Aleksandrova *et al. BMC Medicine* (2021) 19:1
https://doi.org/10.1186/s12916-020-01826-0

BMC Medicine

RESEARCH ARTICLE Open Access

Development and validation of a lifestyle-based model for colorectal cancer risk prediction: the LiFeCRC score

Krasimira Aleksandrova[1,2,3*], Robin Reichmann[1,2], Rudolf Kaaks[4], Mazda Jenab[5], H. Bas Bueno-de-Mesquita[6,7], Christina C. Dahm[8], Anne Kirstine Eriksen [9], Anne Tjønneland[9], Fanny Artaud[10,11], Marie-Christine Boutron-Ruault[10,11], Gianluca Severi[10,11,12], Anika Hüsing[4], Antonia Trichopoulou[13], Anna Karakatsani[13,14], Eleni Peppa[13], Salvatore Panico[15], Giovanna Masala[16], Sara Grioni[17], Carlotta Sacerdote[18], Rosario Tumino[19], Sjoerd G. Elias[20], Anne M. May[20], Kristin B. Borch[21], Torkjel M. Sandanger[21], Guri Skeie[21], Maria-Jose Sánchez[22,23,24,25], José María Huerta[24,26], Núria Sala[27,28], Aurelio Barricarte Gurrea[24,29,30], José Ramón Quirós[31], Pilar Amiano[24,32], Jonna Berntsson[33], Isabel Drake[34], Bethany van Guelpen[35,36], Sophia Harlid[35], Tim Key[37], Elisabete Weiderpass[5], Elom K. Aglago[5], Amanda J. Cross[7], Konstantinos K. Tsilidis[7,38], Elio Riboli[7] and Marc J. Gunter[5]

Abstract

Background: Nutrition and lifestyle have been long established as risk factors for colorectal cancer (CRC). Modifiable lifestyle behaviours bear potential to minimize long-term CRC risk; however, translation of lifestyle information into individualized CRC risk assessment has not been implemented. Lifestyle-based risk models may aid the identification of high-risk individuals, guide referral to screening and motivate behaviour change. We therefore developed and validated a lifestyle-based CRC risk prediction algorithm in an asymptomatic European population.

Methods: The model was based on data from 255,482 participants in the European Prospective Investigation into Cancer and Nutrition (EPIC) study aged 19 to 70 years who were free of cancer at study baseline (1992–2000) and were followed up to 31 September 2010. The model was validated in a sample comprising 74,403 participants selected among five EPIC centres. Over a median follow-up time of 15 years, there were 3645 and 981 colorectal cancer cases in the derivation and validation samples, respectively. Variable selection algorithms in Cox proportional hazard regression and random survival forest (RSF) were used to identify the best predictors among plausible predictor variables. Measures of discrimination and calibration were calculated in derivation and validation samples. To facilitate model communication, a nomogram and a web-based application were developed.

(Continued on next page)

* Correspondence: aleksandrova@leibniz-bips.de
[1]Nutrition, Immunity and Metabolism Senior Scientist Group, Department of Nutrition and Gerontology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany
[2]Institute of Nutritional Science, University of Potsdam, Potsdam, Germany
Full list of author information is available at the end of the article

(Continued from previous page)

Results: The final selection model included age, waist circumference, height, smoking, alcohol consumption, physical activity, vegetables, dairy products, processed meat, and sugar and confectionary. The risk score demonstrated good discrimination overall and in sex-specific models. Harrell's C-index was 0.710 in the derivation cohort and 0.714 in the validation cohort. The model was well calibrated and showed strong agreement between predicted and observed risk. Random survival forest analysis suggested high model robustness. Beyond age, lifestyle data led to improved model performance overall (continuous net reclassification improvement = 0.307 (95% CI 0.264–0.352)), and especially for young individuals below 45 years (continuous net reclassification improvement = 0.364 (95% CI 0.084–0.575)).

Conclusions: LiFeCRC score based on age and lifestyle data accurately identifies individuals at risk for incident colorectal cancer in European populations and could contribute to improved prevention through motivating lifestyle change at an individual level.

Keywords: Colorectal cancer, Risk prediction, Lifestyle behaviour, Risk screening, Cancer prevention

Background

Colorectal cancer accounted for over 1.8 million new cases or 10% of all new cases of cancer worldwide in 2018 [1]. Worryingly, the global burden of colorectal cancer is expected to rise by 60% reaching 2.2 million new cases and 1.1 million deaths in 2030, with European countries ranking highest in the global statistics of colorectal cancer incidence and mortality [2]. The projected increase in colorectal cancer burden necessitates improved assessment of primary prevention strategies [2, 3]. Targeted prevention in an asymptomatic population that addresses potentially modifiable factors has potential for reducing lifestyle-associated long-term risk of colorectal cancer and represents a cost-effective approach to reduce the cancer burden [4, 5].

Lifestyle behaviours such as smoking, alcohol consumption, and poor diet have long been recognized to be associated with a higher risk of colorectal cancer [6–15]. Updated evidence on nutrition and cancer risk further highlighted the importance of risk factors such as body fatness (i.e. abdominal adiposity), adult-attained height, physical activity, high intake of red and processed meat and low intakes of whole grains, dairy products and fish [15, 16]. Despite accumulation of evidence, translation of lifestyle information into individualized colorectal cancer risk assessment strategies has not been implemented so far. Risk stratification may aid the identification of high-risk individuals, guide referral to screening and motivate lifestyle modification [17]. Individualized risk estimates in primary care may essentially aid behaviour change and complement preventive approaches to shifting population distributions of risk factors [17].

A number of colorectal cancer risk prediction models have been published over the last decade [18–21]. Most published models have been predominantly developed using data from American and Asian populations [18, 19]. We have previously validated several models in European populations based on data from UK Biobank and the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort studies [20]; however, several gaps remain to be addressed. First, only a few previous models have been developed based on prospective cohort data with long enough follow-up time to account for the potentially long latency period of colorectal cancer development [18]. Second, important emerging predictors related to nutrition and lifestyle such as abdominal fatness have not been considered [22]. Third, most models focused only on model development and did not address the full continuum of model development, validation and communication recommended in recent methodological guidelines for research on risk prediction (i.e. TRIPOD, Transparent Reporting of a multivariable Prediction model for Individual Prognosis or Diagnosis) [19, 23]. Fourth, previous models were mostly developed using logistic regression and did not account for time-to-event. New approaches such as penalized regression methods (i.e. elastic net regression) and machine learning algorithms (i.e. random survival forest) might offer additional means for model improvement [24, 25]. Finally, model communication to the wider public was generally not addressed by previous studies and was restricted to providing a formula to calculate individual absolute risk of colorectal cancer [18]. Graphical nomograms and web-based applications could further aid in facilitating model communication [26].

In this context, we aimed to develop and validate a lifestyle-based risk prediction model for the prevention of colorectal cancer in a population-based European cohort. We further aimed to construct a simple and widely applicable user-friendly risk calculator offering an estimate of colorectal cancer risk based on individual's personal data.

Methods

Study design and data source

The lifestyle-based prediction model for colorectal cancer risk (LiFeCRC score) was developed using data collected within EPIC, a multicentre prospective cohort study comprising 521,324 participants aged 17 to 98 years at study baseline (predominantly 35 to 70 years) recruited between 1992 and 2000 across 23 centres in 10 European countries [27]. Participants included blood donors, screening participants, health-conscious individuals and the general population. Written informed consent was obtained from all participants before joining the EPIC study. Approval for the EPIC study was obtained from the ethical review boards of the International Agency for Research on Cancer and from all local institutions through which subjects were recruited for the EPIC study, as previously reported [28].

Case ascertainment

The primary outcome was incident colorectal cancer. Cancer cases were identified through population cancer registries in Denmark, Italy, the Netherlands, Spain, Sweden and the UK. In France, Germany and Greece, a combination of methods was used including health insurance records, cancer pathology registries and active follow-up of study participants. Follow-up began at the date of enrolment and ended at the date of diagnosis of colorectal cancer, death or last complete follow-up. The last update of endpoint information was done up to 31 September 2010. Colon and rectal cancers were defined according to the 10th Revision of the International Statistical Classification of Diseases, Injuries and Causes of Death (ICD-10), proximal colon tumours include tumours in the cecum, cecal appendix, ascending colon, hepatic flexure, transverse colon and splenic flexure (ICD-10 codes C18.0–18.5); distal colon tumours include those in the descending colon (ICD-10 code C18.6) and sigmoid colon (ICD-10 code C18.7); and rectal tumours are those occurring at the rectosigmoid junction (ICD-10 code C19) or in the rectum (ICD-10 code C20). Only the first primary neoplasm was included in the analysis; non-melanoma skin cancer was excluded.

Study population

Figure 1 presents a flowchart of study population selection for deriving the LiFeCRC score in the EPIC cohort.

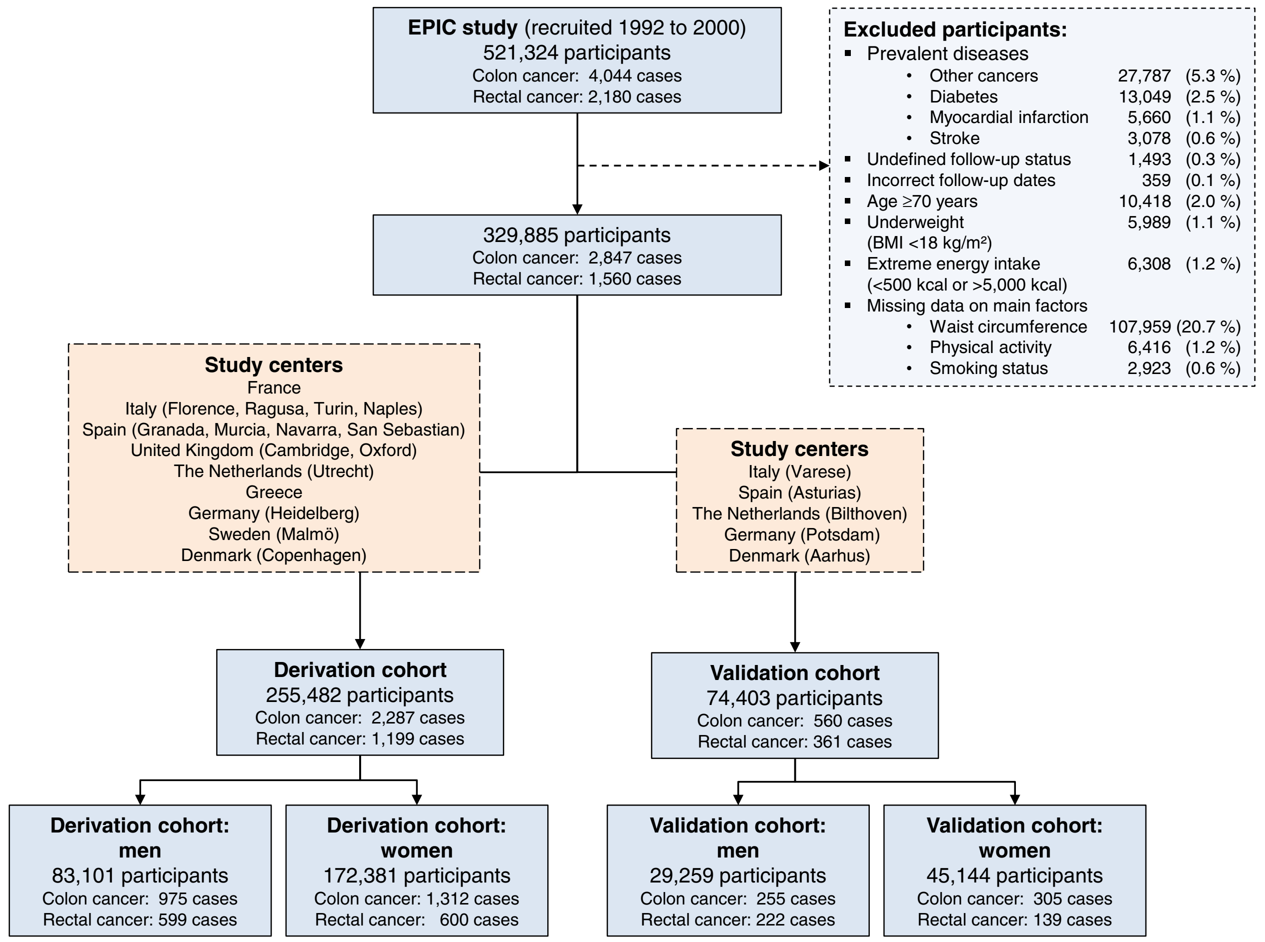

EPIC study (recruited 1992 to 2000)
521,324 participants
Colon cancer: 4,044 cases
Rectal cancer: 2,180 cases
Excluded participants:
Prevalent diseases
Other cancers 27,787 (5.3 %)
Diabetes 13,049 (2.5 %)
Myocardial infarction 5,660 (1.1 %)
Stroke 3,078 (0.6 %)
Undefined follow-up status 1,493 (0.3 %)
Incorrect follow-up dates 359 (0.1 %)
Age ≥70 years 10,418 (2.0 %)
Underweight (BMI <18 kg/m²) 5,989 (1.1 %)
Extreme energy intake (<500 kcal or >5,000 kcal) 6,308 (1.2 %)
Missing data on main factors
Waist circumference 107,959 (20.7 %)
Physical activity 6,416 (1.2 %)
Smoking status 2,923 (0.6 %)
329,885 participants
Colon cancer: 2,847 cases
Rectal cancer: 1,560 cases
Study centers
France
Italy (Florence, Ragusa, Turin, Naples)
Spain (Granada, Murcia, Navarra, San Sebastian)
United Kingdom (Cambridge, Oxford)
The Netherlands (Utrecht)
Greece
Germany (Heidelberg)
Sweden (Malmö)
Denmark (Copenhagen)
Study centers
Italy (Varese)
Spain (Asturias)
The Netherlands (Bilthoven)
Germany (Potsdam)
Denmark (Aarhus)
Derivation cohort
255,482 participants
Colon cancer: 2,287 cases
Rectal cancer: 1,199 cases
Validation cohort
74,403 participants
Colon cancer: 560 cases
Rectal cancer: 361 cases
Derivation cohort: men
83,101 participants
Colon cancer: 975 cases
Rectal cancer: 599 cases
Derivation cohort: women
172,381 participants
Colon cancer: 1,312 cases
Rectal cancer: 600 cases
Validation cohort: men
29,259 participants
Colon cancer: 255 cases
Rectal cancer: 222 cases
Validation cohort: women
45,144 participants
Colon cancer: 305 cases
Rectal cancer: 139 cases

Fig. 1 Flowchart of study population selection

Participants with prevalent cancer, diabetes, myocardial infarction or stroke at recruitment and participants without follow-up were excluded. Missing information on main risk factors (sex, anthropometric measurements, lifestyle and dietary data) was present in 22.5% of the data, and therefore, entries with missing data were excluded for complete case analysis. Based on this, participants from EPIC-Umeå and EPIC-Norway were excluded from the current analyses due to lack of data on waist circumference measurements. The resulting study sample comprised 329,885 participants among which 4626 incident colorectal cancer cases (2847 colon cancer/1560 rectal cancer) were diagnosed during study follow-up. This sample was split into a derivation cohort ($N = 255{,}482$) and a validation cohort ($N = 74{,}403$) on a non-random principle following the TRIPOD recommendations [23]. The derivation sample included participants from 21 EPIC centres in France, Italy, Spain, UK, the Netherlands, Greece, Germany, Sweden and Denmark. The validation sample included participants representing Southern and Northern European populations from 5 EPIC centres in Italy, Spain, the Netherlands, Germany and Denmark (Fig. 1).

Baseline data collection

At baseline, participants completed extensive medical, dietary and lifestyle questionnaires, including questions on alcohol use, smoking status, physical activity, education and previous illnesses. Body weight, height and waist circumference were measured in all centres except for EPIC-Oxford (health-conscious population) and EPIC-France where anthropometric measurements were self-reported [27]. Usual food intakes were measured by using country-specific validated dietary questionnaires [29]. All dietary variables used in the present study were calibrated by using an additive calibration method as previously described [30]. Non-steroidal anti-inflammatory drug (NSAID) use was only assessed in the Cambridge study center, and family history of colorectal cancer was assessed only in study centres in France, Spain and the UK. Baseline characteristics of participants with available information on NSAID use and family history of colorectal cancer are presented in Supplementary Table 1, Additional File 1.

Model development

The model development and model validation were performed and reported following the TRIPOD guidelines [23, 31] (Supplementary Table 2, Additional File 1). The general workflow of model derivation, performance evaluation, validation and model communication are presented in Supplementary Fig. 1, Additional File 2.

Overall, the LiFeCRC score was derived based on beta coefficients for colorectal cancer risk estimated in Cox proportional hazard models within the derivation dataset. Time -to - event was defined as time from baseline assessment to first cancer event. Supplementary Table 3, Additional File 1 presents the variable names and measurement scales of a predefined set of 16 predictors selected based on published literature reflecting latest evidence from systematic reviews (i.e. World Cancer Research Fund/American Institute for Cancer Research reports) and based on availability of data in the EPIC cohort. Analyses based on Schoenfeld residuals and stratified Kaplan-Meier curves revealed no violation of the proportional hazard assumption of the Cox model. To test whether the predictive performance of each variable is the same, regardless of the values of other predictors, statistical interactions between different combinations of predictor variables on the multiplicative scale were tested using the likelihood ratio test. Since model discrimination was not improved by including significant interaction terms, the inclusion of interaction terms in the final Cox models was disregarded to avoid overfitting.

Elastic net selection

Predictor variable selection was performed using bootstrapped elastic net regularization [32]. Elastic net regularization is a penalized regression method, combining least absolute shrinkage and selection operator (LASSO) and ridge regression. A penalty parameter λ is used to shrink predictor regression coefficients, eventually removing predictor variables from the model by setting their respective regression coefficient to zero. A mixing parameter α is used to fix the proportion for combining LASSO and ridge regression. Optimal values for both parameters λ and α were determined based on minimal mean error of 10-fold cross-validation using 100 possible λ values for α values between 0.5 and 1 (0.5, 0.6, 0.7, 0.8, 0.9, 1). The selected parameters were then used to bootstrap the elastic net regularization of each predictor's Cox regression coefficient with 1000 replications. Based on all bootstrap replications, mean coefficient values and 95% confidence intervals were calculated for each predictor coefficient. Predictors with confidence intervals including zero were removed. All remaining predictors were then used to generate reduced elastic net penalized Cox regression models. The model selection was conducted for colorectal cancer as a single endpoint (LiFeCRC score) and according to sex and cancer subsite (colon/rectum). Variable selection and Cox regression modeling were performed using R 3.6.1 (R Core Team) [33], and the *glmnet* (version 2.0-18) [34] and *survival* (version 2.44-1.1) [35] packages.

Absolute risk assessment

The individual 10-year absolute risk P (10y) for colorectal cancer was calculated using the following formula:

$$P(10\text{y}) = 1 - S_m(10\text{y})^{\exp(\text{Risk Score}_i\ -\ \text{Risk Score}_m)}$$

The 10-year survival function estimate S_m (10y) was calculated for average predictor variable values. The average Risk Score$_m$ and the individual Risk Score$_i$ were computed using the following formulas:

$$\text{Risk Score}_m = \sum_j \beta_j \cdot \text{predictor mean value}_j$$

$$\text{Risk Score}_i = \sum_j \beta_j \cdot \text{predictor value}_{ij}$$

The j index stands for a predictor variable of a Cox regression model and β_j is the beta estimate.

In additional analyses, the study population was stratified according to predefined risk categories of low, intermediate and high risk, based the 50th and 90th percentile of predicted risk in the derivation cohort. Incidence rates and model selection characteristics across the so defined risk categories in both the derivation and validation samples have been assessed.

Model performance: discrimination and calibration

Model discrimination

Model discrimination was assessed based on Harrell's C-index as a measure similar to the receiver operating characteristic statistic that takes the censored nature of data into account. This value represents the odds of the predicted probability of developing colorectal cancer being higher for those who actually develop colorectal cancer compared to those who do not develop the disease. To account for model optimism in terms of overfitting, bootstrapping with 1000 replications was performed. In bootstrapping, entries are randomly drawn with replacement from a data set until the bootstrap sample has the size as the original dataset. For each bootstrap sample, an elastic net penalized Cox regression model was fitted. Harrell's C-index of each bootstrap model was then calculated for the bootstrap sample and the original data in each bootstrap replication. The difference of these values was averaged over all 1000 bootstrap replications to calculate the amount of optimism for the C-index of the original model, which was used to calculate an optimism-corrected C-index. This analysis was performed in R [33] with the package *rms* (version 5.1-3.1) [36].

Model calibration

Calibration plots of estimated individual predicted risks of developing colorectal cancer in the next 10 years were derived from the penalized Cox regression model. These values were divided into deciles, and each decile's mean value was computed. The Kaplan-Meier survival function at 10 years with 95% confidence interval was calculated for each decile group. Subsequently, the trend of the mean predicted risks and the observed complement of the Kaplan-Meier survival of each decile was visually compared as a measure of calibration. Model performance, including Harrell's C-index and calibration plots, was also evaluated in the validation cohort.

Model communication

In order to assist the translation of the generated statistical model into an individual risk prediction equation, we created a 10-year risk assessment nomogram as a graphical model representation that allows risk estimation. For this purpose, we used the R [33] package *rms* (version 5.1-3.1) [36]. In addition, we developed a user-friendly risk calculator application using the R [33] packages *shiny* (version 1.2.0) [37] and *shinydashboard* (version 0.7.1) [38] that can be adapted for a web-based use. This application allows the prediction of individual colorectal cancer risk by including characteristics into input fields. The input values are then evaluated using the validated colorectal cancer risk prediction model.

Random survival forest

Random survival forest was used as an alternative machine learning method in order to prove model robustness, i.e. assess whether the same set of predictors will be selected. Each random survival forest was generated with a total number of 500 decision trees with 100 unique data points on average in each terminal node and a maximum of 10 possible random split points to consider at each branch of a decision tree. A variable importance measure for each predictor variable, describing the impact of using randomly permuted values of this variable instead of observed values for the prediction of known entries, was then extracted from the random survival forest. For the computation of random survival forests, the package "randomForestSRC" (version 2.6.1) was used. Model performance was evaluated in the derivation and validation cohort using Harrell's C-index and calibration plots.

Sensitivity analyses

In sensitivity analyses, we evaluated the added predictive value of lifestyle data beyond age, using the following statistics: (1) improvement in model discrimination—based on goodness of fit (likelihood ratio test), estimated net change in Harrell's C-index and continuous net reclassification improvement (NRI$^{>0}$); (2) improvement in model calibration based on comparison of calibration plots and (3) net benefit of the model based on decision curve analysis. We also stratified the study population in the derivation and validation sample according to age groups: < 45 years; 45–65 years; > 65 years and calculated model performance characteristics (Harrell's C-index and NRI$^{>0}$) for the lifestyle-based model across these categories. In addition, we also calculated the predicted 10-year absolute risk of colorectal cancer for a predefined “healthy” and “unhealthy” lifestyle pattern across different age groups and a constant body height. In subsample of the derivation cohort with available information, Harrell’s C-index was compared between models with and without inclusion of NSAID use or family history. To address model generalizability, we further evaluated model performance across subgroups by selected variables, i.e. waist circumference, education, smoking status (including level of smoking intensity) and level of alcohol consumption. Finally, to account for the potential influence of competing risk of death ($N = 23{,}774$), we calculated the cumulative incidence adjusted for mortality and evaluated the discrimination of the reduced model based on Fine-Gray subdistribution hazard regression [39] in both the derivation and validation samples.

Results

Baseline characteristics

Table 1 shows the baseline characteristics of men and women in the derivation and validation cohorts. Overall, the distribution of risk factors was similar across both cohorts. In the derivation cohort, the mean age at study baseline was 51.4 years, 67.5% of the participants were women, and mean age at colorectal cancer diagnosis was 66.0 years in women and 66.4 years in men. Never-smokers, physically active and highly educated people comprised 49.1%, 10.3% and 24.6% of the derivation cohort, respectively. The median follow-up time was 15.4 (interquartile range 13.2 to 16.9) years in the derivation cohort and 14.1 (interquartile range 10.5 to 16.0) years in the validation cohort.

Model development

Figure 2 illustrates the distribution of Cox regression coefficients of all predictor variables based on the bootstrapped elastic net regularization. Selected variables in the reduced model are highlighted based on the selection criterion of having a coefficient value of 0 not included in the 95% confidence interval. Table 2 shows derived colorectal cancer hazard ratios for all risk factors (full model) and risk factors that remained after elastic net selection (reduced model). The selected predictors of the overall colorectal cancer risk in men and women included age, waist circumference, height, daily alcohol consumption, smoking, physical activity, vegetables, dairy products, processed meat, and sugar and confectionary (Table 2). The models derived separately for men and women confirmed age, waist circumference, smoking and vegetable intake as consistent predictors across both genders. Additional predictors retained in the reduced model in men were daily alcohol consumption, dairy intake, dark bread and red meat, and in women, height and processed meat. The estimated 10-year mean absolute risk for colorectal cancer of the derivation cohort was 0.78% in both sexes, 1.07% in men and 0.64% in women (Table 2). Table 3 provides an overview of selected variables by anatomical subsite, colon and rectal cancer, overall and separately in men and women. An additional predictor that was retained in the model for rectal cancer was the intake of soft drinks. Notably, selected predictors in women were somewhat different for colon and rectal cancer. For colon cancer, the model included age, waist circumference, height, smoking and vegetable intake, whereas for rectal cancer it included age, processed meat and soft drinks (Table 3).

Model performance: discrimination and calibration

Overall model discrimination was good with Harrell’s C-index of 0.709 for the derived colorectal cancer risk model. Optimism-adjusted Harrell's C index ranged from 0.667 for the model for rectal cancer in women to 0.716 for the model for colon cancer in both sexes (Table 4). Reduced models showed similar predictive performance as the “full models” suggesting that obtaining data on selected predictors would yield sufficient information and additional factors are not adding predictive value to the model. The performance in the validation cohort was similar for all models, suggesting a high level of stability and a lack of overfitting. Calibration plots of derived colorectal cancer risk models in the derivation and validation sample overall and by sex are presented in Fig. 3. An overall good calibration was observed based on the comparable intercepts for models across derivation and validation samples.

Model communication

Absolute risk formula

To provide assessment of the absolute 10-year risk of colorectal cancer for individuals with various combinations of risk factors, we prepared a formula with the following selected predictors:

$$\text{Absolute risk}\begin{pmatrix}\text{Colorectal cancer}\\ \text{within 10 years}\end{pmatrix} = 1 - S_m(10\,\text{years})^{\exp(\text{Risk Score}_i\ -\ \text{Risk Score}_m)} = 1 - 0.9943^{\exp(\text{Risk Score}_i\ -\ 6.8089)}$$

Table 1 Baseline characteristics of participants in the derivation and validation cohorts

Characteristics	Derivation cohort			Validation cohort		
	All participants	Men	Women	All participants	Men	Women
N	255,482	83,101	172,381	74,403	29,259	45,144
Age at recruitment, years, mean (SD)	51.4 (9.7)	52.3 (9.0)	51.0 (9.9)	49.7 (9.6)	50.7 (9.3)	49.0 (9.8)
Age range, years	19.5 to 70.0	19.5 to 70.0	20.0 to 70.0	19.9 to 70.0	20.1 to 69.2	19.9 to 70.0
BMI, kg/m^2, mean (SD)	25.8 (4.3)	26.6 (3.6)	25.4 (4.5)	26.0 (4.2)	26.6 (3.6)	25.7 (4.5)
Waist, cm, mean (SD)	84.6 (12.9)	94.6 (10.0)	79.8 (11.3)	86.0 (12.5)	94.0 (10.2)	80.8 (11.0)
Height, cm, mean (SD)	165.8 (9.1)	174.3 (7.3)	161.7 (6.7)	167.3 (9.6)	175.5 (7.3)	162.1 (6.9)
Postmenopausal status, %			49.5			41.8
Ever use of hormone for menopause, %			26.0			23.8
Smoking status, %						
Smoker	23.5	31.3	19.7	27.0	32.3	23.5
Former	27.4	36.2	23.1	27.8	37.3	21.6
Never	49.1	32.4	57.2	45.3	30.4	54.9
Physical activity, %						
Inactive	19.1	29.4	14.2	18.3	25.3	13.7
Moderately inactive	30.3	32.9	29.0	27.2	30.0	25.4
Moderately active	40.2	29.3	45.5	43.5	33.7	49.9
Active	10.3	8.5	11.2	11.0	11.0	11.0
Education, %						
None	5.9	4.9	6.3	3.1	2.4	3.6
Primary school completed	27.0	30.6	25.2	29.2	24.8	32.1
Technical school/professional school	22.4	22.7	22.3	31.1	29.8	31.9
Secondary school	16.2	11.9	18.3	13.4	11.4	14.7
University degree	24.6	27.4	23.3	23.0	31.5	17.5
Not specified	4.0	2.5	4.7	0.2	0.2	0.3
Dietary intake, g/day, median (IQR)						
Alcohol	6.6 (1.1 to 17.1)	14.4 (5.0 to 32.3)	4.1 (0.6 to 11.9)	7.8 (1.5 to 19.7)	16.4 (6.7 to 32.9)	4.2 (0.6 to 11.8)
Vegetables	196.9 (124.4 to 301.8)	179.7 (110.8 to 287.8)	204.9 (131.5 to 307.4)	130.0 (92.8 to 183.2)	124.0 (88.1 to 174.7)	133.9 (95.9 to 189.1)
Fruits	215.1 (117.1 to 340.8)	176.6 (88.5 to 313.3)	232.7 (132.4 to 351.8)	161.0 (94.1 to 264.7)	127.4 (73.9 to 225.3)	186.0 (105.6 to 290.3)
Dark bread	28.6 (0.0 to 91.8)	34.9 (0.0 to 112.5)	27.9 (0.0 to 87.8)	91.4 (15.0 to 150.2)	115.8 (49.3 to 179.6)	73.2 (8.0 to 128.0)
Dairy products	283.0 (159.5 to 447.5)	257.3 (130.0 to 434.3)	295.2 (173.1 to 451.7)	265.1 (150.0 to 434.7)	256.8 (136.7 to 447.0)	270.2 (158.8 to 428.7)
Red meat	38.2 (17.3 to 65.4)	49.7 (24.8 to 80.5)	33.8 (13.9 to 57.3)	44.8 (23.9 to 74.5)	61.9 (33.5 to 93.1)	36.9 (19.8 to 61.1)
Poultry	16.1 (5.8 to 30.9)	16.4 (7.3 to 34.3)	15.8 (4.8 to 29.3)	13.2 (6.5 to 24.3)	14.6 (7.3 to 25.5)	12.4 (5.9 to 23.4)

Table 1 Baseline characteristics of participants in the derivation and validation cohorts *(Continued)*

Characteristics	Derivation cohort			Validation cohort		
	All participants	**Men**	**Women**	**All participants**	**Men**	**Women**
Processed meat	19.2 (6.4 to 37.9)	27.9 (10.8 to 51.9)	16.4 (5.4 to 31.8)	34.5 (17.9 to 59.6)	47.6 (27.4 to 76.7)	27.8 (14.6 to 48.7)
Fish	21.4 (9.3 to 37.0)	24.3 (12.6 to 41.2)	19.7 (7.7 to 34.9)	16.1 (6.0 to 30.3)	17.6 (6.3 to 32.2)	15.3 (5.9 to 27.9)
Sugar and confectionary	31.6 (16.3 to 55.2)	36.9 (19.2 to 65.0)	29.6 (15.2 to 51.0)	35.3 (18.7 to 60.0)	41.0 (21.7 to 70.3)	32.3 (17.1 to 54.1)
Soft drinks	9.5 (0.0 to 85.7)	16.4 (0.0 to 100.0)	6.6 (0.0 to 74.3)	8.7 (0.0 to 85.7)	19.3 (0.0 to 118.7)	3.8 (0.0 to 66.0)

BMI body mass index, *IQR* interquartile range, *SD* standard deviation

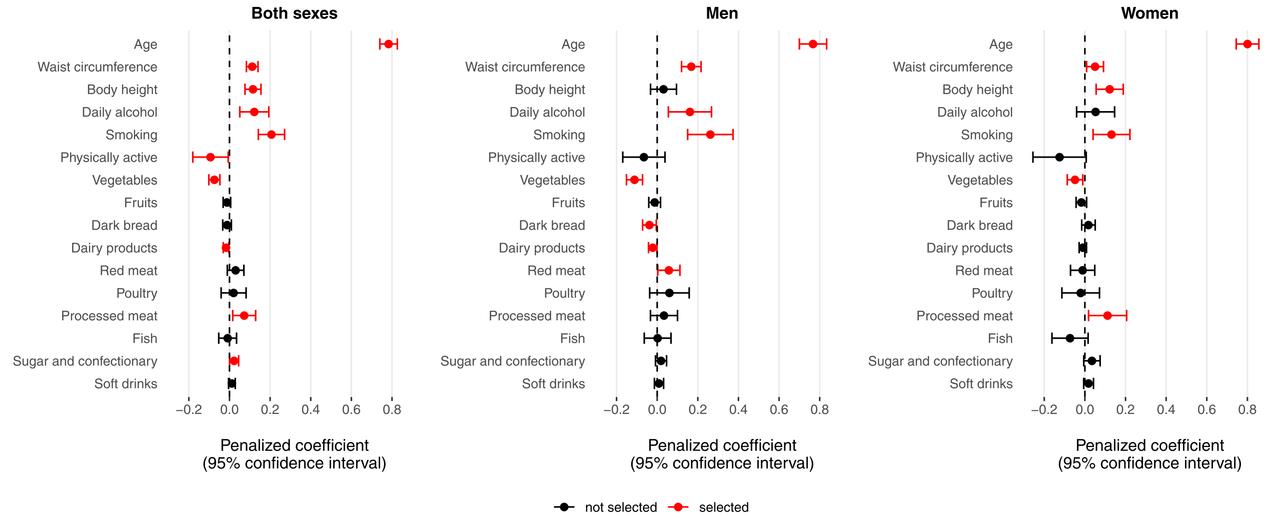


Fig. 2 Average Cox regression coefficients with 95% confidence intervals after bootstrapped elastic net regularization. Bootstrapping was performed over 1000 repetitions. Selected variables with a confidence interval not including 0 are highlighted in red

$$\begin{aligned}
\text{Risk Score}_i = {} & 0.0781 \times \text{Age}_i\,(\text{years}) \\
& +0.0117 \times \text{Waist circumference}_i\,(\text{cm}) \\
& +0.0115 \times \text{Body height}_i\,(\text{cm}) \\
& +0.1292 \times \text{Daily alcohol}_i\,(\text{yes} = 1, \text{no} = 0) \\
& +0.2125 \times \text{Smoking}_i\,(\text{yes} = 1, \text{no} = 0) \\
& -0.0964 \times \text{Physically active}_i\,(\text{yes} = 1, \text{no} = 0) \\
& -0.0773 \times \text{Vegetable intake}_i\,(\text{per } 100\text{g/day}) \\
& -0.0166 \times \text{Dairy products intake}_i\,(\text{per } 100\text{g/day}) \\
& +0.0808 \times \text{Processed meat intake}_i\,(\text{per } 50\text{g/day}) \\
& +0.0268 \times \text{Sugar and confectionary}_i\,(\text{per } 50\text{g/day})
\end{aligned}$$

Values for S_m (10 years) and Risk Score$_m$ are given in Table 2. Absolute risk for different timespans can be calculated by replacing S_m in the formula accordingly. The survival function estimates for timespans between 0 and 20 years are shown in Supplementary Fig. 2, Additional File 2. Incidence rates and model selection characteristics across predefined risk categories (low, intermediate and high risk) with cut points at 0.62% and 1.60% 10-year absolute risk are presented in Supplementary Table 4, Additional File 1, for both the derivation and validation sample.

Nomogram

Figure 4 shows a nomogram of the weights and points of the colorectal cancer risk prediction score allowing estimation of an individual's probability to develop colorectal cancer over a 10-year period. The nomogram is characterized by a scale corresponding to each variable, a point scale, a total point scale and a probability scale. The use of the nomogram is simple and involves 3 steps. First, on the scale for each variable, the value corresponding to a specific individual is read and the point scale is used to calculate the points for all variable values. Second, the total number of points is calculated by adding up all the points obtained in the previous step, and its value is identified on the total point scale. Finally, the probability of an event corresponding to the total points of the individual is represented on the risk scale. As a practical example, we estimated the 10-year risk of colorectal cancer, for individuals with two different combinations of ages and lifestyle factors, representing low-risk and high-risk extremes: individual 1 was 45 years old (50 points) with a body height of 166 cm (7.5 points), a waist circumference of 70 cm (3 points) and healthy lifestyle behaviour (low daily alcohol consumption (0 points), non-smoker (0 points), physically active (0 points), 430 g daily vegetable intake (7 points), 630 g daily dairy products intake (2.5 points), 0 g daily processed meat intake (0 points), and 5 g daily sugar and confectionary intake (0 points)), and individual 2 was 65 years old (90 points) with a body height of 166 cm (7.5 points), a waist circumference of 100 cm (12 points) and rather unhealthy lifestyle behaviour (high daily alcohol consumption (3 points), smoker (5 points), physically inactive (2.5 points), 80 g daily vegetable intake (14.5 points), 70 g daily dairy products intake (5 points), 60 g daily processed meat intake (2.5 points), and 90 g daily sugar and confectionary intake (1.5 points)). The total number of points of the various prediction indicators was ~ 70 and ~ 143.5 and the corresponding absolute predicted 10-year risk of colorectal cancer was ~ 0.2% (risk score of ~ 5.7) and ~ 3–3.5% (risk score of ~ 8.6) for individual 1 and individual 2, respectively.

Web-based calculator

As an alternative approach to model communication, we developed a web-based calculator for the estimation of a

Table 2 Characteristics of colorectal cancer prediction models developed in the derivation cohort, overall and by sex

Predictor variables	**Hazard ratios (95% confidence intervals), *p* value**											
	Both sexes				**Men**				**Women**			
	Full model*		**Reduced model† (LiFeCRC score)**		**Full model***		**Reduced model†**		**Full model***		**Reduced model†**	
Age at recruitment, per 10 years	2.23 (2.13 to 2.33)	< .001	2.18 (2.09 to 2.29)	< .001	2.19 (2.04 to 2.35)	< .001	2.12 (1.98 to 2.27)	< .001	2.26 (2.13 to 2.40)	< .001	2.21 (2.08 to 2.34)	< .001
Waist circumference, per 10 cm	1.12 (1.09 to 1.15)	< .001	1.12 (1.09 to 1.16)	< .001	1.19 (1.13 to 1.25)	< .001	1.19 (1.13 to 1.25)	< .001	1.06 (1.01 to 1.10)	0.008	1.05 (1.01 to 1.09)	0.023
Height, per 10 cm	1.13 (1.08 to 1.17)	< .001	1.12 (1.08 to 1.17)	< .001	1.04 (0.96 to 1.11)	0.333			1.14 (1.06 to 1.22)	< .001	1.16 (1.08 to 1.24)	< .001
Daily alcohol consumption, high	1.14 (1.06 to 1.23)	< .001	1.14 (1.06 to 1.22)	< .001	1.18 (1.07 to 1.31)	0.002	1.18 (1.06 to 1.30)	0.002	1.06 (0.96 to 1.18)	0.245		
Ever smoker, yes	1.24 (1.16 to 1.33)	< .001	1.24 (1.15 to 1.32)	< .001	1.31 (1.17 to 1.47)	< .001	1.31 (1.17 to 1.46)	< .001	1.15 (1.05 to 1.26)	0.003	1.16 (1.06 to 1.27)	0.001
Physically active, yes	0.89 (0.82 to 0.97)	0.010	0.91 (0.83 to 0.99)	0.026	0.93 (0.83 to 1.04)	0.185			0.87 (0.76 to 0.99)	0.029		
Vegetables, per 100 g/day	0.92 (0.90 to 0.95)	< .001	0.93 (0.90 to 0.95)	< .001	0.89 (0.85 to 0.93)	< .001	0.89 (0.85 to 0.92)	< .001	0.95 (0.92 to 0.99)	0.009	0.93 (0.90 to 0.97)	< .001
Fruits, per 100 g/day	0.99 (0.97 to 1.01)	0.180			0.99 (0.96 to 1.02)	0.424			0.98 (0.96 to 1.01)	0.196		
Dark bread, per 50 g/day	0.98 (0.96 to 1.01)	0.147			0.96 (0.93 to 0.99)	0.015	0.97 (0.94 to 1.00)	0.070	1.02 (0.98 to 1.06)	0.329		
Dairy products, per 100 g/day	0.98 (0.97 to 0.99)	0.006	0.98 (0.97 to 1.00)	0.017	0.98 (0.96 to 1.00)	0.016	0.98 (0.96 to 1.00)	0.048	0.99 (0.97 to 1.01)	0.198		
Red meat, per 50 g/day	1.03 (0.99 to 1.08)	0.158			1.06 (1.00 to 1.13)	0.046	1.08 (1.02 to 1.14)	0.010	0.98 (0.91 to 1.06)	0.577		
Poultry, per 50 g/day	1.03 (0.95 to 1.12)	0.461			1.07 (0.96 to 1.20)	0.210			0.99 (0.87 to 1.12)	0.845		
Processed meat, per 50 g/day	1.08 (1.02 to 1.14)	0.006	1.08 (1.03 to 1.14)	0.004	1.04 (0.97 to 1.11)	0.296			1.13 (1.03 to 1.24)	0.010	1.12 (1.02 to 1.23)	0.020
Fish, per 50 g/day	0.99 (0.93 to 1.05)	0.665			1.00 (0.93 to 1.09)	0.914			0.92 (0.83 to 1.02)	0.109		
Sugar and confectionary, per 50 g/day	1.03 (1.00 to 1.05)	0.028	1.03 (1.00 to 1.05)	0.022	1.02 (0.99 to 1.05)	0.118			1.04 (0.99 to 1.09)	0.088		
Soft drinks, per 100 g/day	1.02 (1.00 to 1.03)	0.097			1.01 (0.99 to 1.04)	0.392			1.02 (1.00 to 1.05)	0.095		
Survival function$_m$ (10 years)‡	0.9944		0.9943		0.9920		0.9919		0.9953		0.9952	
Risk Score$_m$§	6.8953		6.8089		6.2600		5.5356		6.4758		6.7039	
Absolute Risk$_m$ (10 years)#	0.78%		0.78%		1.07%		1.07%		0.64%		0.64%	

*Full models were derived by using all available predictor variables

†Reduced models were derived by using bootstrapped elastic net variable selection with all predictor variables of the full model

‡Survival function$_m$ (10 years): 10-year survival function estimate of average predictor values of the derivation cohort. Estimates for timespans between 0 and 20 years are shown in Supplementary Fig. 2, Additional File 2

§Risk Score$_m$: Mean risk score calculated based on the sum of beta coefficient products of average predictor values of the derivation cohort

#Absolute Risk$_m$: 10-year mean absolute risk of the derivation cohort

Table 3 Characteristics of colon and rectal cancer prediction models developed in the derivation cohort, overall and by sex

Predictor variables	Hazard ratios (95% confidence intervals), p-value											
	Colon cancer						Rectal cancer					
	Both sexes		Men		Women		Both sexes		Men		Women	
Age at recruitment, per 10 years	2.30 (2.17 to 2.44)	< .001	2.11 (1.93 to 2.30)	< .001	2.31 (2.15 to 2.49)	< .001	2.02 (1.87 to 2.18)	< .001	2.10 (1.88 to 2.35)	< .001	1.86 (1.68 to 2.04)	< .001
Waist circumference, per 10 cm	1.14 (1.10 to 1.18)	< .001	1.27 (1.19 to 1.35)	< .001	1.07 (1.02 to 1.12)	0.006	1.10 (1.05 to 1.15)	< .001				
Height, per 10 cm	1.14 (1.08 to 1.19)	< .001			1.17 (1.08 to 1.28)	< .001	1.15 (1.08 to 1.23)	< .001				
Daily alcohol consumption, high	1.12 (1.02 to 1.22)	0.017					1.21 (1.07 to 1.37)	0.002	1.27 (1.08 to 1.50)	0.004		
Ever smoker, yes	1.18 (1.08 to 1.28)	< .001	1.18 (1.02 to 1.35)	0.024	1.14 (1.02 to 1.28)	0.018	1.32 (1.17 to 1.49)	< .001	1.44 (1.19 to 1.74)	< .001		
Physically active, yes	0.88 (0.80 to 0.98)	0.025			0.84 (0.72 to 0.99)	0.035						
Vegetables, per 100 g/day	0.91 (0.89 to 0.94)	< .001	0.90 (0.85 to 0.94)	< .001	0.93 (0.89 to 0.97)	0.001	0.92 (0.88 to 0.96)	< .001	0.88 (0.82 to 0.93)	< .001		
Fruits, per 100 g/day												
Dark bread, per 50 g/day	0.96 (0.93 to 0.99)	0.017	0.94 (0.90 to 0.98)	0.008								
Dairy products, per 100 g/day	0.98 (0.96 to 0.99)	0.007	0.96 (0.93 to 0.98)	0.001								
Red meat, per 50 g/day												
Poultry per, 50 g/day												
Processed meat, per 50 g/day							1.20 (1.10 to 1.31)	< .001			1.27 (1.09 to 1.47)	0.002
Fish, per 50 g/day												
Sugar and confectionary, per 50 g/day												
Soft drinks, per 100 g/day							1.03 (1.00 to 1.06)	0.025			1.06 (1.02 to 1.10)	0.006
Survival function$_m$ (10 years)*	0.9966		0.9953		0.9970		0.9979		0.9968		0.9982	
Risk Score$_m$†	7.156		5.802		7.137		6.8764		3.9117		3.2985	
Absolute Risk$_m$ (10 years)‡	0.48%		0.62%		0.41%		0.28%		0.42%		0.21%	

*Survival function$_m$ (10 years): 10-year survival function of average predictor values of the derivation cohort

†Risk Score$_m$: Mean risk score calculated based on the sum of beta coefficient products of average predictor values of the derivation cohort

‡Absolute Risk$_m$: 10-year mean absolute risk of the derivation cohort

Table 4 Model selection and discrimination in the derivation and validation cohorts

	Colorectal cancer			Colon cancer			Rectal cancer		
Selected predictors	**Both sexes**	**Men**	**Women**	**Both sexes**	**Men**	**Women**	**Both sexes**	**Men**	**Women**
Age at recruitment, per 10 years	●	●	●	●	●	●	●	●	●
Waist circumference, per 10 cm	●	●	●	●	●	●	●		
Height, per 10 cm	●		●	●		●	●		
Daily alcohol consumption, high	●	●		●			●	●	
Ever smoker, yes	●	●	●	●	●	●	●	●	
Physically active, yes	●			●		●			
Vegetables, per 100 g/day	●	●	●	●	●	●	●	●	
Fruits, per 100 g/day									
Dark bread, per 50 g/day		●		●	●				
Dairy products, per 100 g/day	●	●		●	●				
Red meat, per 50 g/day		●							
Poultry, per 50 g/day									
Processed meat, per 50 g/day	●		●				●		●
Fish, per 50 g/day									
Sugar and confectionary, per 50 g/day	●								
Soft drinks, per 100 g/day							●		●
Harrell's C-index									
Full model									
Derivation cohort	0.710	0.700	0.702	0.718	0.708	0.718	0.705	0.705	0.677
Optimism corrected *	0.708	0.697	0.700	0.716	0.707	0.715	0.704	0.703	0.668
Validation cohort	0.715	0.707	0.700	0.708	0.727	0.700	0.730	0.689	0.693
Reduced model									
Derivation cohort	0.710	0.699	0.700	0.717	0.705	0.717	0.703	0.700	0.668
Optimism corrected*	0.709	0.698	0.699	0.716	0.704	0.715	0.701	0.698	0.667
Validation cohort	0.714	0.708	0.699	0.708	0.727	0.698	0.728	0.687	0.696

*Harrell's C-index for the derivation cohort corrected for optimism by bootstrapping with 1000 replications. For each bootstrap sample a new model is fitted and the C-index calculated for the bootstrap sample and the original derivation cohort. The difference between these two C-indices is then averaged over all bootstrap replications and then subtracted from the original C-index

personalized colorectal cancer risk based on the validated LiFeCRC score. A graphical illustration of the application layout with predicted and absolute risk values for a modifiable time span is presented in Fig. 5. Of note, the results produced by the web-based calculator should be interpreted considering that competing risk of mortality was not included in the absolute risk calculation.

Random survival forest

Results of random survival forest-based relative variable importance for colorectal cancer risk prediction are presented in Supplementary Fig. 3, Additional File 2. The main selected predictors remained similar as in the Cox regression model, confirming model robustness. The highest relative importance was observed for age, followed by waist circumference, red and processed meat intake, height and vegetable consumption. The model for women showed, in addition, height, dark bread and dairy products intake as additional important predictors, whereas the model for men showed smoking and sweets and confectionary consumption as additional important predictors. Overall, the discrimination (Supplementary Fig. 3, Additional File 2) and calibration (Supplementary Fig. 4, Additional File 2) of the random survival forest based colorectal cancer risk prediction model was comparable to the Cox regression model.

Sensitivity analysis

In a sensitivity analysis, we evaluated to what extent lifestyle data added predictive value to the colorectal cancer risk model based on age only. The addition of the lifestyle variables resulted in a statistically significantly increased goodness of fit (likelihood ratio test $p < 0.001$). The estimated $\text{NRI}^{>0}$ was 0.307 (95% confidence interval 0.264 to 0.352) indicating an improvement in model performance. Supplementary Fig. 5, Additional File 2

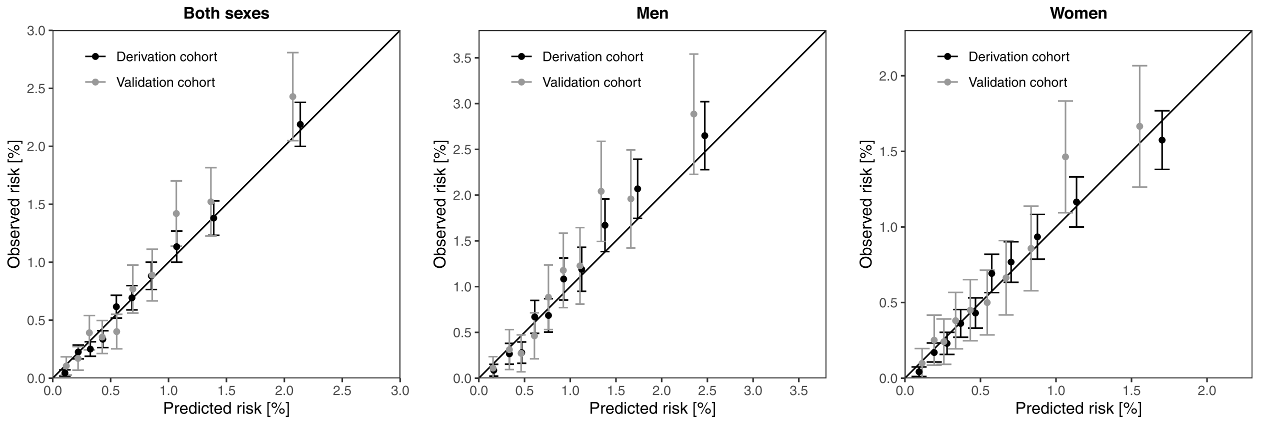


Fig. 3 Calibration plots of 10-year colorectal cancer risk. Predicted risk is compared against observed risk in the derivation and validation cohorts, overall and by sex. Observed risk is based on the complement of the Kaplan-Meier survival curve

displays the model calibration and net benefit curves for an aged-based model and the LiFeCRC model that additionally included lifestyle factors for overall colorectal cancer. An improved calibration and higher net benefit were observed for colorectal cancer risk thresholds between 0.7 and 2.5% for the LiFeCRC model compared to the age-based model. In analyses stratified according to age groups, model performance was higher in individuals < 45 years and adding lifestyle data contributed to improved reclassification statistics, i.e. higher $\mathrm{NRI}^{>0}$, suggesting relative importance of lifestyle data assessment for risk prediction at younger ages (< 45 years), i.e. $\mathrm{NRI}^{>0} = 0.364$ (95% confidence interval 0.084 to 0.575) (Supplementary Table 5, Additional File 1). We further estimated the predicted 10-year absolute risk of colorectal cancer for an arbitrary predefined "healthy" and "unhealthy" lifestyle, across different age groups and a constant body height (Supplementary Fig. 6, Additional File 2). For example, an individual aged 45 years with a body height of 166 cm adopting a predefined "unhealthy lifestyle" (waist circumference of 100 cm, high daily alcohol consumption, smoker, physically inactive, 80 g daily vegetable intake, 70 g daily dairy products intake, 60 g daily processed meat intake and 90 g daily sugar and confectionary intake) has a 3.6 times higher absolute risk of colorectal cancer within the next 10 years compared

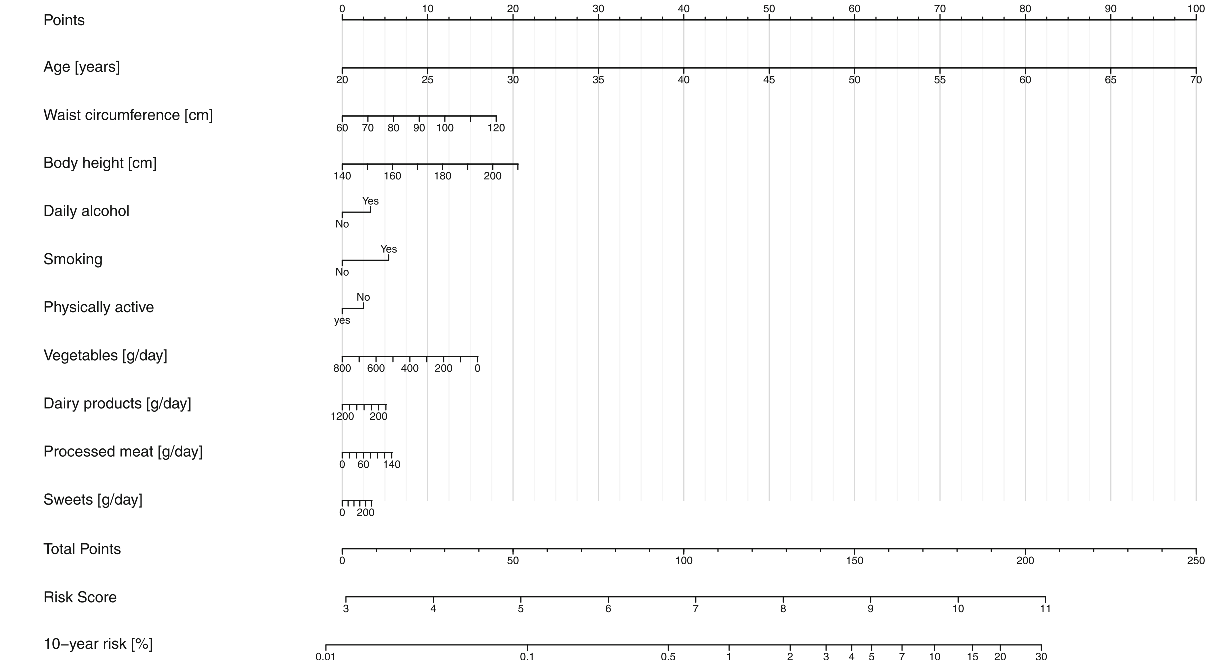
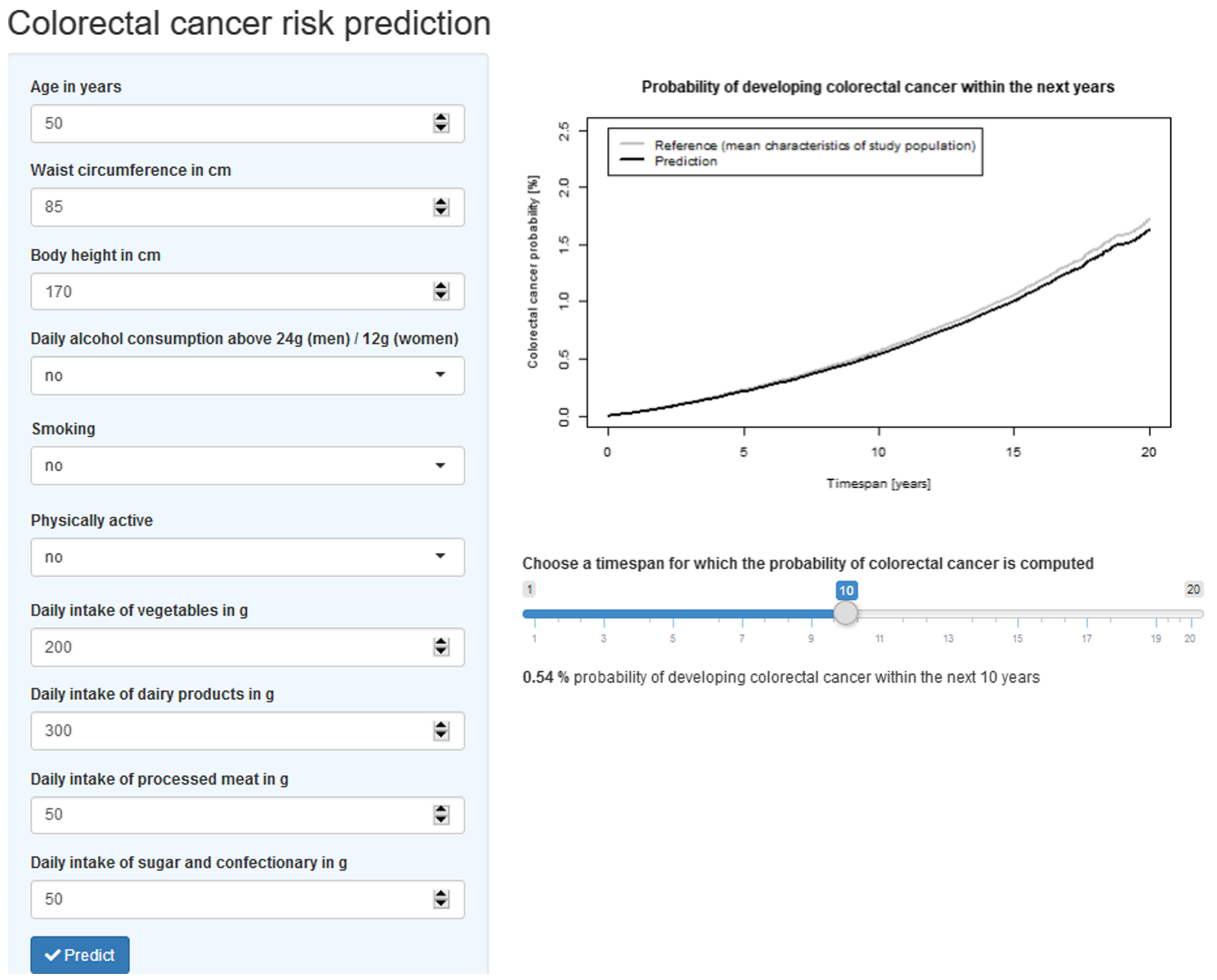


Fig. 4 Nomogram of colorectal cancer absolute risk prediction over 10 years



Colorectal cancer risk prediction

Age in years: 50

Waist circumference in cm: 85

Body height in cm: 170

Daily alcohol consumption above 24g (men) / 12g (women): no

Smoking: no

Physically active: no

Daily intake of vegetables in g: 200

Daily intake of dairy products in g: 300

Daily intake of processed meat in g: 50

Daily intake of sugar and confectionary in g: 50

✓ Predict

Probability of developing colorectal cancer within the next years
Reference (mean characteristics of study population)
Prediction
Colorectal cancer probability [%]
0.0
0.5
1.0
1.5
2.0
2.5
Timespan [years]
0
5
10
15
20

Choose a timespan for which the probability of colorectal cancer is computed: 10

0.54 % probability of developing colorectal cancer within the next 10 years

Fig. 5 Application for the colorectal cancer risk model. Example for a hypothetical individual data entry and risk calculation

to a person of the same age and body height, adopting a predefined "healthy lifestyle" (waist circumference of 70 cm, low daily alcohol consumption, non-smoker, physically active, 430 g daily vegetable intake, 630 g daily dairy products intake, 0 g daily processed meat intake and 5 g daily sugar and confectionary intake). In a subsample with available information, addition of information on NSAID use or family history of colorectal cancer to the list of predictors did not further improve model performance beyond main lifestyle variables (Supplementary Fig. 7, Additional File 2). The results did not reveal marked differences in model discrimination among subgroups by waist circumference, education, smoking status and levels of alcohol consumption (Supplementary Table 6, Additional File 1). Furthermore, no substantial differences could be seen between the Kaplan-Meier survival function and the cumulative incidence function taking competing risk into account (data not shown). Also, no differences in the discrimination ability of the Fine-Gray model taking competing risk of death into account could be observed (C-index = 0.710).

Discussion

In this large European prospective cohort study, we developed and validated the LiFeCRC score, as a lifestyle-based prediction model for the prevention of colorectal cancer in asymptomatic populations across Europe. Beyond age, the variables retained in the model were waist circumference, height, daily alcohol consumption, smoking status, physical activity and dietary intakes of vegetables, dairy products, processed meat and sugar and confectionary. Separate models were also developed for men and women and for colon and rectal cancer subtypes. The model showed good calibration and discrimination properties to identify individuals at all levels of colorectal cancer risk. Modifiable lifestyle factors contributed to model performance and accuracy beyond age alone and could improve reclassification statistics especially in younger age groups (< 45 years). A user-friendly colorectal cancer risk nomogram and a web calculator were developed to facilitate model communication.

Currently, the target population for colorectal cancer screening is mainly selected based on age alone (i.e. 50 years or above). Although age is undoubtedly an important predictor of colorectal cancer as shown in our data, information on modifiable lifestyle factors allows provision of preventive health recommendations for individuals at risk [40]. Lifestyle-based models have been suggested in medical practice as important tools that could be used to identify those most likely to benefit from lifestyle interventions and to contribute to behaviour change interventions [41]. A number of intervention studies focusing on changing lifestyle for colorectal cancer prevention reported significant effects on the target behaviours [42–46]. In those studies, tailored approaches that enable personalized feedback regarding individual lifestyle patterns were suggested as more successful compared to generic approaches [42–47]. Despite lifestyle interventions representing a powerful cost-effective strategy for colorectal cancer prevention, there has been little incentive on the side of health professionals to advocate lifestyle-based recommendations [48]. Risk assessment tools such as the LifeCRC score could facilitate improved advocacy on the side of health professionals and motivate or empower individuals to implement behaviour changes [47, 49]. Targeting lifestyle factors in those at highest risk may be particularly relevant for younger age groups that may profit most from early preventive interventions aimed at encouraging behavioural changes [47].

A number of previous models incorporated lifestyle data with common covariates including self-reported BMI (body mass index), alcohol consumption and smoking [18–21]. Recently, a model based on BMI, smoking, alcohol, red and processed meat, fruits, vegetables and physical activity demonstrated C-statistics of 0.66 and 0.68 in men and women, respectively [41]. Compared with this and other published models that also include family history and more complex variables [18, 19, 50, 51], the EPIC lifestyle-based model showed a comparable and even improved performance based on Harrell's C-index of 0.710 in both derivation and validation cohort. As previously reported, the highest C-statistic for colorectal cancer risk prediction model ranged from 0.67 in UK Biobank to 0.69 EPIC validation samples [20]. Compared to our model, that model included 13 variables: age, ethnicity, education, BMI, family history, diabetes, oestrogen exposure, non-steroidal anti-inflammatory use, physical activity, smoking, alcohol, red meat intake and multivitamin use. Having the strong discrimination statistics for models based on age alone, additional predictors were shown to add little improvement to model C-statistics in previous studies as well as in our data [18, 20, 51]. To address the question whether lifestyle information is important for absolute risk assessment beyond age, we evaluated the model performance across different age groups. These results showed that the model performance was highest in the group of participants < 45 years old and suggested this age period as a relevant time window for early cancer prevention. We further calculated the 10-year absolute risk of colorectal cancer across different ages comparing predefined “healthy” versus “unhealthy” lifestyle pattern based on selected model predictors. These analyses suggested that at a given age and height, i.e. for an individual aged 45 years with a body height of 166 cm, following the unhealthy lifestyle pattern would lead to 3.6 times higher absolute risk of colorectal cancer within the next 10 years compared to a person of the same age and body height, adopting a healthy lifestyle. These results highlight the importance of adherence to healthy lifestyle for the long-term reduction of colorectal cancer risk. In support of these data, recent analysis based on a large German population sample showed that healthy lifestyle could improve prospects for avoiding colorectal cancer in the long term even beyond individual genetic risk [52].

The elaborated phenotyping and detailed assessment of nutritional data in the EPIC cohort allowed selection of several factors not commonly depicted in previous colorectal cancer risk prediction models. Compared to previous models that used data on self-reported BMI, in the EPIC cohort data was available on waist circumference measurements and these were among the main predictors [53, 54]. Unlike BMI which does not take body fat distribution into account, waist circumference provides a proxy for the centrally located visceral fat shown especially relevant for colorectal cancer development [53, 55]. Only a few previous models included data on height which was selected as another important predictor by our model [56, 57]. Greater height could provide reflection of an increased standard of living characterized by greater availability of energy and protein-rich foods, lower physical activity and a reduced incidence of childhood infections that follow different patterns across Europe [58]. Physical activity was also selected as a predictor of colorectal cancer risk, particularly in the model for women. These data support recent findings from the Women's Health Initiative [59] and the overall notion of the importance of physical activity for the prevention of colorectal cancer [60]. Beyond red meat [56, 57, 61] and vegetable intake [56, 62–64], additional dietary predictors selected by our model included low dairy intake and high intakes of sugary products, including soft drinks. Guiding individuals towards healthy dietary and lifestyle choices could complement colorectal cancer screening as means for colorectal cancer prevention.

The selected model performed similarly well as the model with the full list of predictors, suggesting that it can be used as a simpler approach for determining high-risk individuals. Thus, individuals and health professionals would need to inquire about fewer lifestyle factors, avoiding the use of long questionnaires and minimizing the burden of data collection on both the patient and clinician side. However, for a comprehensive lifestyle recommendation, all healthy behaviours could be considered in additional counselling. The model performance among women was modest, and better in men, likely because some risk factors were more strongly associated with risk among men. The general distribution and influence of risk factors may differ geographically across populations and additional model elaboration and adaptation of country-specific risk models should be further considered. Ultimately, research is needed to assess the feasibility and effectiveness of the current lifestyle-based risk assessment tool on health behaviour modification, colorectal cancer risk factor improvement, and overall potential for colorectal cancer prevention when incorporated into the primary care setting, particularly as a pre-screening instrument of high-risk patients. More work is also warranted for the refinement of the risk communication tool before its general integration into practice. Finally, in future research, additional predictors, including relevant biomarker and genetic variables, should be further explored on the way towards improved precision prevention of colorectal cancer. For example, in a systematic review of 29 studies, addition of common single nucleotide polymorphisms (SNPs) to other risk factors in models developed in asymptomatic individuals in the general population increased model discrimination by 0.01 to 0.06 [19]. Overall, the reported C-statistic ranged from 0.56 to 0.63 for SNPs alone and in combination with other risk factors, respectively [19]. Further studies are warranted to evaluate whether employing genetic risk profiling beyond established risk factors can be useful to identify individuals at high colorectal cancer risk.

Our work has several strengths. The EPIC study provided an ideal setting to develop a lifestyle-based colorectal cancer risk prediction model, given its large sample size, various population backgrounds and a long follow-up time of over 20 years. Furthermore, the study provided a variety of objectively measured anthropometric data along with dietary and lifestyle information. Therefore, the current model is the first developed on a European-wide study population sample, allowing assessment of risk across a broad range of diet and lifestyle behaviours. Given the large sample size, we were also able to validate the risk scores in an independent subset of the EPIC populations. Additionally, we derived the colorectal cancer risk estimates empirically following state-of-the-art and novel machine learning approaches, i.e. random survival forest, considering various predictors simultaneously and the gradient in risk across the full distribution of risk levels. Finally, we considered model application and suggested a nomogram and a web-tool to enable risk communication. Several potential limitations of our study warrant discussion. First, we derived the risk equations based on a study population comprising of volunteers. Volunteer-based studies are prone to include individuals who are often more likely to have favourable exposure and health profiles compared to those who do not. Thus, higher prevalence of healthy behaviours in our sample as compared to the general population could have resulted in overestimated absolute risk estimates. Second, with the exception of age and the anthropometric measures, we relied on data of self-reported predictors and routinely collected cancer outcomes. Though any risk prediction tool made publicly available online would also rely on self-reported data, more accurate risk factor ascertainment would possibly improve overall model discrimination and calibration. Nevertheless, our model has shown a good discrimination and excellent calibration. Third, dietary data was collected using food frequency questionnaires as a commonly applied dietary assessment method in epidemiology, however future model application should consider further adaptation and feasibility assessment to facilitate model communication in practice. Fourth, we based analyses on lifestyle information collected at study baseline and, therefore, could not account for potential behavioural changes during study follow-up. Finally, the model was developed based on data available in the EPIC cohort and did not include some potentially important predictors, such as NSAID use or family history of colorectal cancer. However, we have conducted a sensitivity analysis using data from study centres that collected these data and the model performance was not altered.

Conclusions

Despite being one of the leading causes of cancer morbidity and mortality, colorectal cancer is largely preventable. LiFeCRC score based on age and lifestyle data accurately identifies individuals at risk for incident colorectal cancer in European populations and could contribute to improved prevention through motivating lifestyle change at the individual level.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12916-020-01826-0.

Additional file 1: Supplementary Table 1. Baseline characteristics of participants with available information on NSAID use and colorectal cancer family history. **Supplementary Table 2.** TRIPOD Checklist - Prediction Model Development and Validation. **Supplementary Table 3.** Factors considered for inclusion in the LiFeCRC score. **Supplementary Table 4.** LifeCRC model selection characteristics across pre-defined risk categories in the derivation and validation samples. **Supplementary Table 5.** Added predictive performance for age and lifestyle-based (LiFeCRC) colorectal cancer risk prediction models. Results are stratified by age groups. **Supplementary Table 6.** Sensitivity analysis by subgroups.

Additional file 2: Supplementary Figure 1. Workflow of risk model development and validation. **Supplementary Figure 2.** Survival function of average predictor values of the derivation cohort. **Supplementary Figure 3.** Discrimination and relative variable importance based on Random Survival Forest models for colorectal cancer prediction. **Supplementary Figure 4.** Random Survival Forest colorectal cancer full model calibration. **Supplementary Figure 5.** Model performance comparison the LiFeCRC score and a colorectal cancer risk model including only age. (a) Calibration plot of predicted 10-year colorectal cancer risk for a model that included only age and the LiFeCRC score model with additional lifestyle predictors (waist circumference, body height, daily alcohol consumption, smoking, physical activity, and daily intake of vegetables, dairy products and red meat). (b) Decision curves illustrating net benefit of prediction models for a range of colorectal cancer risk thresholds, used to decide about further treatment or intervention. Decisions curves are shown for different models: none treatment, all treatment, treatment based on the age-model, treatment based on the LiFeCRC model. **Supplementary Figure 6.** Predicted 10-year absolute risk of colorectal cancer for a healthy and unhealthy lifestyle. Risk across different age-groups and a constant body height of 166 cm. *Unhealthy lifestyle:* waist circumference of 100 cm, high daily alcohol consumption, smoker, physically inactive, 80 g daily vegetable intake, 70 g daily dairy products intake, 60 g daily processed meat intake, and 90 g daily sugar and confectionary intake. *Healthy lifestyle:* waist circumference of 70 cm, low daily alcohol consumption, non-smoker, physically active, 430 g daily vegetable intake, 630 g daily dairy products intake, 0 g daily processed meat intake, and 5 g daily sugar and confectionary intake. **Supplementary Figure 7.** Full model performance including NSAID use and family history of colorectal cancer.

Abbreviations

AICR: American Institute for Cancer Research; BMI: Body mass index; C-index: Harrell's concordance index; CRC: Colorectal cancer; EPIC: European Prospective Investigation into Cancer and Nutrition; ICD: International Statistical Classification of Diseases; IQR: Interquartile range; LASSO: Least absolute shrinkage and selection operator; LiFeCRC score: Lifestyle-based prediction model for colorectal cancer risk; NRI$^{>0}$: Continuous net reclassification improvement; NSAID: Non-steroidal anti-inflammatory drug; RSF: Random survival forest; SD: Standard deviation; SNPs: Single nucleotide polymorphisms; TRIPOD: Transparent Reporting of a multivariable Prediction model for Individual Prognosis or Diagnosis; UK: United Kingdom; WCRF: World Cancer Research Fund

Acknowledgements

The authors express special thanks to Ellen Kohlsdorf (EPIC-Potsdam, Germany) and Bertrand Hemon (IARC-Lyon, France) for their work on data management. We thank the Human Study Centre (HSC) of the German Institute of Human Nutrition Potsdam-Rehbrücke, and the head of the HSC, Manuela Bergmann, for the contribution to the study design and leading the underlying processes of data generation. We are grateful to all the participants who have been part of the project and to the many members of the study teams at different study centers who have enabled this research.

Authors' contributions

KA designed the current study as the principal investigator; RK, TK, AT, BBM, MJ and ER were involved in the EPIC study conception and design; KA drafted the plans for the data analyses; RR conducted statistical analysis under the supervision of KA; KA drafted the manuscript. All authors were involved in interpretation of the results and revision of the manuscript, and all approved the final version of the manuscript. KA and RR have full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. KA is the guarantor. The corresponding author attests that all the listed authors meet the authorship criteria and that no others meeting the criteria have been omitted.

Funding

This work was supported by the German Research Foundation (DFG) (grant AL 1784/3-1), which funded the research position of Dr. Aleksandrova for organizing study conduct and analysis. The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); Deutsche Krebshilfe, Deutsches Krebsforschungszentrum (DKFZ) and Federal Ministry of Education and Research (BMBF) (Germany); Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS), Instituto de salud Carlos III PI13/00061 to Granada; PI13/01162 to EPIC-Murcia, Regional Governments of Andalucía, Asturias, Basque Country, Murcia (no. 6236), Navarra and Catalonia (Catalan Institute of Oncology – ICO-IDIBELL) (Spain); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (C864/A14136 to EPIC-Norfolk and C8221/A19170 to EPIC-Oxford), Medical Research Council (MR/N003284/1 and MC-UU_12015/1 to EPIC-Norfolk and MR/M012190/1 to EPIC-Oxford) (United Kingdom). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

EPIC data are available for investigators who seek to answer important questions on health and disease in the context of research projects that are consistent with the legal and ethical standard practices of IARC/WHO (World Health Organization) and the EPIC centres. The primary responsibility for accessing the data belongs to the EPIC centres that provided them. For information on how to submit an application for gaining access to EPIC data and/or biospecimens, please follow the instructions at http://epic.iarc.fr/access/index.php.

Ethics approval and consent to participate

This study complies with the Declaration of Helsinki. Ethical review boards of IARC (International Agency for Research on Cancer) and all local institutions where participants had been recruited gave approval for the study, and all participants gave written informed consent for data collection and storage, as well as individual follow-up.

Consent for publication

Not applicable.

Competing interests

The authors have read the BMC Medicine editorial policy on competing interests and declare they have no relevant conflicts of interest to declare. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Author details

[1]Nutrition, Immunity and Metabolism Senior Scientist Group, Department of Nutrition and Gerontology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany. [2]Institute of Nutritional Science, University of Potsdam, Potsdam, Germany. [3]Department of Epidemiological Methods and Etiological Research, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany. [4]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [5]International Agency for Research on Cancer, World Health Organization, Lyon, France. [6]National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. [7]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [8]Department of Public Health, Aarhus University, Aarhus, Denmark. [9]Danish Cancer Society Research Center, Copenhagen, Denmark. [10]CESP, Faculté de Medicine, Université Paris-Saclay, Villejuif, France. [11]Institut Gustave Roussy, Villejuif, France. [12]Dipartimento di Statistica, Informatica e Applicazioni "G. Parenti" (DISIA), University of Florence,

Florence, Italy. [13]Hellenic Health Foundation, Athens, Greece. [14]2nd Pulmonary Medicine Department, School of Medicine, National and Kapodistrian University of Athens, "ATTIKON" University Hospital, Haidari, Greece. [15]EPIC Centre of Naples, Dipartimento di Medicina Clinica e Chirurgia, University of Naples Federico II, Naples, Italy. [16]Cancer Risk Factors and Lifestyle Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network – ISPRO, Florence, Italy. [17]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy. [18]Unit of Cancer Epidemiology, Città della Salute e della Scienza University-Hospital and Center for Cancer Prevention (CPO), Turin, Italy. [19]Cancer Registry and Histopathology Department, Provincial Health Authority (ASP), Ragusa, Italy. [20]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands. [21]Department of Community Medicine, Health Faculty, UiT-the Arctic university of Norway, Tromsø, Norway. [22]Escuela Andaluza de Salud Pública (EASP), Granada, Spain. [23]Instituto de Investigación Biosanitaria ibs. GRANADA, Granada, Spain. [24]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [25]Universidad de Granada, Granada, Spain. [26]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain. [27]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Translational Research Laboratory, Catalan Institute of Oncology (ICO), Barcelona, Spain. [28]Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain. [29]Navarra Public Health Institute, Pamplona, Spain. [30]Navarra Institute for Health Research (IdiSNA), Pamplona, Spain. [31]Public Health Directorate, Asturias, Spain. [32]Ministry of Health of the Basque Government, Public Health Division of Gipuzkoa, Biodonostia Health Research Institute, Donostia-San Sebastian, Spain. [33]Department of Clinical Sciences, Division of Oncology and Pathology, Lund University, Lund, Sweden. [34]Department of Clinical Sciences in Malmö, Lund University, Lund, Sweden. [35]Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden. [36]Wallenberg Centre for Molecular Medicine, Umeå University, Umeå, Sweden. [37]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [38]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece.

Received: 19 June 2020 Accepted: 23 October 2020
Published online: 04 January 2021

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
2. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017; 66(4):683–91.
3. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, Gavin A, Visser O, Bray F. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer. 2018; 103:356–87.
4. Islami F, Goding Sauer A, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ, McCullough ML, Patel AV, Ma J, Soerjomataram I, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA Cancer J Clin. 2018;68(1):31–54.
5. Zhang FF, Cudhea F, Shan Z, Michaud DS, Imamura F, Eom H, Ruan M, Rehm CD, Liu J, Du M, et al. Preventable cancer burden associated with poor diet in the United States. JNCI Cancer Spectr. 2019;3(2):pkz034.
6. Bingham S. The fibre-folate debate in colo-rectal cancer. Proc Nutr Soc. 2006;65(1):19–23.
7. Bingham SA, Day NE, Luben R, Ferrari P, Slimani N, Norat T, Clavel-Chapelon F, Kesse E, Nieters A, Boeing H, et al. Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. Lancet. 2003;361(9368): 1496–501.
8. Ferrari P, Jenab M, Norat T, Moskal A, Slimani N, Olsen A, Tjonneland A, Overvad K, Jensen MK, Boutron-Ruault MC, et al. Lifetime and baseline alcohol intake and risk of colon and rectal cancers in the European Prospective Investigation into Cancer and Nutrition (EPIC). Int J Cancer. 2007;121(9):2065–72.
9. Jenab M, Ferrari P, Slimani N, Norat T, Casagrande C, Overad K, Olsen A, Stripp C, Tjonneland A, Boutron-Ruault MC, et al. Association of nut and seed intake with colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition. Cancer Epidemiol Biomark Prev. 2004;13(10):1595–603.
10. Leufkens AM, van Duijnhoven FJ, Siersema PD, Boshuizen HC, Vrieling A, Agudo A, Gram IT, Weiderpass E, Dahm C, Overvad K, et al. Cigarette smoking and colorectal cancer risk in the EPIC study. Clin Gastroenterol Hepatol. 2011;9(2):137–44.
11. Norat T, Bingham S, Ferrari P, Slimani N, Jenab M, Mazuir M, Overvad K, Olsen A, Tjonneland A, Clavel F, et al. Meat, fish, and colorectal cancer risk: the European Prospective Investigation into Cancer and Nutrition. J Natl Cancer Inst. 2005;97(12):906–16.
12. Pischon T, Lahmann PH, Boeing H, Friedenreich C, Norat T, Tjonneland A, Halkjaer J, Overvad K, Clavel-Chapelon F, Boutron-Ruault MC, et al. Body size and risk of colon and rectal cancer in the European Prospective Investigation Into Cancer and Nutrition (EPIC). J Natl Cancer Inst. 2006; 98(13):920–31.
13. van Duijnhoven FJ, Bueno-De-Mesquita HB, Ferrari P, Jenab M, Boshuizen HC, Ros MM, Casagrande C, Tjonneland A, Olsen A, Overvad K, et al. Fruit, vegetables, and colorectal cancer risk: the European Prospective Investigation into Cancer and Nutrition. Am J Clin Nutr. 2009;89(5):1441–52.
14. Vieira AR, Abar L, Chan DSM, Vingeliene S, Polemiti E, Stevens C, Greenwood D, Norat T. Foods and beverages and colorectal cancer risk: a systematic review and meta-analysis of cohort studies, an update of the evidence of the WCRF-AICR Continuous Update Project. Ann Oncol. 2017; 28(8):1788–802.
15. World Cancer Research Fund/American Institute for Cancer Research Diet, Nutrition, Physical Activity and Cancer: A Global Perspective. Continuous Update Project Expert Report 2018. Available online: http://dietandcancerreport.org. Accessed 2 Mar 2020.
16. Gonzalez CA, Riboli E. Diet and cancer prevention: contributions from the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Eur J Cancer. 2010;46(14):2555–62.
17. Freedman AN, Seminara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, Pfeiffer RM. Cancer risk prediction models: a workshop on development, evaluation, and application. J Natl Cancer Inst. 2005;97(10):715–23.
18. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk prediction models for colorectal cancer: a systematic review. Cancer Prev Res (Phila). 2016;9(1):13–26.
19. McGeoch L, Saunders CL, Griffin SJ, Emery JD, Walter FM, Thompson DJ, Antoniou AC, Usher-Smith JA. Risk prediction models for colorectal cancer incorporating common genetic variants: a systematic review. Cancer Epidemiol Biomark Prev. 2019;8(10):1580–93.
20. Smith T, Muller DC, Moons KGM, Cross AJ, Johansson M, Ferrari P, Fagherazzi G, Peeters PHM, Severi G, Husing A, et al. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. Gut. 2019; 68(4):672–83.
21. Ma GK, Ladabaum U. Personalizing colorectal cancer screening: a systematic review of models to predict risk of colorectal neoplasia. Clin Gastroenterol Hepatol. 2014;12(10):1624–34. e1621.
22. Usher-Smith JA, Harshfield A, Saunders CL, Sharp SJ, Emery J, Walter FM, Muir K, Griffin SJ. External validation of risk prediction models for incident colorectal cancer using UK Biobank. Br J Cancer. 2018;118(5):750–9.
23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Br J Surg. 2015;102(3):148–58.
24. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Stat Med. 2016;35(7):1159–77.
25. Cueto-Lopez N, Garcia-Ordas MT, Davila-Batista V, Moreno V, Aragones N, Alaiz-Rodriguez R. A comparative study on feature selection for a risk prediction model for colorectal cancer. Comput Methods Prog Biomed. 2019;177:219–29.
26. Van Belle V, Van Calster B. Visualizing risk prediction models. PLoS One. 2015;10(7):e0132614.
27. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondière UR, Hémon B, Casagrande C, Vignat J, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2007;5(6b):1113–24.

28. Aleksandrova K, Pischon T, Jenab M, Bueno-de-Mesquita HB, Fedirko V, Norat T, Romaguera D, Knuppel S, Boutron-Ruault MC, Dossus L, et al. Combined impact of healthy lifestyle factors on colorectal cancer: a large European cohort study. BMC Med. 2014;12:168.
29. Slimani N, Deharveng G, Unwin I, Southgate DA, Vignat J, Skeie G, Salvini S, Parpinel M, Moller A, Ireland J, et al. The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study. Eur J Clin Nutr. 2007; 61(9):1037–56.
30. Kaaks R, Riboli E. Validation and calibration of dietary intake measurements in the EPIC project: methodological considerations. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol. 1997;26(Suppl 1): S15–25.
31. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73.
32. Laurin C, Boomsma D, Lubke G. The use of vector bootstrapping to improve variable selection precision in Lasso models. Stat Appl Genet Mol Biol. 2016; 15(4):305–20.
33. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019. URL https://www.R-project.org/.
34. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5): 1–13. URL http://www.jstatsoft.org/v39/i05/.
35. Therneau T. A Package for Survival Analysis in R_. R package version 3.2–3, 2020. URL: https://CRAN.R-project.org/package=survival.
36. Harrell FE Jr. rms: Regression Modeling Strategies. R package version 6.0–0. 2020. https://CRAN.R-project.org/package=rms.
37. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R. R package version 1.5.0. 2020. https://CRAN.R-project.org/package=shiny.
38. Chang W, Ribeiro BB. shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.1. 2018. https://CRAN.R-project.org/package=shinydashboard.
39. Austin PC, Latouche A, Fine JP. A review of the use of time-varying covariates in the Fine-Gray subdistribution hazard competing risk regression model. Stat Med. 2020;39(2):103–13.
40. Muller-Riemenschneider F, Holmberg C, Rieckmann N, Kliems H, Rufer V, Muller-Nordhorn J, Willich SN. Barriers to routine risk-score use for healthy primary care patients: survey and qualitative study. Arch Intern Med. 2010; 170(8):719–24.
41. Usher-Smith JA, Sharp SJ, Luben R, Griffin SJ. Development and validation of lifestyle-based models to predict incidence of the most common potentially preventable cancers. Cancer Epidemiol Biomark Prev. 2019;28(1):67–75.
42. Emmons KM, McBride CM, Puleo E, Pollak KI, Clipp E, Kuntz K, Marcus BH, Napolitano M, Onken J, Farraye F, et al. Project PREVENT: a randomized trial to reduce multiple behavioral risk factors for colon cancer. Cancer Epidemiol Biomark Prev. 2005;14(6):1453–9.
43. Smith-Warner SA, Elmer PJ, Tharp TM, Fosdick L, Randall B, Gross M, Wood J, Potter JD. Increasing vegetable and fruit intake: randomized intervention and monitoring in an at-risk population. Cancer Epidemiol Biomark Prev. 2000;9(3):307–17.
44. Caswell S, Anderson AS, Steele RJ. Bowel health to better health: a minimal contact lifestyle intervention for people at increased risk of colorectal cancer. Br J Nutr. 2009;102(11):1541–6.
45. Robb KA, Power E, Kralj-Hans I, Atkin WS, Wardle J. The impact of individually-tailored lifestyle advice in the colorectal cancer screening context: a randomised pilot study in North-West London. Prev Med. 2010; 51(6):505–8.
46. Craigie AM, Caswell S, Paterson C, Treweek S, Belch JJ, Daly F, Rodger J, Thompson J, Kirk A, Ludbrook A, et al. Study protocol for BeWEL: the impact of a BodyWEight and physicaL activity intervention on adults at risk of developing colorectal adenomas. BMC Public Health. 2011;11:184.
47. Anderson AS, Mackison D, Boath C, Steele R. Promoting changes in diet and physical activity in breast and colorectal cancer screening settings: an unexplored opportunity for endorsing healthy behaviors. Cancer Prev Res (Phila). 2013;6(3):165–72.
48. Usher-Smith JA, Silarova B, Ward A, Youell J, Muir KR, Campbell J, Warcaba J. Incorporating cancer risk information into general practice: a qualitative study using focus groups with health professionals. Br J Gen Pract. 2017; 67(656):e218–26.
49. Demark-Wahnefried W, Rock CL, Patrick K, Byers T. Lifestyle interventions to reduce cancer risk and improve outcomes. Am Fam Physician. 2008;77(11): 1573–8.
50. Guo LW, Li N, Chen HD, Lyu ZY, Feng XS, Wei LP, Li X, Wen Y, Lu M, Dai M. Progress in construction and verification of colorectal cancer risk prediction models: a systematic review. Zhonghua Yu Fang Yi Xue Za Zhi. 2019;53(6): 603–10.
51. Williams TG, Cubiella J, Griffin SJ, Walter FM, Usher-Smith JA. Risk prediction models for colorectal cancer in people with symptoms: a systematic review. BMC Gastroenterol. 2016;16(1):63.
52. Carr PR, Weigl K, Edelmann D, Jansen L, Chang-Claude J, Brenner H, Hoffmeister M. Estimation of absolute risk of colorectal cancer based on healthy lifestyle, genetic risk, and colonoscopy status in a population-based study. Gastroenterology. 2020;159(1):129–38.
53. Aleksandrova K, Nimptsch K, Pischon T. Influence of obesity and related metabolic alterations on colorectal cancer risk. Curr Nutr Rep. 2013;2(1):1–9.
54. Aleksandrova K, Schlesinger S, Fedirko V, Jenab M, Bueno-de-Mesquita B, Freisling H, Romieu I, Pischon T, Kaaks R, Gunter MJ, et al. Metabolic mediators of the association between adult weight gain and colorectal cancer: data from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. Am J Epidemiol. 2017;185(9):751–64.
55. Aleksandrova K, Drogan D, Boeing H, Jenab M, Bas Bueno-de-Mesquita H, Jansen E, van Duijnhoven FJ, Rinaldi S, Fedirko V, Romieu I, et al. Adiposity, mediating biomarkers and risk of colon cancer in the European prospective investigation into cancer and nutrition study. Int J Cancer. 2014;134(3):612–21.
56. Colditz GA, Atwood KA, Emmons K, Monson RR, Willett WC, Trichopoulos D, Hunter DJ. Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. Cancer Causes Control. 2000;11(6):477–88.
57. Shin A, Joo J, Yang HR, Bak J, Park Y, Kim J, Oh JH, Nam BH. Risk prediction model for colorectal cancer: National Health Insurance Corporation study, Korea. PLoS One. 2014;9(2):e88079.
58. Garcia J, Quintana-Domeque C. The evolution of adult height in Europe: a brief note. Econ Hum Biol. 2007;5(2):340–9.
59. Gorczyca AM, Eaton CB, LaMonte MJ, Garcia DO, Johnston JD, He K, Bidulescu A, Goodman D, Groessl E, Lane D, et al. Association of physical activity and sitting time with incident colorectal cancer in postmenopausal women. Eur J Cancer Prev. 2018;27(4):331–8.
60. Aleksandrova K, Jenab M, Leitzmann M, Bueno-de-Mesquita B, Kaaks R, Trichopoulou A, Bamia C, Lagiou P, Rinaldi S, Freisling H, et al. Physical activity, mediating factors and risk of colon cancer: insights into adiposity and circulating biomarkers from the EPIC cohort. Int J Epidemiol. 2017;46(6): 1823–35.
61. Wells BJ, Kattan MW, Cooper GS, Jackson L, Koroukian S. Colorectal cancer predicted risk online (CRC-PRO) calculator using data from the multi-ethnic cohort study. J Am Board Fam Med. 2014;27(1):42–55.
62. Driver JA, Gaziano JM, Gelber RP, Lee IM, Buring JE, Kurth T. Development of a risk score for colorectal cancer in men. Am J Med. 2007;120(3):257–63.
63. Freedman AN, Slattery ML, Ballard-Barbash R, Willis G, Cann BJ, Pee D, Gail MH, Pfeiffer RM. Colorectal cancer risk prediction tool for white men and women without known susceptibility. J Clin Oncol. 2009;27(5):686–93.
64. Park Y, Freedman AN, Gail MH, Pee D, Hollenbeck A, Schatzkin A, Pfeiffer RM. Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. J Clin Oncol. 2009;27(5):694–8.

Publisher's Note