nature communications

Article

https://doi.org/10.1038/s41467-024-52894-2

Plasma proteomic and polygenic profiling improve risk stratification and personalized screening for colorectal cancer

Received: 21 February 2024

Accepted: 20 September 2024

Published online: 15 October 2024

Check for updates

Jing Sun[1,2,9], Yue Liu[1,3,4,5,9], Jianhui Zhao[1,2], Bin Lu[6], Siyun Zhou[1,2], Wei Lu[1,3,4,5], Jingsun Wei[1,3,4,5], Yeting Hu[1,3,4,5], Xiangxing Kong[1,3,4,5], Junshun Gao[7,8], Hong Guan[7,8], Junli Gao[7,8,10], Qian Xiao[1,3,4,5,10] ✉ & Xue Li[1,2,10]

This study aims to identify colorectal cancer (CRC)-related proteomic profiles and develop a prediction model for CRC onset by integrating proteomic profiles with genetic and non-genetic factors (QCancer-15) to improve the risk stratification and estimate of personalized initial screening age. Here, using a two-stage strategy, we prioritize 15 protein biomarkers as predictors to construct a protein risk score (ProS). The risk prediction model integrating proteomic profiles with polygenic risk score (PRS) and QCancer-15 risk score (QCancer-S) shows improved performance (C-statistic: 0.79 vs. 0.71, $P = 4.94E{-}03$ in training cohort; 0.75 vs 0.69, $P = 5.49E{-}04$ in validation cohort) and net benefit than QCancer-S alone. The combined model markedly stratifies the risk of CRC onset. Participants with high ProS, PRS, or combined risk score are proposed to start screening at age 46, 41, or before 40 years old. In this work, the integration of blood proteomics with PRS and QCancer-15 demonstrates improved performance for risk stratification and clinical implication for the derivation of risk-adapted starting ages of CRC screening, which may contribute to the decision-making process for CRC screening.

Colorectal cancer (CRC) is the third most common malignancy and the second cause of cancer death globally[1]. Currently, the most effective way to reduce the burden of CRC remains early detection by population screening. Although the modalities of screening vary internationally, screening programs are mostly based on age and family history and not considered other risk determinants. However, accumulating evidence indicates that substantial variation of CRC risk among populations is not only attributed to age and family history, especially the incidence of early-onset CRC has gradually increased in recent years[2]. Furthermore, given that colonoscopy (the gold standard) is invasive, time consuming, and expensive, further identification of non-invasive early screening and diagnostic biomarkers and development of the risk-based, personalized screening recommendations for improving screening effectiveness is urgently required.

Blood proteins, appearing in circulation due to active secretion or cellular leakage, present a holistic readout of human health states and diseases[3] and act as a major reservoir of biomarkers and therapeutic targets, holding easier accessibility and the most intrinsic predictive

[1]Department of Colorectal Surgery and Oncology (Key Laboratory of Cancer Prevention and Intervention, China National Ministry of Education), The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. [2]Department of Big Data in Health Science, School of Public Health and The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. [3]Center for Medical Research and Innovation in Digestive System Tumors, Ministry of Education, Hangzhou, Zhejiang, China. [4]Zhejiang Provincial Clinical Research Center for CANCER, Hangzhou, Zhejiang, China. [5]Cancer Center of Zhejiang University, Hangzhou, Zhejiang, China. [6]Nursing Department, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. [7]Cosmos Wisdom Mass Spectrometry Center of Zhejiang University Medical School, Hangzhou, Zhejiang, China. [8]Key Laboratory of Precision Medicine in Diagnosis and Monitoring Research of Zhejiang Province, Hangzhou, Zhejiang, China. [9]These authors contributed equally: Jing Sun, Yue Liu. [10]These authors are jointly supervised this work: Junli Gao, Qian Xiao, Xue Li. ✉e-mail: qxiao3@zju.edu.cn

potential for diseases[4]. Several studies have shown the close bond between circulating proteins and CRC risk[5,6] and suggested that protein biomarkers have the potential to discriminate CRC cases from controls[7–9]. However, these proteins-based predictive studies were based on case-control design, with small sample size. Population-based longitudinal study, in theory, is more appropriate to estimate early molecular signatures associated with disease onset for risk stratification. Moreover, genetic and well-established non-genetic factors (e.g., QCancer-[15 yr, colorectal])[10,11] also play an important role in CRC incidence and contribute to risk stratification of CRC[12]. Harnessing circulating proteins, as well as genetic and non-genetic factors simultaneously can be expected to develop predictive tools with better clinical utility in CRC.

In this work, we use a two-stage strategy based on case-control and independent prospective population cohort to identify and validate the CRC-related proteomic profiles (Fig. 1). Then, QCancer-15 risk score and polygenic risk scores (PRSs) are constructed and combined with key protein biomarkers to develop a prediction model for CRC onset. Last, we evaluate the predictive performance of the combined model and potential clinical utility in the aspect of net benefit, risk stratification, and personalized initial screening age for CRC.

Results

Proteomic signatures of CRC

In the discovery stage, a total of 421 differentially expressed plasma proteins were found between CRC cases and controls, with 243

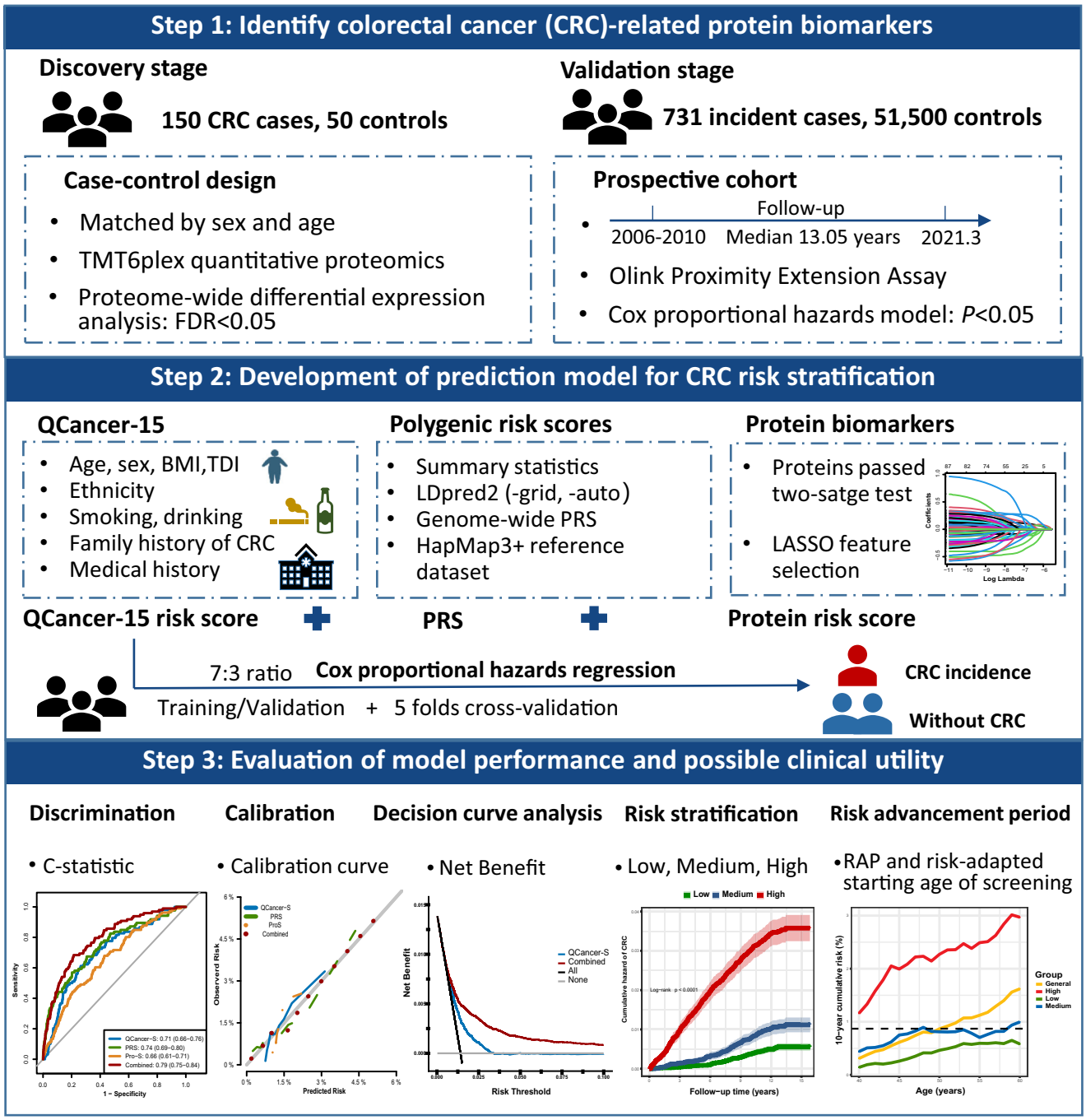

Fig. 1 | Flowchart showing the overall study design. BMI body mass index, TDI Townsend deprivation index, PRS polygenic risk score, RAP risk advancement period.

upregulated proteins (Log2 fold change, $Log_2FC > 0$, false discovery rate, FDR < 0.05) and 178 downregulated proteins ($Log_2FC < 0$, FDR < 0.05) (Supplementary Data 1). In the validation stage (UK Biobank, UKBB), during a median follow-up time of 13.05 (interquartile range, IQR: 12.38–13.75) years, 731 participants were diagnosed with CRC (Table 1). 253 of 421 proteins were also measured in UKBB, and a total of 88 were validated. 75 proteins were positively associated with incident CRC (hazard ratio, HR > 1, $P < 0.05$) and 13 proteins were inversely associated with incident CRC (HR < 1, $P < 0.05$) (Supplementary Data 1).

Construction of risk scores

Of the 88 proteins identified by two-stage strategy, the least absolute shrinkage and selection operator (LASSO)-Cox regression was applied to optimize the selection of protein signatures associated with CRC risk. Consequently, 15 protein biomarkers (IGFBP4, WFDC2, TFF1, LTBP2, COL18A1, CCN3, SELL, PON3, ADGRG2, IGF2R, PZP, CDON, TNXB, APLP1, CR2) were finally selected to construct the protein risk score (ProS) for CRC risk (Fig. 2, Supplementary Fig. 1). The baseline characteristics included in QCancer-15 predictors of UKBB are shown in Table 1. The constructed ProS (HR: 1.24 per SD increment, 95% confidence interval [CI]: 1.21–1.28, $P = 1.85 \times 10^{-48}$) and QCancer-15 risk score (QCancer-S) (HR: 1.55 per SD increment, 95% CI: 1.48–1.62, $P = 1.51 \times 10^{-72}$) were positively associated with incident CRC (Supplementary Data 2). Using LDpred2, a total of 2 PRSs were constructed. Both of the two PRSs were positively associated with incident CRC, and HR per SD of PRS for CRC risk was 1.46 for LDpred2-auto and 2.31 for LDpred2-grid (95% CI: 2.15–2.48, $P = 4.36 \times 10^{-115}$) (Supplementary Data 2).

Development and evaluation of prediction model

The prediction model based on QCancer-S, PRSs, and ProS were constructed in UKBB, respectively. In the training cohort, the C-statistics for both QCancer-S and QCancer-S (including geographical region) was 0.71 (95% CI: 0.66, 0.76), and there was no statistically significant difference ($P = 0.63$) (Fig. 3a, Supplementary Data 2). For generalizability, the QCancer-S was used in subsequent modeling and analyses. The C-statistics for ProS was 0.66 (95% CI: 0.61, 0.71) (Fig. 3a). Of the 2 PRS models, LDpred2-grid performed best, with a C-statistic of 0.74 (95% CI: 0.69, 0.80) (Supplementary Data 2). Then the PRS (LDpred2-grid) with the best performance was combined with QCancer-S and ProS to construct a combined model, and the C-statistic reached 0.79 (95% CI: 0.75, 0.84), which was significantly improved than the QCancer-S alone model in both training (C-statistic: 0.79 vs. 0.71, $P = 4.94E{-}03$) and validation (C-statistic: 0.75 vs 0.69, $P = 5.49E{-}04$) cohorts (Fig. 3a, Supplementary Data 2).

The results of stratified analyses are shown in Supplementary Data 3. In the training cohort, the C-statistic of combined model was 0.76 (95% CI: 0.69, 0.83) for females and 0.77 (95% CI: 0.70, 0.84) for males. When considering CRC risk separately for colon and rectum, the C-statistic of the combined model reached 0.79 (95% CI: 0.74, 0.85) for colon cancer and 0.82 (95% CI: 0.73, 0.91) for rectal cancer. The combined model tended to perform better than QCancer-S alone regardless of whether the participants had previously been screened for bowel cancer or a family history of CRC. The C-statistic of the combined model reached 0.79 (95% CI: 0.73, 0.85) for participants without screening, 0.79 (95% CI: 0.62, 0.95) for participants who had previously been screened before they were recruited into the UKBB one year and earlier, and 0.91 (95% CI: 0.85, 0.98) for participants with family history. Similar results were also observed in the validation cohort. For participants without a family history of CRC and no screening, the combined model still showed improved discrimination than QCancer-S alone in both training (C-statistic, 0.82 vs. 0.71, $P < 0.05$) and validation (C-statistic, 0.75 vs. 0.71, $P < 0.05$) cohorts. Calibration curves indicated that the observed and predicted rates had good concordance (Fig. 3b, Supplementary Fig. 2).

Table 1 | Baseline characteristics of participants

Characteristic	All participants	Incident CRC cases
Number of participants	52,231	731
Follow-up (years), median (IQR)	13.05 (12.38, 13.75)	6.69 (3.49, 9.34)
Sex		
Female	28,196 (54%)	305 (42%)
Male	24,035 (46%)	426 (58%)
Age (years), median (IQR)	58 (50, 64)	62 (57, 66)
Ethnicity		
White/not recorded	48,951 (94%)	702 (96%)
Black African	684 (1.3%)	8 (1.1%)
Caribbean	502 (1.0%)	3 (0.4%)
Indian	581 (1.1%)	2 (0.3%)
Pakistani	172 (0.3%)	3 (0.4%)
Bangladeshi	29 (<0.1%)	0 (0%)
Other Asian	192 (0.4%)	3 (0.4%)
Chinese	147 (0.3%)	4 (0.5%)
Other	973 (1.9%)	6 (0.8%)
Townsend deprivation index, median (IQR)	−2.1 (−3.6, 0.8)	−2.1 (−3.5, 0.6)
Body mass index (BMI), median (IQR)	26.8 (24.2, 29.9)	27.3 (24.8, 29.9)
Smoking		
Non-smoker	28,304 (54%)	344 (47%)
Ex-smoker	18,160 (35%)	319 (44%)
Light smoker	2258 (4.3%)	23 (3.1%)
Moderate smoker	1617 (3.1%)	16 (2.2%)
Heavy smoker	1379 (2.6%)	20 (2.7%)
Missing	513 (1.0%)	9 (1.2%)
Alcohol intake		
Non-drinker	4504 (8.6%)	61 (8.3%)
Trivial drinker	6189 (12%)	79 (11%)
Light drinker	15,584 (30%)	203 (28%)
Moderate drinker	10,514 (20%)	168 (23%)
Heavy drinker	1946 (3.7%)	37 (5.1%)
Very heavy drinker	1285 (2.5%)	35 (4.8%)
Missing	12,209 (23%)	148 (20%)
Family history of colorectal cancer		
No	47,614 (91%)	633 (87%)
Yes	3569 (6.8%)	81 (11%)
Missing	1048 (2.0%)	17 (2.3%)
Medical history		
Diabetes	742 (1.4%)	13 (1.8%)
Ulcerative colitis	114 (0.2%)	3 (0.4%)
Bowel polyps	26 (<0.1%)	0 (0%)
Breast cancer	1105 (2.1%)	17 (2.3%)
Uterine cancer	113 (0.2%)	2 (0.3%)
Ovarian cancer	80 (0.2%)	1 (0.1%)
Cervical cancer	220 (0.4%)	0 (0%)
Lung cancer	46 (<0.1%)	1 (0.1%)
Blood cancers	263 (0.5%)	6 (0.8%)
Oral cancers	99 (0.2%)	3 (0.4%)

CRC colorectal cancer, *IQR* interquartile range.

Net benefit of prediction model

Decision curve analyses showed that the combined model had greater net benefit than QCancer-S alone model across a range of risk probabilities in both training and validation cohorts, as well as whole population (Fig. 3c, Supplementary Fig. 3). For instance, when the risk probability was 1%, the net benefit for combined model was 0.65% in whole population (0.62% in training cohort, 0.71% validation cohort),

https://doi.org/10.1038/s41467-024-52894-2

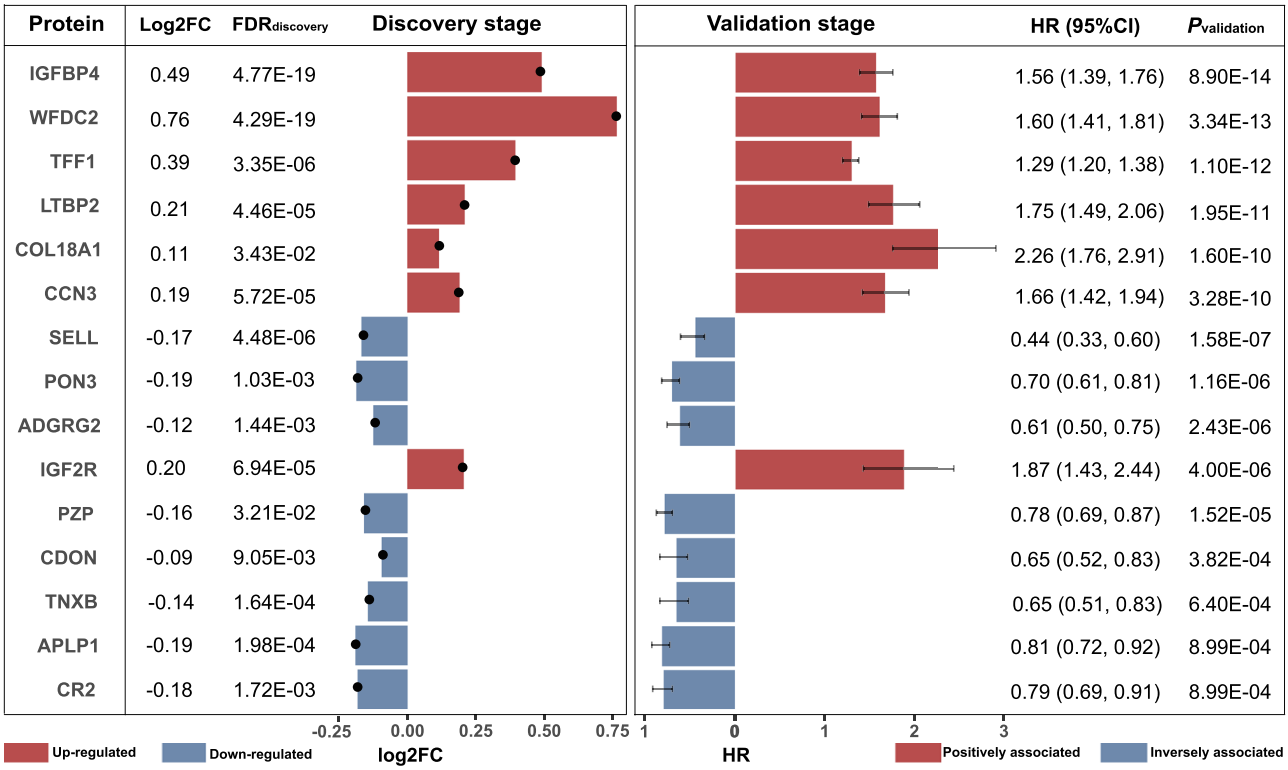

Fig. 2 | The results of 15 plasma proteins that were used to construct protein risk score (ProS) with colorectal cancer from proteome-wide differential expression analysis (discovery stage) and Cox proportional hazards regression analysis (validation stage). The *P*-value was derived corresponding to a two-sided test, and false discovery rate (FDR) was used for multiple testing correction. In the validation stage, the centers of the lines represent estimated hazard ratios, and the lines represent the 95% confidence intervals. Source data are provided as a Source Data file. Log_2FC Log_2 fold change, FDR false discovery rate, HR hazard ratio, CI confidence interval.

meaning that compared with assuming that none of participants were intervened, 65 (62 in training cohort, 71 in validation cohort) net detect CRCs without an increase of false-positives and unnecessary colonoscopies per 10,000 participants. The net benefit for the combined model was 0.24% (0.24% in training cohort, 0.21% in validation cohort) greater than assuming that all participants were intervened and 0.10% (0.11% in training cohort, 0.06% in validation cohort) greater than the QCancer-S model (Supplementary Data 4). Furthermore, there were 24 (24 in training cohort, 21 in validation cohort) fewer false-positives for the combined model whereas only 14 (13 in training cohort, 15 in validation cohort) fewer false-positives for the QCancer-S model per 100 individuals. In stratification by sex or tumor site, the combined model still performed better than QCancer-S alone (Supplementary Fig. 3). In participants without a family history of CRC and without screening, these values were 26 fewer false-positives for the combined models whereas only 16 fewer false-positives for the QCancer-S model per 100 individuals (Supplementary Fig. 3, Supplementary Data 4).

Risk stratification for CRC across risk groups

The combined model was then used to assess the risk of CRC for participants, and a nomogram based on the information was established across the whole population, training cohort, and validation cohort as shown in Fig. 3d and Supplementary Fig. 4. After further dividing the components of combined model (QCancer-S, PRS, and ProS) and combined risk score (ComS) into three groups (low, medium, and high), respectively, compared with medium group, participants in high QCancer-S group had more than 1.9-fold increased risk of CRC (HR: 1.91, 95% CI: 1.60–2.29) among whole population (HR [95% CI]: 2.06 [1.65–2.56] in training cohort, 1.66 [1.23–2.25] in validation cohort), 2.5-fold increased risk of CRC for high PRS group (HR: 2.71, 95% CI: 2.24–3.27) among whole population (HR [95% CI]: 2.84 [2.26–3.59] in training cohort, 2.46 [1.78–3.40] in validation cohort), 1.5-fold increased risk of CRC for high ProS group (HR: 1.67, 95% CI: 1.40–1.99) among whole population (HR [95% CI]: 1.78 [1.43–2.22] in training cohort, 1.48 [1.09–1.99] in validation cohort), and threefold increased risk of CRC for high ComS group (HR: 3.27, 95% CI: 2.71–3.96) among whole population (HR [95% CI]: 2.94 [2.35–3.67] in training cohort, 4.27 [2.96–6.18] in validation cohort) (Supplementary Data 5). Figure 4a–d and Supplementary Fig. 5 show Kaplan–Meier cumulative incidence curves across three risk groups based on four risk scores among whole population, training cohort, and validation cohort, respectively, showing distinct separation between different risk groups (Log-rank $P < 0.0001$) and the risk groups based on ComS showing improved separation. In participants without a family history of CRC and without screening, the ComS also performed better in risk stratification (Supplementary Data 5, Supplementary Fig. 6).

Risk advancement period of CRC across different stratification groups

Furthermore, participants in high PRS group reached the equivalent CRC risk ~17.9 years (risk advancement period, RAP: 17.87, 95% CI: 12.91, 22.83) earlier than participants with PRS in the medium group among whole population, whereas participants in low PRS group were 8.9 years later than medium group (RAP: −8.93, 95% CI: −13.34, −4.52) (Table 2). Compared with the medium ProS group, participants in the high ProS group reached the equivalent CRC risk of ~4.7 years (RAP: 4.67, 95% CI: 1.08, 8.27) earlier, whereas 4.4 years (RAP: −4.42, 95% CI: −8.17, −0.66) later for participants in the low ProS group. When compared with the medium ComS group, participants in high ComS group even reached the equivalent CRC risk 27.5 years (RAP: 27.52, 95% CI: 16.66, 38.37) earlier, while ~15.2 years (RAP: −15.15, 95% CI: −23.78, −6.52) later for participants in the low ComS group. The RAP estimates

https://doi.org/10.1038/s41467-024-52894-2

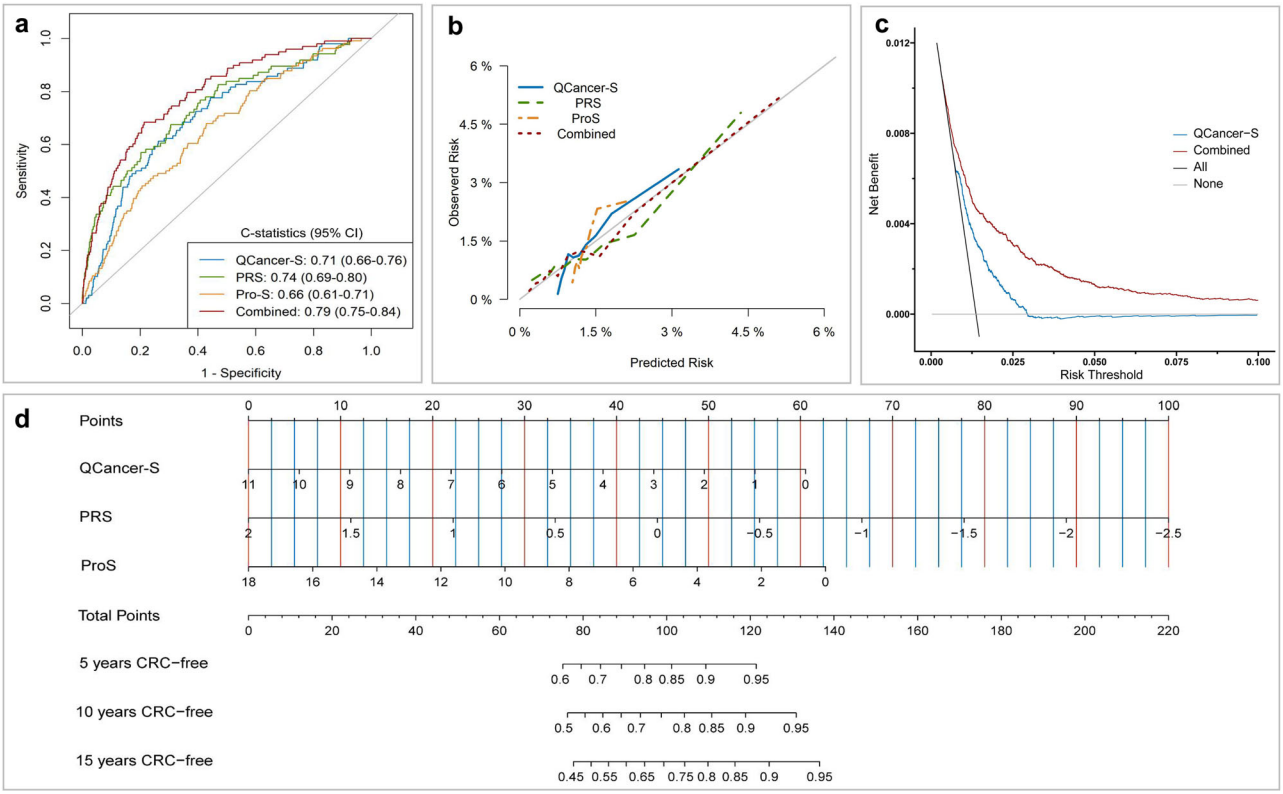

Fig. 3 | Assessment of the prediction models of the QCancer-15 risk score (QCancer-S)-based, polygenic risk score (PRS)-based, protein risk score (ProS)-based, and Combined model for colorectal cancer (CRC). a Receiver operating characteristic (ROC) curves for QCancer-S, PRS, ProS, and Combined model in training cohort. **b** The calibration curves of four models predict the probability of CRC in training cohort. **c** Decision curve analysis for QCancer-S and Combined model in training cohort, calculated at median follow-up time (13.05 years). The gray line in net benefit curves indicates no intervention, and the black line indicates intervention for all. **d** Nomogram for the prediction of the 5-, 10-, and 15-year being free of CRC in whole population. Source data are provided as a Source Data file.

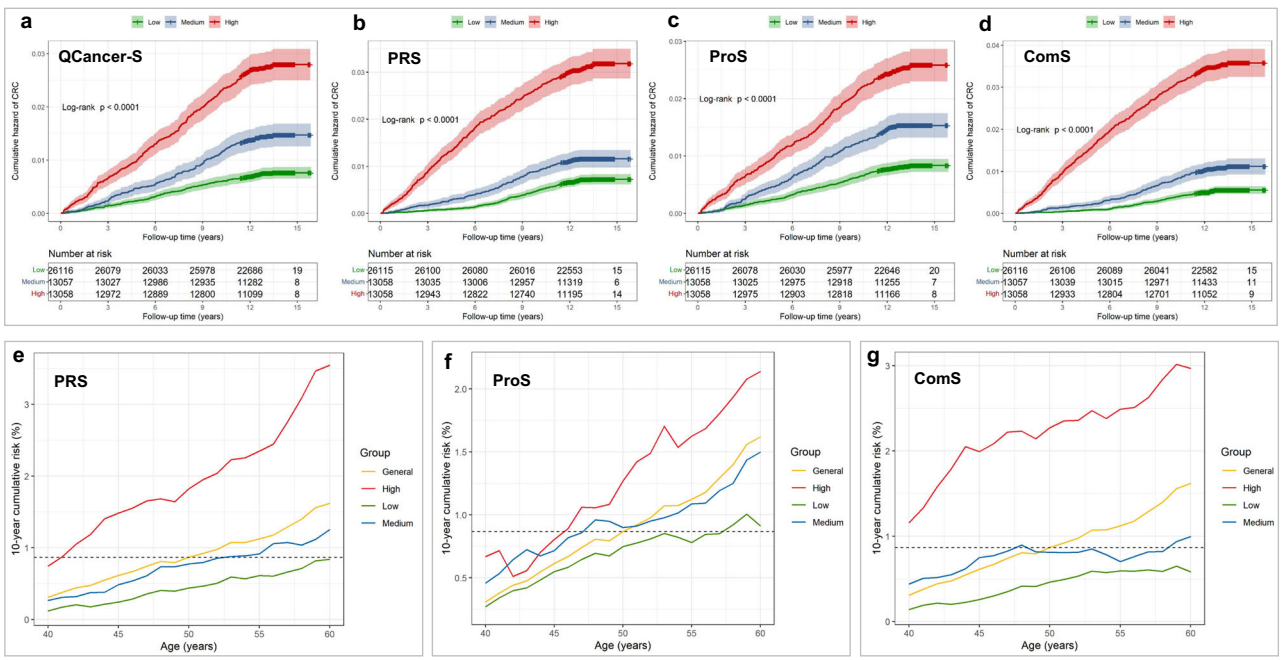

Fig. 4 | Risk stratification and risk-adapted starting age of colorectal cancer (CRC) screening in whole population across risk groups. Kaplan–Meier cumulative incidence curves across risk groups based on (**a**) QCancer-S (log-rank $p = 4.13E{-}54$), (**b**) polygenic risk score (PRS) (log-rank $p = 3.98E{-}79$), (**c**) protein risk score (ProS) (log-rank $p = 5.08E{-}40$), and (**d**) combined risk score (ComS) (log-rank $p = 1.82E{-}121$). The shadow bands represent confidence intervals. Age-specific 10-year cumulative risk of CRC for participants across risk groups based on (**e**) PRS, (**f**) ProS, and (**g**) ComS. The dashed line in age-specific 10-year cumulative risk curves indicates the risk at age 50 years (the starting age of screening for average-risk adults recommended by the current guidelines) in the general population. Source data are provided as a Source Data file.

Table 2 | Risk advancement period (RAP) and risk-adapted starting age of colorectal cancer (CRC) screening in participants across different risk levels

Risk score	Case	Control	RAP (95% CI) (years)	Risk-adapted starting age (years) of screening
PRS				
Low	183	25,932	−8.93 (−13.34, −4.52)	>60
Medium	149	12,909	ref	53
High	399	12,659	17.87 (12.91, 22.83)	41
ProS				
Low	210	25,905	−4.42 (−8.17, −0.66)	57
Medium	196	12,862	ref	47
High	325	12,733	4.67 (1.08, 8.27)	46
ComS				
Low	140	25,976	−15.15 (−23.78, −6.52)	>60
Medium	140	12,917	ref	48
High	451	12,607	27.52 (16.66, 38.37)	<40

RAP risk advancement period, *CI* confidence interval, *PRS* polygenic risk score, *ProS* protein risk score, *ComS* combined risk score.

and 95% CI across different risk groups in training and validation cohorts are shown in Supplementary Data 6. When the analysis was limited to participants without a family history of CRC and without screening, similar results were observed (Supplementary Data 7).

Risk-adapted starting age of CRC screening

The 10-year cumulative risk was calculated subsequently to estimating the personalized initial screening age for whole population of different risk groups. The ages of participants with low-, medium-, or high-PRS reached 10-year cumulative CRC risk equal to the average risk at age 50 years old in the general population were >60, 53, and 41 years old (Fig. 4e, Table 2). The ages of participants with low-, medium-, or high-ProS reached 10-year cumulative CRC risk equal to the average risk of general population at age 50 years old were 57, 47, and 46 years old (Fig. 4f, Table 2). When integration of ProS, PRS, and QCancer-S, the ages of participants with low-, medium-, or high-ComS reached 10-year cumulative CRC risk equal to the average risk at age 50 years old in the general population were >60, 48, and <40 years old (Fig. 4g, Table 2). Similar results were also observed in participants without a family history of CRC and without screening (Supplementary Fig. 7, Supplementary Data 7).

Discussion

Using a two-stage strategy, we identified CRC-related circulating proteomic profiles and prioritized 15 protein biomarkers for risk stratification. The prediction model integrating proteomic profiles with genetic (i.e., PRS) and non-genetic factors (i.e., QCancer-15) performed better across all metrics than QCancer-15 alone. In addition, the combined model markedly stratified the risk of CRC onset and derived risk-adapted starting ages for CRC screening. Compared with the medium risk group, participants with high ProS, PRS, or ComS reached the equivalent CRC risk of 4.7, 17.9, or even 27.5 years earlier, while 4.4, 8.9 or 15.2 years later in the low ProS, PRS, or ComS group. When referencing the general population aged 50 years old, participants with high ProS, PRS, or ComS were proposed to start screening at age 46, 41, or before 40 years old, while the corresponding low-score groups were proposed to start screening at age 57 or after 60 years. Similar results were observed in participants without a family history of CRC and no screening.

The two-stage study design identified and prioritized 15 circulating protein biomarkers associated with CRC incidence. Most of them have reported evidence to be associated with either CRC risk, prognosis, or response to the therapy in either protein level (i.e., TFF1, LTBP2, PZP)[13–15], mRNA expression (i.e., TFF1, LTBP2, WFDC2, CDON)[14,16–18], DNA methylation (i.e., PON3)[19], or gene polymorphisms (i.e., CR2)[20]. TFF1 (Trefoil factor 1), belongs to the members of the trefoil family, which are stable secretory proteins expressed in gastrointestinal mucosa. In line with our findings, previous studies have reported that serum TFF1 levels were significantly higher in CRC patients than healthy controls[13], and the mRNA level of TFF1 was elevated in CRC tissues than the adjacent normal tissues and promoted the malignant behavior of colon cancer via activation of the epithelial-mesenchymal transition process[16]. LTBP2 (Latent-transforming growth factor beta-binding protein 2) is a member of the fibrillin/LTBP superfamily. Both mRNA and protein levels of LTBP2 in CRC tissues were reported to be remarkably higher than those in adjacent normal tissues, and high LTBP2 protein level was linked to poor overall survival in CRC patients, suggesting that LTBP2 might act as an oncogene in CRC development and an important biomarker for predicting CRC prognosis[14]. We also indicated a protective effect of plasma PZP (Pregnancy zone protein) on the risk of CRC onset. Similarly, Wang et al. reported that serum PZP protein level was related to the response to neoadjuvant chemoradiation (nCRT), and rectal cancer patients with a higher level of PZP tended to pathological complete response, with potential predictive value for rectal cancer with nCRT[15].

A proactive approach on the basis of individualized risk prediction has been described as the future of early detection for cancers, highlighting multifactorial cancer risk assessment of epidemiological factors, inherited genetic variants, and omics biomarkers (especially protein biomarkers)[21]. Previous studies have developed prediction models for CRC risk based on only PRS or non-genetic predictors[22,23] or combining both and observed that models integrating PRS and non-genetic predictors had better performance than non-genetic alone or PRS alone[12,24]. Specifically, our previous study based on 116 GWAS significant SNPs produced a C-statistic of 0.6[23]. Thomas et al. reported a C-statistic of 0.65 for CRC risk based on a cross-ancestry PRS alone model in non-Hispanic White[22]. Briggs et al. integrated the QCancer and PRS to produce a C-statistic of 0.73 in predicting CRC risk for males and 0.69 for females[12], and another study reported a C-statistic of 0.72 after combining PRS and non-genetic predictors[24]. Furthermore, a case-control study in the American population has shown improved discrimination after adding 11 plasma biomarkers reported by previous studies based on candidate strategy into a non-genetic risk factors model, with a C-statistic of 0.66–0.73[9]. Consistently, our work supported and extended these findings. Even though there were differences in protein coverage (candidate strategy vs. proteome), study designs, and study population, the protein (IGFBP3) included in their model and the protein (IGFBP4) included in our model both belong to the insulin-like growth factor-binding protein family, which indicated the generalizability in some degree. By integrating proteomic profiles with PRS and QCancer-15, we constructed a combined model with a

C-statistic of 0.79 for prediction of the risk of CRC incidence in a prospective cohort, which showed better performance than the QCancer-15 or PRS alone models in both training and validation cohorts. In addition, the results from the decision curve analysis and Kaplan–Meier curves showed an increased net benefit of the combined model to the QCancer-15 alone model and suggested the potential clinical utility in risk stratification for CRC.

The risk-adapted starting age of screening has important clinical and public health significance for individualized prevention of cancer. A previous study based on the family history of CRC estimated that participants with different family histories reached a 10-year cumulative risk of CRC equal to the risk at age 50 years old in the average-risk population 3–29 years earlier[25]. Another study focusing on sex and PRS using the approach of RAP estimated that individuals in the lowest or highest PRS decile reached an equivalent risk of CRC onset at 8–10 years later or earlier than individuals in middle PRS deciles, and males in the highest PRS decile reached comparable risk at 16 years earlier than females in middle PRS deciles[26]. Consistently, our study employed both RAP and absolute risk (i.e., 10-year cumulative risk) methods and obtained similar estimates for risk-adapted starting ages of CRC screening. We found that participants in high PRS group reached the equivalent CRC risk 17.9 years earlier than participants in the medium PRS group, whereas participants in low PRS group were 8.9 years later than medium group. Similarly, when using the average-risk adults at 50 years old which is the screening age recommended by the current guidelines as ref. 27, the age of participants with high- or low-PRS reached 10-year cumulative CRC risk equal to comparable risk was 41 or >60 years old. Additionally, we extended these approaches to proteomics and estimated the risk-adapted starting age of CRC screening for participants with ProS in high or low group to be 46 or 57 years old. When integrating QCancer-15, PRS, and ProS, the ages reached equivalent risk across different risk groups showed greater separation. Participants in high ComS group reached the equivalent CRC risk 27.5 years earlier than participants in medium group, while 15.2 years later for participants in low ComS group than medium group. The risk-adapted starting age of CRC screening for participants in high- or low-risk group was estimated to be before 40 or after 60 years old. Similar results were observed in participants without a family history of CRC and no screening who usually have a relatively lower risk and are easily overlooked. Overall, these findings supported strong evidence for promoting proteomics in clinical practice, which could contribute to improving personalized risk assessment and screening of CRC and the development of precision medicine for CRC.

This study has some strengths. First, a two-stage design based on high-throughput proteomic technologies was used to identify the CRC-related proteomic profiles, which ensured that the proteins with stable associations with CRC across different populations and proteomic platforms were used for modeling, improving the generalizability of these findings. Second, the community-based prospective cohort with a large sample size, good representation, and long-term follow-up allowed us to build a population-based prediction model for CRC onset. Third, integrating the circulating proteomic profiles with genetic and well-established non-genetic factors provided a holistic readout of human health states, which contributed to the more precise and comprehensive risk stratification for CRC to help focus limited resources on individuals with high risk. Fourth, RAP and 10-year cumulative risk provided similar results for the recommendation of risk-adapted starting ages of CRC screening, which implied the stability and reliability of the current findings. Additionally, the parameters needed for the model building can be obtained from self-reported and a single non-invasive biospecimen (i.e., blood), which ensured the feasibility and acceptance. Nevertheless, several limitations of this study require careful consideration. First, although the CRC-related proteomic biomarkers were identified by a two-stage design based on both Chinese and UKBB populations, the prediction model was established in the UKBB (including European and non-European ancestry participants) with only internal verification. Further research based on independent external populations was needed to assess the generalizability before applying the current findings in routine care. Second, the RAP and 10-year cumulative risk could not be used in assessing risk-adapted starting age of CRC screening across different QCancer-S groups due to the age information itself included and predominated in QCancer-S. Further research to evaluate the contribution of well-established non-genetic factors (excluding age) in risk-adapted starting age of CRC screening may be interesting. Additionally, although most of the estimates of RAPs were similar between the training cohort and validation cohort, the RAP estimates for several risk score groups were not significant in the validation cohort, which may be partly attributed to the small number of cases. Lastly, not only the initial screening age but also screening intervals are included in screening strategies. The screening intervals for risk-adapted CRC screening need to be investigated by future studies.

In conclusion, the integration of blood proteomics with PRS and QCancer-15 demonstrated improved performance for risk stratification and clinical implication for the derivation of risk-adapted starting ages of CRC screening. This may contribute to the decision-making process for CRC screening, with the potential to reduce the screening burden of people with lower risk, while improving screening detection rates. Future research to evaluate the acceptance, feasibility, and cost-effectiveness of the combined model and risk-adapted screening age is needed.

Methods

Ethics approval and consent to participate

Ethics approval for the discovery stage of this study was obtained from the Ethics Committees of Second Affiliated Hospital of Zhejiang University School of Medicine (2023-1190). The UKBB was approved by North West Multicenter Research Ethics Committee (11/NW/0382). All participants provided written informed consent, and this study was performed in accordance with the Declaration of Helsinki.

Study design and participants

A flowchart shows the overall study design (Fig. 1). Briefly, using a two-stage strategy, we first explored CRC-related protein biomarkers by proteome-wide differential expression analysis based on a case-control design and validated them in an independent prospective population cohort. Then, the ProS, QCancer-S, and PRSs were constructed and integrated to develop a prediction model for the risk of CRC onset based on the prospective population cohort. Last, the combined model was evaluated in predictive performance and possible clinical utility (e.g., net benefit, risk stratification, and personalized initial screening age).

The discovery stage comprised 150 newly diagnosed CRC cases (age range: 44–89 years old) and 50 sex- and age-matched controls from the Second Affiliated Hospital of Zhejiang University School of Medicine. The validation cohort comprised CRC incident cases and controls from UKBB, recruited from 2006 to 2010. Participants with proteomics measurements were selected. After excluding individuals with CRC at recruitment and missing genotype data, a total of 52,231 individuals (731 CRC incident cases, 51,500 controls) aged 39–70 years old from UKBB remained.

Definition of outcome

The primary outcome was CRC diagnosis, which was defined as malignant neoplasms of the colon and rectum, diagnosed by a clinician and pathologist in the discovery cohort, or using the International Classification of Diseases (ICD, ICD-9: 153, 154.0, 154.1; or ICD-10: C18-C20) in UKBB through linkage to hospital data and cancer and death registries. The date of enrolment for a participant to UKBB be considered as the start of the follow-up period, and follow-up was

censored at the date of incident CRC, death, loss to follow-up, or end of available registry follow-up (March 31, 2021), whichever came first.

Plasma proteomics measurement

Blood samples were collected from participants at enrolment, and plasma samples were prepared and stored at −80 °C. Protein measurements were generated using the iodo Tandem Mass Tags (TMT)-6plex quantitative proteomics in the discovery cohort. The peptide segments were quantitative by the ratio of CRC sample channels to the channels of the control group and were Log_2 transformation and normalized using the quantiles method. Details on proteomic measurement, processing, and quality control (QC) are described in Supplementary Methods. After QC, a total of 606 proteins were identified and included in subsequent analyses (Supplementary Data 8).

In UKBB, details on sample selection, proteomic measurement, processing, and QC have been described elsewhere[28]. Briefly, the relative abundance of 2923 proteins was quantified by Olink Proximity Extension Assay, and measurements were expressed as normalized protein expression values by Log_2 transforming. Protein values less than the limit of detection (LOD) were replaced by dividing the LOD by the square root of 2. The missing measurements of protein were imputed by mean values. For proteins with differential expression between CRC cases and controls in discovery stage, we further verified the relationships of them with CRC risk based on prospective cohort design in UKBB.

Statistical analysis

Baseline characteristics description. The Kolmogorov–Smirnov normality test was used to identify the distribution types of continuous variables. Median (IQR) was utilized to describe continuous variables, and number (percentage) was used to describe categorical variables.

Identification of CRC-related protein biomarkers

A two-stage strategy was adopted to identify CRC-related proteomic signatures. In the discovery stage based on case-control design, principal component analysis and Pearson correlation coefficient matrix were used to identify outlier samples (Supplementary Methods), and 20 samples with quantitative anomalies were identified (Supplementary Fig. 8). After removing outlier samples, proteome-wide differential expression analysis based on 'limma' package[29] was performed to identify upregulated ($Log_2FC > 0$) or downregulated ($Log_2FC < 0$) proteins between CRC cases and controls. The FDR was used for multiple testing correction, with FDR < 0.05 as the significance level. For significant proteins (FDR < 0.05) in discovery stage, we further employed Cox proportional hazards (CPH) models to assess the relationships of them with CRC incidence in validation cohort. The protein with a consistent direction of effect and significant P-value ($P < 0.05$) in the validation stage was considered successful verification.

Definition of QCancer-15

The QCancer-15 predictors were coded based on baseline characteristics in UKBB matched as closely as possible to the original model[11,12]. QCancer-15 predictors included age (age at recruitment), sex, Townsend deprivation index, geographical region, body mass index (BMI), smoking status, alcohol intake, family history of CRC, and previous medical history. Individuals with missing data in these variables were assigned to a "Missing" category for each respective variable. Details in the coding of these predictors are shown in Supplementary Data 9. For generalizability in other populations, we excluded the geographical region variable that was tailored to the UK population from QCancer-15 predictors.

Approaches for deriving polygenic risk score (PRS)

The largest trans-ancestry CRC meta-GWASs of European ancestry (21,731 cases and 47,444 controls of East Asian ancestry; 78,473 cases and 107,143 controls of European ancestry)[30] to date were employed to provide the association effect size of each single nucleotide polymorphism (SNP). LDpred2 was used to construct genome-wide PRS, which is based on a Bayesian approach and accounts for linkage disequilibrium (LD) between the SNPs[31]. Two different LDpred2 models were performed, including auto model (LDpred2-auto) and grid model (LDpred2-grid), using HapMap3+ dataset as reference. A total of 2 PRSs were developed. Details are described in Supplementary Methods.

Construction of risk scores and development of prediction model

The CRC-related protein signatures were further selected by LASSO-Cox regression among proteins that passed the two-stage test, and a total of 15 proteins remained. A weighted ProS or QCancer-S (risk Score=h0(t)*exp(β1×1 + β2×2 + …+βnXn)) was constructed based on the 15 selected proteins or QCancer-15 predictors in UKBB by employing the 'predict' function of R package 'caret' (version 6.0.94)[32], respectively. Where Xn was the level of protein (or QCancer-15 predictor) and βn was the coefficient of protein (or QCancer-15 predictor) associated with CRC risk derived from a CPH model, and the h0(t) was the baseline hazard. Details are described in Supplementary Methods. Then, UKBB participants were randomly split into training and validation cohorts with a 7:3 ratio by using 'caret' package. The prediction models for CRC risk were developed based on QCancer-S, QCancer-S (including geographical region), 2 PRSs, or ProS, respectively, using the CPH model with fivefold cross-validation in the training cohort. Next, the QCancer-S, ProS, and PRS with the best performance were combined to construct the combined model for predicting the risk of CRC onset. These models were also evaluated by the stratification of sex, tumor sites (colon or rectum), bowel cancer screening (yes or no), and family history of CRC (yes or no).

Evaluation of model performance and potential clinical utility

The model performance was evaluated using a wide range of metrics, including the area under the receiver-operating characteristics curve (AUC, known as C-statistic), relative risk calibration, net benefit, hazard ratio estimates, risk stratification, risk advancement period, and 10-year cumulative risk for developing CRC.

Discrimination and calibration of model

The discrimination of these models was assessed by C-statistic (95% confidence interval, CI) with fivefold cross-validation in the training cohort and was further assessed in the validation cohort. The differences in the performance of different models were compared by using bootstrap method, with 500 stratified bootstrap replicates. Calibration curves were drawn using 'riskRegression' package[33] to visually depict the concordance between observed event rates and predicted risks based on CPH models. The nomogram of the QCancer-S, PRS, and ProS was set up for the prediction of the 5-, 10-, and 15-year being free of CRC.

Decision curve analysis

To assess the potential clinical utility of the prediction models on recommended interventions for CRC (e.g., screening), the decision curve analysis was performed using 'ggDCA' package[34] to calculate the net benefit (NB) obtained using QCancer-S alone or combined model (QCancer-S + PRS + ProS) to select individuals for screening colonoscopy. $NB = ((\text{true positives})/n) - ((\text{false positives})/n) * (P_t/(1-P_t))$, where n is the total number of individuals, and P_t is the risk (or probability) threshold (e.g., at $P_t = 1\%$, it is willing to perform colonoscopy for 100 participants to detect one CRC case)[34]. We compared the combined model with the model based on QCancer-S alone, as well as intervention for none and intervention for all. The reduction in the number of false positives per 100 individuals was estimated by the formula: 100*(NB of the model−NB of intervention all)/(P_t/(1−P_t)).

The NBs across relevant risk thresholds over median follow-up time (13.05 years) were plotted in decision curves. The values for NB at a range of prespecified risk thresholds (1%, 1.5%, 2%, 2.5%, and 3%) were reported.

Hazards ratios estimate and risk stratification analysis

HRs were first derived from CPH regression for per SD increase in QCancer-S, PRS, and ProS for overall, sex-specific, and site-specific CRC risk. Then, QCancer-S, PRS, ProS, and ComS (Supplementary Methods) were divided into quartiles (Q1-lowest, Q2, Q3, and Q4-highest) based on their distributions in the present UKBB population, respectively. Because of the similar trends in the 10-year cumulative risk curves, risk scores were further classified into 3 groups (low, medium, and high) (Supplementary Fig. 9). Specifically, using the 10-year cumulative risk in the general population as a reference, participants in Q3 with a similar risk to the general population were classified into the medium-risk group, participants in Q1 and Q2 with lower risk than the general population were classified into low-risk group, and participants in Q4 were classified into high-risk group. The HRs (95% CIs) were also estimated for the categorical variables using the medium group as the reference. The risk stratification value of the risk scores was assessed by Kaplan–Meier curves with log-rank tests.

Risk advancement period (RAP) analysis

To translate the HRs for PRS, ProS, and ComS into how many years of age later or earlier for individuals in the low or high PRS, ProS, and ComS groups would reach risks comparable to the reference group (i.e., those in the medium PRS, ProS, or ComS level), the RAP[35] analysis was performed. The QCancer-S was not assessed by RAP due to the age information itself included and predominated in QCancer-S. The point estimates of RAPs were estimated from the multivariable CPH model, which included age, PRS (low, medium, high), and ProS (low, medium, high), and sex and the first 10 genetic principal components (PCs) as covariates, by dividing the regression coefficient of the PRS or ProS by the coefficient of age. When estimating the RAP of ComS, the model included age, ComS (low, medium, high), sex, and the first 10 PCs, in which RAP was estimated by dividing the regression coefficient of ComS by the coefficient of age. The 95% CIs for RAPs were calculated based on bootstrap method, with 500 stratified bootstrap replicates.

Estimating the risk-adapted starting age of screening

For further estimating the risk-adapted starting age of screening for populations of different risk groups, the 10-year cumulative risk was calculated. Considering the whole UKBB population to be more informative and representative, we estimate the risk-adapted starting age of screening based on the whole population. 10-year cumulative risk = 1−exp(−10-year cumulative incidence rate), where the 10-year cumulative incidence rate equaled the sum of each subsequent 10 years age-specific annual incidence rate of each age[36]. The age-specific annual incidence rate was calculated by dividing the number of cases for each age by person-years for that age. Given that starting age of screening is 50 years old for average-risk adults recommended by the current guidelines[27], we defined the risk-adapted starting age of screening as the age at which individuals with a particular risk of CRC reached a similar level of 10-year cumulative risk to the general population of 50 years old[36]. All statistical tests were two-sided and were conducted using R version 4.2.2.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw data of plasma proteomics generated in this study have been deposited in the iProX database under accession code IPX0008947000 (the equivalent ProteomeXChange PXD code: PXD054569) and IPX0008947001. The UK Biobank data are available under restricted access for data protocol, access can be obtained by registering and applying at http://ukbiobank.ac.uk/register-apply/. This study used the UK Biobank data under application number 66354. The result data generated in this study are provided in the Supplementary Information/Source Data file. The CRC GWAS summary statistics[30] are available through the GWAS catalog (accession no. GCST90129505). Source data are provided with this paper.

Code availability

Analysis code is available at https://github.com/shuaidexue/Proteomics-predicted-colorectal-cancer-risk.

References

1. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Sinicrope, F. A. Increasing incidence of early-onset colorectal cancer. *N. Engl. J. Med.* **386**, 1547–1558 (2022).
3. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell Proteom.* **1**, 845–867 (2002).
4. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).
5. Sun, J. et al. Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome. *Genome Med.* **15**, 75 (2023).
6. Sun, X. et al. Prospective proteomic study identifies potential circulating protein biomarkers for colorectal cancer risk. *Cancers* **14**, 3261 (2022).
7. Ahn, S. B. et al. Potential early clinical stage colorectal cancer diagnosis using a proteomics blood test panel. *Clin. Proteom.* **16**, 34 (2019).
8. Harlid, S. et al. A two-tiered targeted proteomics approach to identify pre-diagnostic biomarkers of colorectal cancer risk. *Sci. Rep.* **11**, 5151 (2021).
9. Fang, Z. et al. Risk prediction models for colorectal cancer: evaluating the discrimination due to added biomarkers. *Int. J. Cancer* **149**, 1021–1030 (2021).
10. Usher-Smith, J. A. et al. External validation of risk prediction models for incident colorectal cancer using UK Biobank. *Br. J. Cancer* **118**, 750–759 (2018).
11. 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* **5**, e007825 (2015).
12. Briggs, S. E. W. et al. Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study. *BMJ* **379**, e071707 (2022).
13. Vocka, M. et al. Trefoil factor family (TFF) proteins as potential serum biomarkers in patients with metastatic colorectal cancer. *Neoplasma* **62**, 470–477 (2015).
14. Huang, Y. et al. High expression of LTBP2 contributes to poor prognosis in colorectal cancer patients and correlates with the mesenchymal colorectal cancer subtype. *Dis. Mark.* **2019**, 5231269 (2019).
15. Wang, H. et al. Predictive value of proteomic markers for advanced rectal cancer with neoadjuvant chemoradiotherapy. *BMC Cancer* **22**, 868 (2022).
16. Yusufu, A. et al. TFF3 and TFF1 expression levels are elevated in colorectal cancer and promote the malignant behavior of colon cancer by activating the EMT process. *Int. J. Oncol.* **55**, 789–804 (2019).

17. Shi, L. P. et al. MicroRNA-149 sensitizes colorectal cancer to radiotherapy by downregulating human epididymis protein 4. *Am. J. Cancer Res.* **8**, 30–38 (2018).
18. Delloye-Bourgeois, C. et al. Sonic Hedgehog promotes tumor cell survival by inhibiting CDON pro-apoptotic activity. *PLoS Biol.* **11**, e1001623 (2013).
19. Baharudin, R. et al. Identification of predictive DNA methylation biomarkers for chemotherapy response in colorectal cancer. *Front Pharm.* **8**, 47 (2017).
20. Zhao, Q. et al. Comprehensive profiling of 1015 patients' exomes reveals genomic-clinical associations in colorectal cancer. *Nat. Commun.* **13**, 2342 (2022).
21. Fitzgerald, R. C. et al. The future of early cancer detection. *Nat. Med.* **28**, 666–677 (2022).
22. Thomas, M. et al. Combining Asian and European genome-wide association studies of colorectal cancer improves risk prediction across racial and ethnic populations. *Nat. Commun.* **14**, 6147 (2023).
23. Li, X. et al. Prediction of colorectal cancer risk based on profiling with common genetic variants. *Int. J. Cancer* **147**, 3431–3437 (2020).
24. Kachuri, L. et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat. Commun.* **11**, 6084 (2020).
25. Tian, Y. et al. Calculating the starting age for screening in relatives of patients with colorectal cancer based on data from large nationwide data sets. *Gastroenterology* **159**, 159–168.e3 (2020).
26. Chen, X. et al. Personalized initial screening age for colorectal cancer in individuals at average risk. *JAMA Netw. Open* **6**, e2339670 (2023).
27. Qaseem, A. et al. Screening for colorectal cancer in asymptomatic average-risk adults: a guidance statement from the American College of Physicians (Version 2). *Ann. Intern. Med.* **176**, 1092–1100 (2023).
28. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
29. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
30. Fernandez-Rozadilla, C. et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat. Genet.* **55**, 89–99 (2023).
31. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
32. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
33. Gerds, T., Ohlendorff, J. & Ozenne, B. *riskRegression: risk regression models and prediction scores for survival analysis with competing risks*. R package version 2023.09.08 (2023). https://CRAN.R-project.org/package=riskRegression.
34. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574 (2006).
35. Liese, A. D. et al. Assessing the impact of classical risk factors on myocardial infarction by rate advancement periods. *Am. J. Epidemiol.* **152**, 884–888 (2000).
36. Zheng, Y. et al. Use of breast cancer risk factors to identify risk-adapted starting age of screening in China. *JAMA Netw. Open* **5**, e2241441 (2022).

Acknowledgements

The authors thank all the participants from the Second Affiliated Hospital of Zhejiang University School of Medicine and UK Biobank, the institutions and their staff for providing data. Q.X. is supported by the National Natural Science Foundation of China (No. 82072360, 82373234) and Project of the regional diagnosis and treatment center of the Health Planning Committee (No. JBZX-201903). X.L. is supported by the Natural Science Fund for Distinguished Young Scholars of Zhejiang Province (LR22H260001), the National Nature Science Foundation of China (82204019), and Zhejiang Provincial Clinical Research Center for CANCER (2022E50008, 2024ZY01056). Y.H. is supported by Zhejiang Province Clinical Research Center for CANCER (2022E50008.2024ZY01056) and Key R&D Program of Zhejiang (2023C03049). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

These authors are joint corresponding authors: Xue Li, Qian Xiao, Junli Gao. X.L., Q.X., J.-L.G., and J.S. contributed to the study conception and design. Material preparation, data collection, methodology, formal analysis, and visualization were performed by J.S., Y.L., J.Z., B.L., and S.Z. H.G., J.-S.G., W.L., J.W., Y.-T.H., and X.K. contributed to the project administration. The first draft of the manuscript was written by J.S., X.L., Q.X., J-L.G. and all the other authors revised and commented on previous versions of the manuscript. X.L. is the study guarantor and attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52894-2.

Correspondence and requests for materials should be addressed to Qian Xiao.

Peer review information *Nature Communications* thanks Jun Zhong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.