors, retaining variables meeting dual thresholds: (1) p value <0.2 and (2) β coefficient absolute values exceeding 0.05. Subsequently, multivariable logistic regression analysis employing backward elimination refined the model through iterative exclusion of non-significant covariates. Logistic regression analysis and backward elimination procedures were conducted using the glm() and step() functions in R V.4.1.0.

Random forest analysis and gradient boosting were also used for our mathematical model establishment. The random forest model was created by the RandomForestClassifier function in Python's sklearn.ensemble package, and the gradient boosting model was created by the XGBClassifier function in Python's xgboost package. The optimisation of parameters was conducted using the GridSearchCV method in sklearn.model_selection. The screening range of model parameters and the final optimal parameters are provided in the online supplemental material.

Statistical analysis

Student's t-test or non-parametric test was used to compare data corresponding to normal distribution or skewed distribution. The χ^2 test was performed to compare rates, and univariable logistic regression analysis was used for correlation analysis. The discrimination and calibration of the model were assessed by receiver operating characteristic (ROC) curve analysis and Hosmer-Lemeshow test, respectively. DeLong's test was used to compare the discrimination between the two models. The Youden index was evaluated for cut-off value determination.

IBM SPSS Statistics V.26.0, R V.4.1.0, and Python V.3.8.10 were used for data analysis. GraphPad V.8.0.2, R V.4.1.0, and Python V.3.8.10 were used for graph plotting. Generally, statistical significance was defined at a p value less than 0.05.

Patient and public involvement

We prioritised patient input throughout our study. Patients helped design the research question and lifestyle habits evaluation system based on their priorities and experiences. A patient advisory group contributed to study design, ensuring it was accessible and relevant. They were not involved in the recruitment in our study. Following completion of the research, we planned to share outcomes with participants through personalised, accessible formats, fostering research awareness within this group.

Colonoscopy and histopathological diagnosis, inclusion and exclusion criteria, evaluation of lifestyle habits, stool collection and DNA extraction, 16S rRNA gene sequencing, and quantitative real-time PCR (qPCR) are provided in the online supplemental material.

RESULTS

Significant differences in lifestyle habits and gut pathogens between patients and healthy controls

A total of 683 eligible and consenting participants were enrolled in cohort 1 (figure 1A), comprising 202 patients with CRA and 481 healthy controls (HC). The collection of lifestyle information and faecal samples strictly followed standardised protocols (online supplemental material). Comparative analysis of basic and lifestyle characteristics (table 1 and figure 1B) demonstrated that age was significantly associated with CRA risk (OR 1.06; 95% CI 1.04, 1.08; p=7.95E-14), with men exhibiting elevated

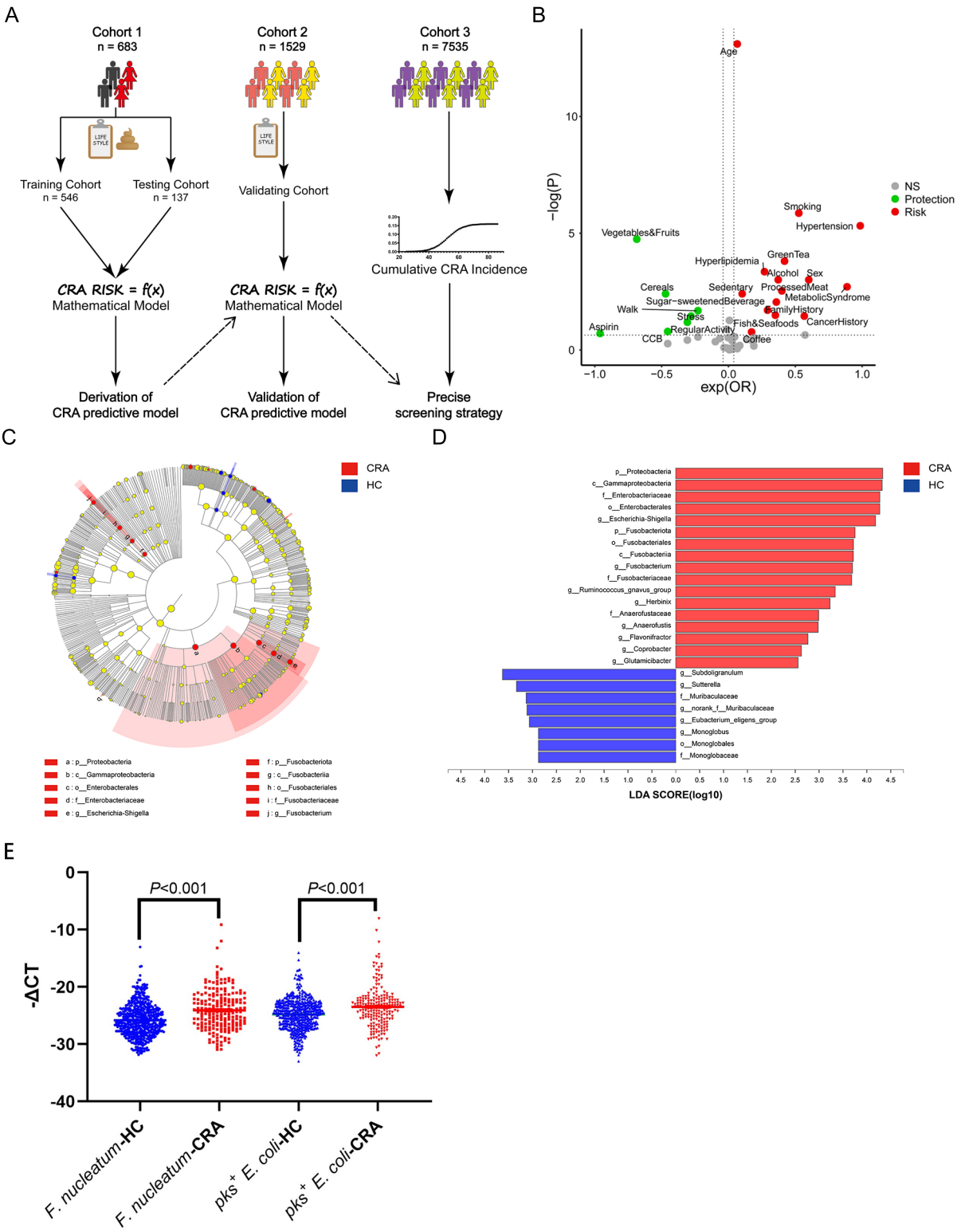

Figure 1 There are significant differences in lifestyle habits and gut pathogens between patients with CRA and HC. (A) Study flow chart. (B) Volcano plot demonstrating the risk and protective factors for CRA. (C) LEfSe cladogram defining microbiological differences between CRA and HC. (D) LEfSe histogram showing microbiological differences between CRA and HC. (E) The relative abundance of *Fusobacterium nucleatum* and *pks*$^{+}$ *Escherichia coli* in the faeces of patients with CRA (n=481) compared with those of the HC group (n=202). CRA, colorectal adenoma; HC, healthy control; NS, not significant; CCB, calcium channel blockers; LDA, linear discriminant analysis; CT, cycle threshold; LEfSe, linear discriminant analysis effect size.

susceptibility (OR 1.81; 95% CI 1.30, 2.53; p=0.001). Chronic health conditions, particularly hypertension (OR 2.66; 95% CI 1.84, 3.83; p=4.87E-06), hyperlipidaemia (OR 1.30; 95% CI 1.12, 1.50; p=4.52E-04), and metabolic syndrome (OR 2.41; 95% CI 1.38, 4.18; p=0.002), emerged as strong risk factors for CRA. Furthermore,

Table 1 Associated factors for CRA in cohort 1

	CRA group (n=202)	HC group (n=481)	OR (95% CI)	P value	P value for trend
Sex, n (%)			1.81 (1.30, 2.53)	0.001	
Male	122/202 (60.40)	220/481 (45.74)			
Female	80/202 (39.60)	261/481 (54.26)			
Age (years)	59.67±10.81	51.23±12.92	1.06 (1.04, 1.08)	7.95E-14	
Body mass index (kg/m^2)	23.61±3.00	23.22±3.24	1.02 (0.97, 1.07)	0.406	
Hypertension, n (%)	76/202 (37.62)	89/481 (18.50)	2.66 (1.84, 3.83)	4.87E-06	
Angiotensin-Converting Enzyme Inhibitors (ACEI)	38/76 (50.00)	44/89 (49.44)	1.02 (0.55, 1.89)	0.943	
Calcium Channel Blockers (CCB)	22/76 (28.95)	33/89 (37.08)	0.63 (0.32, 1.21)	0.164	
Diabetes, n (%)	13/202 (6.44)	31/481 (6.44)	1.04 (0.97, 1.13)	0.277	
Metformin	3/13 (23.07)	10/31 (32.26)	0.63 (0.14, 2.81)	0.544	
Hyperlipidaemia, n (%)	35/202 (17.33)	46/481 (9.56)	1.30 (1.12, 1.50)	4.52E-04	
Statin	19/35 (54.29)	25/46 (54.35)	1.00 (0.41, 2.41)	0.996	
Cardiovascular disease, n (%)	11/202 (5.45)	22/481 (4.57)	1.01 (0.90, 1.13)	0.871	
Metabolic syndrome, n (%)	27/202 (13.37)	29/481 (6.03)	2.41 (1.38, 4.18)	0.002	
Aspirin, n (%)	3/202 (1.49)	11/481 (2.29)	0.38 (0.08, 1.80)	0.200	
Cancer history, n (%)			1.75 (1.04, 2.94)	0.036	
None	175/202 (86.63)	442/481 (91.89)			
History of other cancers	27/202 (13.37)	39/481 (8.11)			
Family history, n (%)			1.42 (1.09, 1.85)		0.009
None	161/202 (79.20)	422/481 (87.73)	Reference	–	
Family history of other gastrointestinal cancers	17/202 (8.42)	26/481 (5.41)	1.71 (0.91, 3.24)	0.098	
Family history of CRC	24/202 (11.88)	33/481 (6.86)	1.91 (1.09, 3.33)	0.023	
Thyroidectomy, n (%)	7/202 (3.47)	15/481 (3.12)	0.93 (0.80, 1.08)	0.327	
Appendectomy, n (%)	8/202 (3.96)	18/481 (3.74)	1.06 (0.45, 2.48)	0.892	
Cholecystectomy, n (%)	8/202 (3.96)	16/481 (3.33)	1.20 (0.51, 2.85)	0.682	
Smoking, n (%)			1.68 (1.36, 2.07)		1.40E-06
Never smoked	128/202 (63.37)	391/481 (81.29)	Reference	–	
Ever smoked	25/202 (12.38)	34/481 (7.07)	2.25 (1.29, 3.91)	0.004	
Current smoking	49/202 (24.26)	56/481 (11.64)	2.67 (1.74, 4.12)	0.001	
Alcohol, n (%)			1.44 (1.16, 1.77)		0.001
Never drunk	145/202 (71.78)	406/481 (84.41)	Reference	–	
Ever drunk	14/202 (6.93)	12/481 (2.49)	3.27 (1.48, 7.23)	0.003	
Current drinking	43/202 (21.29)	63/481 (13.10)	1.91 (1.24, 2.943)	0.003	
Sleeping duration (hours/day)	7.07±1.13	7.09±1.00	0.98 (0.84, 1.15)	0.834	
Stress, n (%)			0.75 (0.57, 0.98)		0.036
Mild	53/202 (26.24)	107/481 (22.25)	Reference	–	
Moderate	128/208 (63.37)	291/481 (60.50)	0.89 (0.60, 1.31)	0.550	
Severe	21/202 (10.40)	83/481 (17.25)	0.51 (0.29, 0.91)	0.023	
Processed meat, n (%)			1.48 (1.14, 1.93)		0.003
Rarely	74/202 (36.63)	219/481 (45.53)	Reference	–	
Occasionally, <25 g/day	102/202 (50.50)	232/481 (48.23)	1.30 (0.92, 1.85)	0.142	
Usually, ≥25 g/day	26/202 (12.87)	30/481 (6.24)	2.57 (1.42, 4.62)	0.002	
Red meat, n (%)			1.20 (0.85, 1.69)		0.308
Rarely	8/202 (3.96)	14/481 (2.91)	Reference	–	
Occasionally, <50 g/day	130/202 (64.36)	284/481 (59.04)	0.80 (0.33, 1.96)	0.626	

Continued

Table 1 Continued

	CRA group (n=202)	HC group (n=481)	OR (95% CI)	P value	P value for trend
Usually, ≥50 g/day	64/202 (31.68)	183/481 (38.05)	0.61 (0.25, 1.53)	0.292	
White meat, n (%)			0.96 (0.69, 1.33)		0.803
Rarely	12/202 (5.94)	24/481 (4.99)	Reference	–	
Occasionally, <50 g/day	144/202 (71.29)	347/481 (72.14)	0.83 (0.40, 1.71)	0.612	
Usually, ≥50 g/day	46/202 (22.77)	110/481 (22.87)	0.84 (0.39, 1.81)	0.651	
Fish and seafood, n (%)			1.41 (1.03, 1.94)		0.033
Rarely	10/202 (4.95)	32/481 (6.65)	Reference	–	
Occasionally, <50 g/day	132/202 (65.35)	343/481 (71.31)	1.23 (0.59, 2.58)	0.580	
Usually, ≥50 g/day	60/202 (29.70)	106/481 (22.04)	1.81 (0.83, 3.94)	0.134	
Vegetables and fruits, n (%)			0.50 (0.36, 0.68)		1.83E-05
Rarely	4/202 (1.98)	4/481 (0.83)	Reference	–	
Occasionally, <500 g/day	124/202 (61.39)	215/481 (44.70)	0.58 (0.14, 2.38)	0.442	
Usually, ≥500 g/day	74/202 (36.63)	262/481 (54.47)	0.28 (0.07, 1.16)	0.079	
Cereals, n (%)			0.62 (0.45, 0.86)		0.004
Rarely	22/202 (10.89)	28/481 (5.82)	Reference	–	
Occasionally, <25 g/day	144/202 (71.29)	329/481 (68.40)	0.56 (0.31, 1.01)	0.053	
Usually, ≥25 g/day	36/202 (17.82)	124/481 (25.78)	0.37 (0.19, 0.72)	0.004	
Eggs, n (%)			0.90 (0.68, 1.19)		0.448
Rarely	13/202 (6.44)	16/481 (3.33)	Reference	–	
Occasionally, <50 g/day	86/202 (42.57)	217/481 (45.11)	0.49 (0.23, 1.06)	0.069	
Usually, ≥50 g/day	103/202 (50.99)	248/481 (51.56)	0.51 (0.24, 1.10)	0.086	
Milk, n (%)			1.04 (0.84, 1.30)		0.705
Rarely	63/202 (31.19)	144/481 (29.94)	Reference	–	
Occasionally, <200 mL/day	83/202 (41.09)	221/481 (45.95)	0.86 (0.58, 1.27)	0.442	
Usually, ≥200 mL/day	56/202 (27.72)	116/481 (24.11)	1.10 (0.71, 1.71)	0.657	
Sugar-sweetened beverage, n (%)			1.33 (1.05, 1.70)		0.020
Rarely	69/202 (34.16)	190/481 (39.50)	Reference	–	
Occasionally, <200 mL/day	94/202 (46.53)	236/481 (49.06)	1.10 (0.76, 1.58)	0.620	
Usually, ≥200 mL/day	39/202 (19.31)	55/481 (11.44)	1.95 (1.19, 3.20)	0.008	
Coffee, n (%)			1.18 (0.93, 1.50)		0.172
Rarely	122/202 (60.40)	321/481 (66.74)	Reference	–	
Occasionally, <200 mL/day	58/202 (28.71)	114/481 (23.70)	1.34 (0.92, 1.96)	0.131	
Usually, ≥200 mL/day	22/202 (10.89)	46/481 (9.56)	1.26 (0.73, 2.18)	0.412	
Green tea, n (%)			1.51 (1.22, 1.88)		1.59E-04
Rarely	79/202 (39.11)	256/481 (53.22)	Reference	–	
Occasionally, <200 mL/day	73/202 (36.14)	154/481 (32.02)	1.54 (1.06, 2.24)	0.025	
Usually, ≥200 mL/day	50/202 (24.75)	71/481 (14.76)	2.28 (1.47, 3.55)	0.001	
Carbonated drinks, n (%)			1.08 (0.77, 1.51)		0.657
Rarely	153/202 (75.74)	360/481 (74.84)	Reference	–	
Occasionally, <200 mL/day	41/202 (20.30)	115/481 (23.91)	0.84 (0.56, 1.26)	0.394	
Usually, ≥200 mL/day	8/202 (3.96)	6/481 (1.25)	3.14 (1.07, 9.20)	0.037	
Sweet foods, n (%)			1.14 (0.86, 1.50)		0.361
Rarely	68/202 (33.66)	174/481 (36.18)	Reference	–	
Occasionally, <50 g/day	116/202 (57.43)	275/481 (57.17)	1.08 (0.76, 1.54)	0.673	
Usually, ≥50 g/day	18/202 (8.91)	32/481 (6.65)	1.44 (0.76, 2.74)	0.266	
Calcium, n (%)	36/202 (17.82)	74/481 (15.38)	1.19 (0.77, 1.85)	0.429	

Continued

Table 1 Continued

	CRA group (n=202)	HC group (n=481)	OR (95% CI)	P value	P value for trend
Vitamin D, n (%)	11/202 (5.45)	35/481 (7.28)	0.73 (0.37, 1.48)	0.385	
Folic acid, n (%)	8/202 (3.96)	11/481 (2.29)	1.76 (0.70, 4.45)	0.231	
Antioxidation, n (%)	32/202 (15.84)	93/481 (19.33)	0.79 (0.51, 1.22)	0.282	
Regular activity, n (%)	97/202 (48.20)	268/481 (55.72)	0.73 (0.53, 1.02)	0.066	
Vigorous activity (hours/week)	1.03±2.89	1.02±2.61	1.00 (0.94, 1.06)	0.977	
Moderate activity (hours/week)	2.87±5.81	3.71±6.94	0.98 (0.95, 1.01)	0.134	
Walking (hours/day)	1.10±0.91	1.30±1.08	0.79 (0.65, 0.97)	0.021	
Sedentary (hours/day)	5.75±2.47	5.15±2.46	1.10 (1.03, 1.18)	0.004	
Metabolic equivalent	2701.62±2736.44	3185.84±3046.08	1.00 (1.00, 1.00)	0.055	
Perceived Stress Scale-14 (PSS-14) score	33.21±8.08	33.92±8.19	0.99 (0.97, 1.01)	0.301	

Data are presented as percentages except in age, sleeping duration, body mass index, vigorous activity, moderate activity, walk, sedentary, metabolic equivalent, PSS-14 score which are presented as mean ± SDs
CRA, colorectal adenoma; CRC, colorectal cancer; HC, healthy control.

individuals with a personal history of cancer (OR 1.75; 95% CI 1.04, 2.94; p=0.036) or a family history of CRC (OR 1.91; 95% CI 1.09, 3.33; p=0.023) were also at higher risk. Lifestyle factors such as smoking (OR 1.68; 95% CI 1.36, 2.07; p=1.40E-06), alcohol consumption (OR 1.44; 95% CI 1.16, 1.77; p=0.001), and a diet rich in processed meat (OR 1.48; 95% CI 1.14, 1.93; p=0.003) were found to significantly elevate CRA risk. Conversely, a diet high in fibre, particularly from cereals (OR 0.62; 95% CI 0.45, 0.86; p=0.004) and vegetables and fruits (OR 0.50; 95% CI 0.36, 0.68; p<0.001), was associated with a reduced risk. Intriguingly, urbanisation-related lifestyle habits, including sedentary behaviours (OR 1.10; 95% CI 1.03, 1.18; p=0.004), decreased walking time (OR 0.79; 95% CI 0.65, 0.97; p=0.021), and consumption of sugar-sweetened beverages (OR 1.33; 95% CI 1.05, 1.70; p=0.02), were also identified as contributors to increased CRA risk.

Meanwhile, 16S rRNA gene analysis was performed in faecal samples sourced from 104 age-matched and gender-matched donors within cohort 1, encompassing 45 CRA and 69 HC. This analysis uncovered a notable enrichment of Proteobacteria and Fusobacteriota at the phylum level, as well as *Escherichia–Shigella* and *Fusobacterium* at the genus level, in the stool samples of patients with CRA compared with those of HC participants (figure 1C,D). In the previous studies, *pks*$^+$ *Escherichia coli* and *Fusobacterium nucleatum* were the known CRC-related species belonging to *Escherichia–Shigella* and *Fusobacterium*. Therefore, the relative abundance of *pks*$^+$ *E. coli* and *F. nucleatum* was measured by qPCR in the faeces of every participant from cohort 1 (figure 1E). There was a significantly higher abundance of *F. nucleatum* and *pks*$^+$ *E. coli* in patients with CRA ($p_{F.\ nucleatum}$=7.62E-10, $p_{pks+\ E.\ coli}$=7.11E-06).

Our findings demonstrated distinct lifestyle patterns and gut microbial pathogen profiles differentiating patients with CRA and HC, establishing an empirical foundation for the derivation and validation of a CRA risk prediction model.

Lifestyle-based CRA prediction model demonstrated promising predictive performance in external validation

To develop a non-invasive predictive model for CRA, we randomly allocated 80% of cohort 1 participants to the training cohort (n=546), with the remaining 20% assigned to the testing cohort (n=137). There was no significant difference in the demographic characteristics between the training cohort and the testing cohort (online supplemental table 1). 23 significantly relevant variables were selected by prescreening procedure before logistic regression modelling (figure 1B). The prescreening selection follows the criterion that the p value has to be less than 0.2 and the absolute value of β should be more than 0.05 (OR greater than 1.05 or less than 0.95). Then, a backward elimination procedure resulted in reducing the pool of 23 retained candidates to 12 variables in the logistic regression model, including age, sex,[1] sedentary time, hyperlipidaemia, cancer history, processed meat consumption, vegetable and fruit consumption, cereal consumption, sugar-sweetened beverage consumption, coffee consumption, smoking status, and regular use of aspirin. Given that family history[11] (OR 1.42; 95% CI 1.09, 1.85; p=0.009) and alcohol use[5] (OR 1.44; 95% CI 1.16, 1.77; p=0.001) are established risk factors that are also significant in the univariable logistic regression analysis, we included them in the final model (figure 2A and online supplemental table 2). The equation for the final CRA absolute risk model is as follows:

$$Y = exp\left(-6.05 + 0.69^{*}X_1 + 0.61^{*}X_2 + 0.31^{*}X_3 + 0.29^{*}X_4 + 0.28^{*}X_5 + 0.25^{*}X_6 + 0.26^{*}X_7 + 0.24^{*}X_8 + 0.22^{*}x_9 + 0.2^{*}X_{10} + 0.08^{*}X_{11} - 0.53^{*}X_{12} - 0.82^{*}x_{13} - 1.93^{*}X_{14}\right)$$

$$CRA\ absolute\ risk\ index = \frac{Y}{(1+Y)}$$

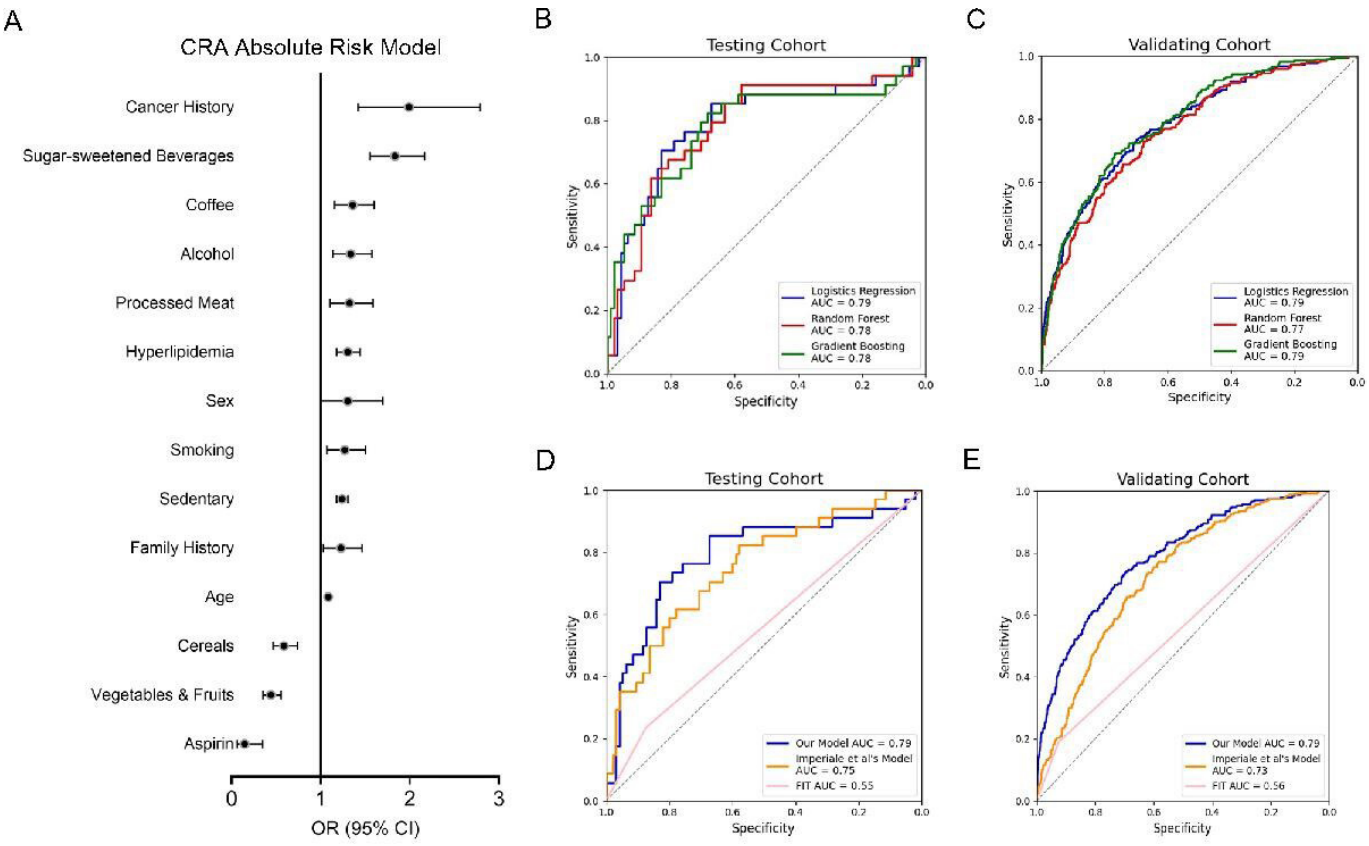

Figure 2 Lifestyle-based CRA prediction model demonstrated promising predictive performance in external validation. (A) Forest plot of CRA absolute risk model. (B) The predictive performance of Logistics regression model, random forest model and gradient boosting model indicated by ROC curve analysis for CRA in Testing Cohort (n=137). (C) The predictive performance of Logistics regression model, random forest model and gradient boosting model indicated by ROC curve analysis for CRA in Validating Cohort (n=1,529). (D) The predictive performance of FIT, Imperiale et al's model and our CRA absolute risk model indicated by ROC curve analysis for CRA in Testing Cohort (n=137). (E) The predictive performance of FIT, Imperiale et al's model and our CRA absolute risk model indicated by ROC curve analysis for CRA in Validating Cohort (n=1,529).

Our model performed well in the testing cohort (n=137) in terms of discrimination (c-statistic=0.79; blue curve in figure 2B and online supplemental table 3) and calibration (Hosmer-Lemeshow test: p=0.62). For further validation in a larger population, 1529 eligible and consenting participants were prospectively enrolled in cohort 2 as a validating cohort (figure 1A). Excellent discrimination (c-statistic=0.79; blue curve in figure 2C and online supplemental table 3) and calibration (Hosmer-Lemeshow test: p=0.65) was also demonstrated in the validating cohort (n=1529). To facilitate clinical implementation, we divided cohort 2 into absolute high-risk and absolute low-risk groups with an optimal sensitivity and specificity of 74.0% and 69.4% for CRA prediction (the optimal cut-off value is 0.32). In cohort 2, the CRA incidence was 30.14% and 11.39% for the high-risk and low-risk group, respectively (online supplemental table 4).

Meanwhile, random forest analysis and gradient boosting were employed within the training cohort for CRA prediction. Following parameter optimisation and validation, the random forest model displayed excellent discrimination in both the testing cohort (c-statistic=0.78; red curve in figure 2B and online supplemental table 3) and the validating cohort (c-statistic=0.77; red curve in figure 2C and online supplemental table 3). Similarly, the gradient boosting model exhibited strong predictive performance, achieving c-statistics of 0.78 (green curve in figure 2B and online supplemental table 3) and 0.79 (green curve in figure 2C and online supplemental table 3) in the testing and validating cohorts, respectively.

Imperiale et al created a predictive model for advanced colorectal neoplasms showing great discrimination in white American populations,[12] incorporating 13 variables (age, sex, smoking, alcohol abuse, metabolic syndrome, red meat consumption, marriage status, advanced education, regular use of aspirin and nonsteroidal anti-inflammatory drugs, regular activity, and moderate and vigorous physical activity). To assess its generalisability to Chinese populations, we examined these 13 variables in our training cohort by multivariable logistic regression. DeLong's test was used to compare the validation separately in the testing cohort (c-statistic=0.75, Hosmer-Lemeshow test: p=0.02; orange curve in figure 2D and online supplemental table 3) and in the validating cohort (c-statistic=0.73, Hosmer-Lemeshow test: p=0.34; orange curve in figure 2E and online supplemental table 3) with our model, and our model showed significantly greater

discrimination in both the testing cohort (DeLong's test: p=0.36) and the validating cohort (DeLong's test: p=4.41E-05).

Given the established role of faecal immunochemical test (FIT) in colorectal neoplasia screening, we retrospectively collected precolonoscopy FIT results from both the testing cohort and the validating cohort. The FIT results demonstrated a relatively weak ability to distinguish between CRA and HC ($c\text{-statistic}_{\text{testing cohort}}$=0.55, $c\text{-statistic}_{\text{validating cohort}}$=0.56; pink curve in figure 2D,E). Comparative analysis revealed significant superiority of our lifestyle-based model over FIT for CRA risk stratification (DeLong's test: $p_{\text{testing cohort}}$=0.002, $p_{\text{validating cohort}}$ <0.001).

Therefore, lifestyle-based CRA prediction models have been developed by multivariable logistic regression analysis, random forest analysis, and gradient boosting. These non-invasive models demonstrated optimal equilibrium between model calibration and discriminative accuracy within Chinese population cohorts for CRC early warning.

Adjunction of gut microbiota can improve the predictive performance of the lifestyle-based CRA prediction model

Analysis of CRA-associated microbial biomarkers (figure 1E) revealed the optimal cut-off value to be at −19.48 (*F. nucleatum*) and −23.61 (*pks*$^+$ *E. coli*) from the ROC curve analysis. The study participants were subsequently stratified into binary groups according to their *F. nucleatum* and *pks*$^+$ *E. coli* abundance. These microbiological stratification parameters were subsequently integrated into the CRA precise risk model (figure 3A and online supplemental table 5). The mathematical expression for the CRA precise risk model is as follows:

$$Y = exp\,(-6.42 + 0.73^{*}X_1 + 0.61^{*}X_2 + 0.56^{*}X_3 + 0.52^{*}X_4 + 0.35^{*}X_5 + 0.32^{*}X_6 + 0.3^{*}X_7 + 0.28^{*}X_8 + 0.25^{*}X_9 + 0.22^{*}X_{10} + 0.21^{*}X_{11} + 0.15^{*}X_{12} + 0.08^{*}X_{13} - 0.55^{*}X_{14} - 0.82^{*}X_{15} - 1.79^{*}X_{16})$$

$$CRA\ precise\ risk\ index = \frac{Y}{(1+Y)}$$

The integrated model incorporating *F. nucleatum* and *pks*$^+$ *E. coli* abundance stratification demonstrated robust performance within the testing cohort (n=137), showing strong discriminative capacity (c-statistic=0.84; blue curve in figure 3B and online supplemental table 6) and adequate calibration (Hosmer-Lemeshow test: p=0.22). Incorporation of microbial biomarkers enhanced the predictive performance of the logistic regression model ($c\text{-statistic}_{\text{with microbe}}$=0.84, $c\text{-statistic}_{\text{without microbe}}$=0.79, DeLong's test: p=0.006; blue curves in figure 3B and in figure 2B

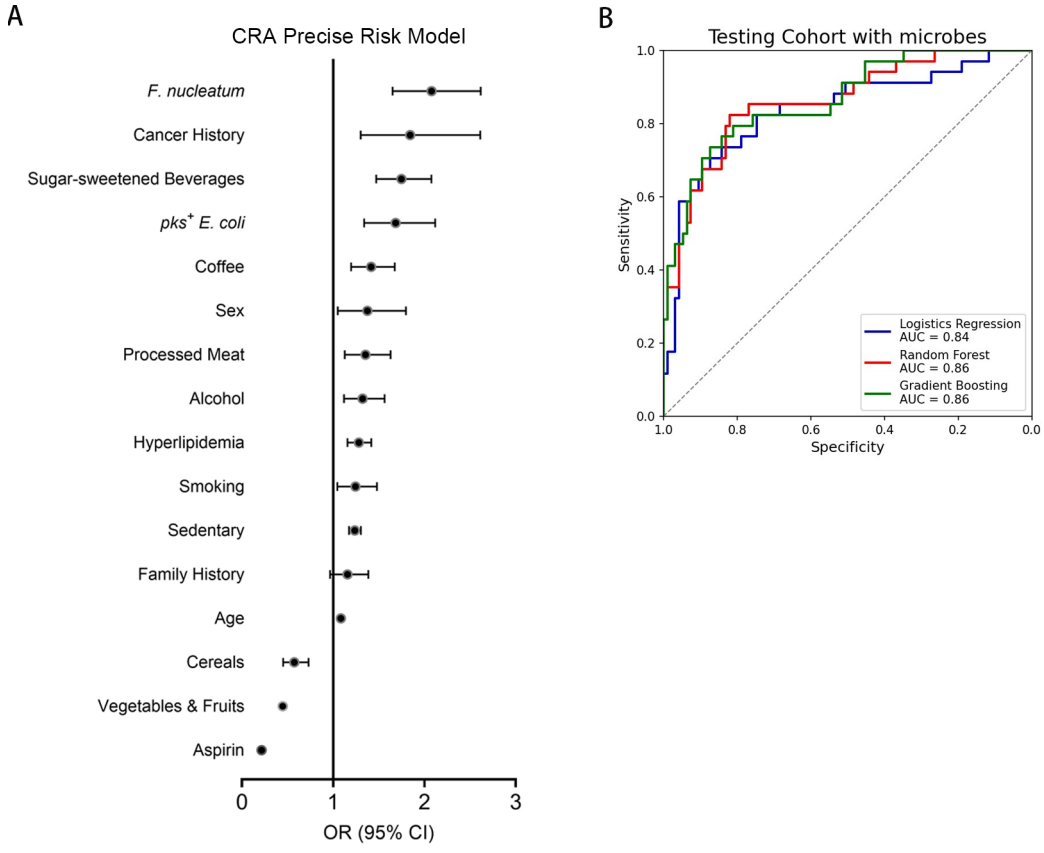

Figure 3 The adjunction of gut microbiota can improve the predictive performance of the lifestyle-based CRA prediction model. (A) Forest plot of the CRA precise risk model. (B) The predictive performance of logistic regression model, random forest model, and gradient boosting model indicated by the ROC curve analysis for CRA in the testing cohort with microbe information (n=137). CRA, colorectal adenoma; t; AUC, area under curve.

and online supplemental table 6). On the other hand, the benefit from combining gut microbe information was also significant in the random forest model (c-statistic$_{\text{with microbe}}$=0.86, c-statistic$_{\text{without microbe}}$=0.78, DeLong's test: p=0.012; red curves in figure 3B and in figure 2B and online supplemental table 6) and gradient boosting model (c-statistic$_{\text{with microbe}}$=0.86, c-statistic$_{\text{without microbe}}$=0.78, DeLong's test: p=0.004; green curves in figure 3B and in figure 2B and online supplemental table 6).

These results demonstrated that incorporation of CRA-associated microbial biomarkers (*F. nucleatum* and *pks*$^{+}$ *E. coli*) increased our models' predictive performance, supporting their potential utility in non-invasive screening strategies for CRC prevention.

Individualised screening strategies based on CRA risk and facilitation of home self-testing

For better application of CRA risk prediction to clinical practice, the development of age-specific predictive models of lifestyle habits was pursued. Age was deliberately excluded from the logistic regression model, which included the remaining 23 significantly associated factors. Subsequent backward elimination procedure was performed to develop the CRA relative risk model (figure 4A and online supplemental

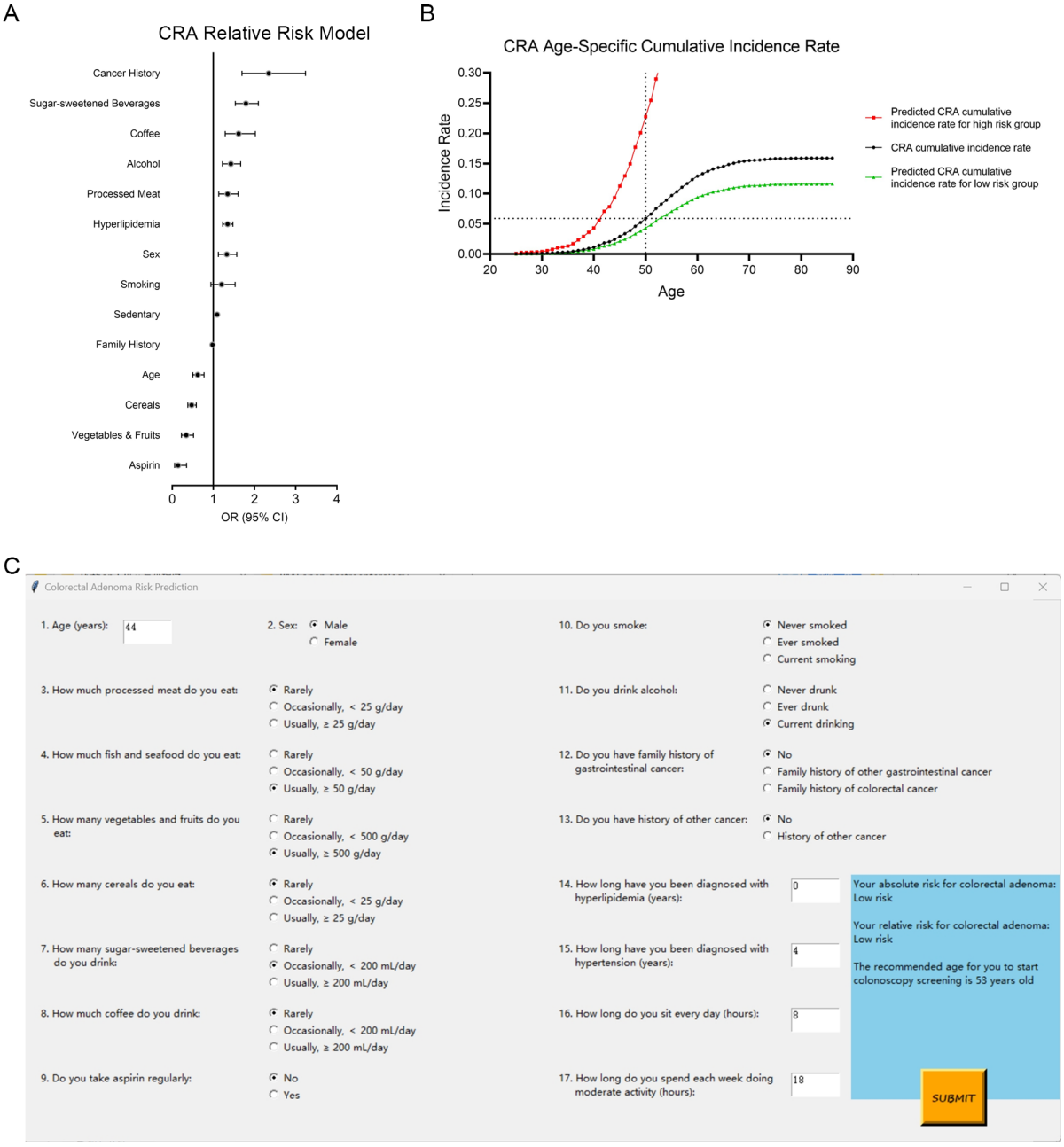

Figure 4 Individualised screening strategies based on CRA risk and facilitation of home self-testing. (A) Forest plot of CRA relative risk model. (B) CRA age-specific cumulative incidence curves for different risk groups. (C) User interface of the colorectal adenoma risk prediction application. CRA, colorectal adenoma.

table 7). The mathematical expression for the CRA relative risk model is as follows:

$$Y = \exp\left(-1.08 + 0.85^{*}X_1 + 0.58^{*}X_2 + 0.48^{*}X_3 + 0.35^{*}X_4 + 0.3^{*}X_5 + 0.3^{*}X_6 + 0.28^{*}X_7 + 0.18^{*}X_8 + 0.09^{*}X_9 - 0.03^{*}X_{10} - 0.47^{*}X_{11} - 0.75^{*}X_{12}\right)$$

$$\textit{CRA relative risk index} = \frac{Y}{(1+Y)}$$

This model indicated discrimination (c-statistic$_{\text{testing cohort}}$=0.63, c-statistic$_{\text{validating cohort}}$=0.69) and calibration (Hosmer-Lemeshow test: $p_{\text{testing cohort}}$=0.32, Hosmer-Lemeshow test: $p_{\text{validating cohort}}$=0.16) in both the testing cohort (n=137) and the validating cohort (n=1529).

For better application, cohort 2 was stratified into relative high-risk and relative low-risk groups with an optimal sensitivity and specificity of 74.0% and 57.7% for CRA prediction (the optimal cut-off value is 0.34). The observed CRA incidence demonstrated marked interstrata disparity (high-risk: 39.6% vs low-risk: 11.3%; online supplemental table 4), respectively. Age-adjusted regression analysis revealed a 3.76-fold elevated likelihood of adenoma in the relative high-risk group (OR 3.76; 95% CI 2.66, 5.31), equivalent to a 5.26-fold excess risk relative to low-risk counterparts. Combined with the distribution in the population, it was calculated that the risk of occurrence in the relative high-risk group was 3.84 times higher than population-average risk, and the risk of occurrence in the relative low-risk group was 0.73 times higher than the population-average risk.

Cohort 3 comprised 7535 asymptomatic individuals undergoing screening colonoscopy for physical examination (figure 1A). Age-specific CRA cumulative incidence rates were derived from participant age and diagnosis (black curve in figure 4B). The anticipated cumulative incidence rate for each relative risk group was then calculated by their age-adjusted ORs(figure 4B). At the conventional screening initiation age of 50 years,[2] the baseline CRA detection rate stood at 5.9%. Respectively, this risk threshold was reached at ages 42 years and 53 years in individuals with relative high-risk and relative low-risk CRA. Our analysis recommends initiating colonoscopy screening at the age of 42 for individuals in the relative high-risk group, while suggesting delayed initiation until 53 years for the relative low-risk group.

Subsequently, we developed a digital risk assessment tool enabling individuals to self-assess their CRA risk with individualised screening recommendations (figure 4C). This app has been uploaded online (https://github.com/YiluZzzzzz/Colorectal-Adenoma-Risk-Prediction-APP).

As a result, we proposed individualised screening strategies for CRA detection alongside developing a small programme for public CRA risk self-prediction. We aimed at better application of our research findings to clinical practice.

DISCUSSION

As a clinically significant precursor lesion, screening for CRA is crucial to the prevention and early detection of CRC. Modifiable environmental determinants, notably lifestyle parameters and gut microbial profiles, were demonstrated to exhibit mechanistic associations with colorectal tumourigenesis. Based on environmental factors, our investigation established non-invasive CRA risk models for early warning of CRC. Individualised colonoscopy screening strategies and CRA risk prediction app were subsequently proposed to optimise healthcare implementation.

The major achievement of our study is the robust lifestyle-based CRA risk models. In the previous studies, due to the undetailed lifestyle evaluation system, the predictive model performance was beyond satisfactory (c-statistics of about 0.60–0.65).[9] Comprehensive lifestyle evaluation in our research boosted the predictive performance. In the CRA absolute risk model established by logistic regression analysis, variables including age,[1] male sex,[1] hyperlipidaemia,[13] cancer history,[14] processed meat consumption,[3] smoking,[4] family history,[11] and alcohol use[5] are the known risk factors for CRC, while the protective effects of high-fibre intake[7] and regular use of aspirin[8] have been proven. New CRC-relevant habits such as sedentary lifestyle,[15] sugar-sweetened beverages,[16] and coffee consumption,[17] which have been discussed in Western countries for decades, require further investigation in the Chinese population. Comparative analysis with Imperiale *et al*'s lifestyle-based model[12] revealed superior discriminative capacity in our cohorts, substantiating enhanced population-specific applicability for Chinese screening paradigms.

Our two principal translational innovations comprise the development of individualised colonoscopy screening strategies and the CRA risk prediction app. Prior research limitations, particularly the absence of clinically actionable risk stratification,[12] left unresolved the critical challenge of managing high-risk populations. Our screening strategies address this gap by providing individualised clinical practice guidelines postrisk evaluation. Simultaneously, the CRA risk prediction app operationalises our findings for mass population health initiatives. This self-prediction app demonstrates dual utility: operational efficiency in risk stratification alongside enhancement of public health literacy regarding colorectal neoplasia prevention.

The adjunction of gut microbiota information with CRA risk prediction constitutes our foremost methodological innovation. In the past few years, gut microbiota has been considered a promising biomarker for colorectal neoplasm prediction.[18] Mechanistic investigations have specifically implicated *F. nucleatum*[19] and *pks*$^{+}$ *E. coli*[20] in colorectal tumourigenesis through their immunomodulatory effects, metabolic reprogramming, and genotoxic activity, as demonstrated in *Apc*$^{min/+}$ murine models recapitulating human

neoplasia progression. The convergence of epidemiological instruments (lifestyle questionnaires) and microbial profiling may establish a novel paradigm—the combination of behavioural and bacteriological risk stratification emerges as a viable population-level screening modality balancing diagnostic accuracy with practical implementation feasibility.

FIT and colonoscopy remain the two principal screening modalities endorsed by Chinese expert consensus on colorectal neoplasm prevention.[2] While existing evidence[21–23] corroborates the adequate specificity of FIT for adenoma detection, its suboptimal sensitivity fundamentally limits diagnostic utility. Conversely, colonoscopy demonstrates robust diagnostic accuracy but remains constrained by specialist dependency and prohibitive healthcare expenditure, rendering population-level implementation impractical. Our models address these limitations through non-invasive implementation requiring neither endoscopic intervention nor faecal sampling, while achieving competitive predictive performance surpassing FIT benchmarks through smartphone-enabled population outreach. We propose deploying this model as a primary triage tool for asymptomatic populations.

Cologuard is an emerging screening method for CRC,[24] demonstrating 87% specificity through faecal DNA biomarker analysis. However, its clinical utility for adenoma detection remains constrained, with sensitivity plummeting to 42% for precursor lesions. Notably, this technology remains unavailable within China's current screening paradigm. Comparative evaluation of Cologuard against our predictive models presents a critical future research direction. Crucially, our model demonstrates substantial economic advantages. This cost differential becomes particularly salient when considering implementation scalability among low-income to middle-income demographics, where our risk-stratified approach offers superior practical feasibility for population-level deployment.

Our study also has certain limitations. First, as an observational design, it identifies associations rather than causal relationships between environmental factors and adenoma risk. Second, while colorectal carcinogenesis involves genetic–environmental interplay, our model currently focuses solely on environmental parameters. Future models need larger cohorts with genomic data for enhanced precision. Third, self-reported lifestyle data introduce inherent recall biases and subjective interpretation discrepancies, despite standardised questionnaires. We propose smartphone integration for real-time digital phenotyping, capturing sleep patterns, diet, and activity through mobile health platforms, to improve data accuracy and enable dynamic model optimisation. Lastly, our Shanghai cohorts limit generalisability. Subsequent validation must encompass diverse demographics to develop population-specific frameworks.

In conclusion, we have established non-invasive prediction models for CRC early detection and prevention, integrating Chinese lifestyle patterns together with gut microbiome profiling. Individualised colonoscopy screening strategies were proposed for stratified CRA risk groups. A dedicated mini-app was subsequently developed to facilitate nationwide CRA risk assessment across China.

Acknowledgements The authors thank all the experts, professors, and coworkers for supporting this study: Dr Yao Zhang, Dr Hua Xiong, Dr Hui-Min Chen, Dr Chao-Fei Chen, Dr Shu-Liang Zhao, Dr Ji-Lin Wang, Dr Wen-Yu Su, and Dr Hui Ding for providing colonoscopy evaluation of the subjects; and Dr Xiao-Yu Chen and Dr Yun Cui for providing pathological diagnosis of the subjects. The authors also thank the Majorbio Cloud Platform for its assistance in data analysis and figure plotting on its online platform (www.majorbio.com). The authors thank all the participants for their participation and all patient advisors for their assistance in this study.

Contributors JF: guarantor of the article. Y-LZ: methodology, software, formal analysis, investigation, writing—original draft, visualisation. J-WD: validation, investigation, writing—review and editing. Z-HL: investigation. X-YM, C-QZ, Y-HX: investigation. C-BZ: writing—review and editing, supervision, funding acquisition. JF: conceptualisation, methodology, writing—review and editing, supervision, project administration, funding acquisition.

Funding This project was supported by the National Key R&D Program of China (2020YFA0509200), National Natural Science Foundation of China (81830081, 31970718, 82203224), Shanghai Municipal Health Commission, Collaborative Innovation Cluster Project (2019CXJQ02), Clinical Research Plan of SHDC (SHDC2020CR1034B), China Postdoctoral Science Foundation (2022M712124), Shanghai Sailing Program (21YF1425600), Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (22CGA17), Shanghai 'Rising Stars of Medical Talents' Youth Development Specialist Program (2023-62) and Innovative Research Team of High-Level Local Universities in Shanghai.

Competing interests None declared.

Patient consent for publication Obtained.

Ethics approval This study was performed in accordance with the Declaration of Helsinki and was approved by the Ethics Review Committee of Renji Hospital, Shanghai Jiao Tong University School of Medicine (KY2022-064). All faecal samples and lifestyle information from cohorts 1 and 2 were obtained with written informed consent. In cohort 3, where no faecal samples or lifestyle data were collected from the subjects, informed consent was waived as approved by the Ethics Review Committee of Renji Hospital, Shanghai Jiao Tong University School of Medicine (KY2022-064).

Provenance and peer review Not commissioned; internally peer reviewed.

Data availability statement Data are available in a public, open access repository. Faecal 16S rRNA sequencing data that support the findings of this study are openly available in the Sequence Read Archive of the NCBI BioProject Repository at https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA880470 (reference number BioProject ID PRJNA880470).[24] The CRA Risk Prediction app that supports the findings of this study is openly available in GitHub (https://github.com/YiluZzzzzz/Colorectal-Adenoma-Risk-Prediction-APP).

ORCID iD

Jing-Yuan Fang http://orcid.org/0000-0003-2282-0248

REFERENCES

1 Dekker E, Tanis PJ, Vleugels JLA, *et al*. Colorectal cancer. *Lancet* 2019;394:1467–80.
2 Chinese Society of Gastroenterology, Cancer Collaboration Group of Chinese Society of Gastroenterology, Chinese Medical Association. Chinese consensus on prevention of colorectal neoplasia (2021, S hanghai). *J of Digest Diseases* 2022;23:58–90.
3 Huang Y, Cao D, Chen Z, *et al*. Red and processed meat consumption and cancer outcomes: Umbrella review. *Food Chem* 2021;356:129697.
4 Botteri E, Borroni E, Sloan EK, *et al*. Smoking and Colorectal Cancer Risk, Overall and by Molecular Subtypes: A Meta-Analysis. *Am J Gastroenterol* 2020;115:1940–9.
5 McNabb S, Harrison TA, Albanes D, *et al*. Meta-analysis of 16 studies of the association of alcohol with colorectal cancer. *Int J Cancer* 2020;146:861–73.
6 Chen H, Zheng X, Zong X, *et al*. Metabolic syndrome, metabolic comorbid conditions and risk of early-onset colorectal cancer. *Gut* 2021;70:1147–54.
7 Chen H-M, Yu Y-N, Wang J-L, *et al*. Decreased dietary fiber intake and structural alteration of gut microbiota in patients with advanced colorectal adenoma. *Am J Clin Nutr* 2013;97:1044–52.
8 McNeil JJ, Gibbs P, Orchard SG, *et al*. Effect of Aspirin on Cancer Incidence and Mortality in Older Adults. *J Natl Cancer Inst* 2021;113:258–65.
9 Vitellius C, Bertrais S, Antier J, *et al*. Evaluation of a risk score based on dietary and lifestyle factors to target a population at risk in colorectal cancer screening. *Dig Liver Dis* 2021;53:900–7.
10 TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385.
11 Fuchs CS, Giovannucci EL, Colditz GA, *et al*. A prospective study of family history and the risk of colorectal cancer. *N Engl J Med* 1994;331:1669–74.
12 Imperiale TF, Monahan PO, Stump TE, *et al*. Derivation and validation of a predictive model for advanced colorectal neoplasia in asymptomatic adults. *Gut* 2021;70:1155–61.
13 Tie G, Yan J, Khair L, *et al*. Hypercholesterolemia Increases Colorectal Cancer Incidence by Reducing Production of NKT and γδ T Cells from Hematopoietic Stem Cells. *Cancer Res* 2017;77:2351–62.
14 Pruitt SL, Gerber DE, Zhu H, *et al*. Survival of patients newly diagnosed with colorectal cancer and with a history of previous cancer. *Cancer Med* 2021;10:4752–67.
15 Kerr J, Anderson C, Lippman SM. Physical activity, sedentary behaviour, diet, and cancer: an update and emerging new evidence. *Lancet Oncol* 2017;18:e457–71.
16 Hur J, Otegbeye E, Joh H-K, *et al*. Sugar-sweetened beverage intake in adulthood and adolescence and risk of early-onset colorectal cancer among women. *Gut* 2021;70:2330–6.
17 Groessl EJ, Allison MA, Larson JC, *et al*. Coffee Consumption and the Incidence of Colorectal Cancer in Women. *J Cancer Epidemiol* 2016;2016:6918431.
18 Coker OO, Liu C, Wu WKK, *et al*. Altered gut metabolites and microbiota interactions are implicated in colorectal carcinogenesis and can be non-invasive diagnostic biomarkers. *Microbiome* 2022;10:35.
19 Hong J, Guo F, Lu S-Y, *et al*. *F. nucleatum* targets lncRNA ENO1-IT1 to promote glycolysis and oncogenesis in colorectal cancer. *Gut* 2021;70:2123–37.
20 Arima K, Zhong R, Ugai T, *et al*. Western-Style Diet, pks Island-Carrying Escherichia coli, and Colorectal Cancer: Analyses From Two Large Prospective Cohort Studies. *Gastroenterology* 2022;163:862–74.
21 Wong SH, Kwong TNY, Chow T-C, *et al*. Quantitation of faecal *Fusobacterium* improves faecal immunochemical test in detecting advanced colorectal neoplasia. *Gut* 2017;66:1441–8.
22 Imperiale TF, Gruber RN, Stump TE, *et al*. Performance Characteristics of Fecal Immunochemical Tests for Colorectal Cancer and Advanced Adenomatous Polyps: A Systematic Review and Meta-analysis. *Ann Intern Med* 2019;170:319–29.
23 Niedermaier T, Weigl K, Hoffmeister M, *et al*. Diagnostic performance of flexible sigmoidoscopy combined with fecal immunochemical test in colorectal cancer screening: meta-analysis and modeling. *Eur J Epidemiol* 2017;32:481–93.
24 A Stool DNA Test (Cologuard) for Colorectal Cancer Screening. *JAMA* 2014;312:2566.