


# A Risk Prediction Model for Sporadic CRC Based on Routine Lab Results

Ben Boursi<sup>1,2,3,5</sup>  · Ronac Mamtani<sup>2,3,4</sup> · Wei-Ting Hwang<sup>2,3</sup> · Kevin Haynes<sup>2,3</sup> · Yu-Xiao Yang<sup>1,2,3</sup>

Received: 15 September 2015 / Accepted: 5 February 2016 / Published online: 19 February 2016  
© Springer Science+Business Media New York 2016

## Abstract

**Background** Current risk scores for colorectal cancer (CRC) are based on demographic and behavioral factors and have limited predictive values.

**Aim** To develop a novel risk prediction model for sporadic CRC using clinical and laboratory data in electronic medical records.

**Methods** We conducted a nested case–control study in a UK primary care database. Cases included those with a diagnostic code of CRC, aged 50–85. Each case was matched with four controls using incidence density sampling. CRC predictors were examined using univariate conditional logistic regression. Variables with  $p$  value  $<0.25$  in the univariate analysis were further evaluated in

multivariate models using backward elimination. Discrimination was assessed using receiver operating curve. Calibration was evaluated using the McFadden's R<sup>2</sup>. Net reclassification index (NRI) associated with incorporation of laboratory results was calculated. Results were internally validated.

**Results** A model similar to existing CRC prediction models which included age, sex, height, obesity, ever smoking, alcohol dependence, and previous screening colonoscopy had an AUC of 0.58 (0.57–0.59) with poor goodness of fit. A laboratory-based model including hematocrit, MCV, lymphocytes, and neutrophil–lymphocyte ratio (NLR) had an AUC of 0.76 (0.76–0.77) and a McFadden's R<sup>2</sup> of 0.21 with a NRI of 47.6 %. A combined model including sex, hemoglobin, MCV, white blood cells, platelets, NLR, and oral hypoglycemic use had an AUC of 0.80 (0.79–0.81) with a McFadden's R<sup>2</sup> of 0.27 and a NRI of 60.7 %. Similar results were shown in an internal validation set.

This work was performed in partial fulfillment of the requirements for a Ph.D. degree of Ben Boursi, Sackler Faculty of Medicine, Tel-Aviv University, Israel.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10620-016-4081-x) contains supplementary material, which is available to authorized users.

✉ Ben Boursi  
bben217@gmail.com

Ronac Mamtani  
ronac.mamtani@uphs.upenn.edu

Wei-Ting Hwang  
whwang@mail.med.upenn.edu

Kevin Haynes  
khaynes@mail.med.upenn.edu

Yu-Xiao Yang  
yangy@mail.med.upenn.edu

<sup>2</sup> Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>3</sup> Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>4</sup> Division of Hematology/Oncology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>5</sup> Tel-Aviv University, 69978 Tel Aviv, Israel

<sup>1</sup> Division of Gastroenterology, Perelman School of Medicine at the University of Pennsylvania, 733 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA

**Conclusion** A laboratory-based risk model had good predictive power for sporadic CRC risk.

**Keywords** Colon · Cancer · Risk model · Screening

## Introduction

Colorectal cancer (CRC) is the third most common cancer in men and the second in women worldwide with an estimated lifetime risk among western populations of 5–6 %. According to data from the American Cancer Society (ACS), approximately 132,700 new CRC cases are to be diagnosed in the USA in 2015 and approximately 49,700 are expected to die of the disease [1–3]. Genetic syndromes, such as familial adenomatous polyposis (FAP), hereditary non-polyposis colorectal cancer (HNPCC), or MUTYH-associated polyposis (MAP), are estimated to account for 5 % of CRC cases. Additional 15–20 % of patients have a familial history of the disease that might suggest hereditary contribution. However, the majority of patients (about 75 %) have sporadic disease, with no family history of CRC [4, 5]. In this population, major risk factors that may influence the development of the disease include age, male gender, obesity, diet high in fat and low in fiber, sedentary lifestyle, cigarette smoking, and alcohol consumption as well as medical history of inflammatory bowel disease (IBD), diabetes mellitus, and insulin resistance [6–8]. The use of aspirin and nonsteroidal anti-inflammatory drugs (NSAIDs) and hormone replacement therapy (HRT) were shown to decrease the risk [9].

Despite the clear recommendations in the medical literature, and availability of screening tests proven to reduce the incidence and mortality from CRC [10, 11], compliance among at-risk populations remains low. According to data of the ACS from 2013, only 59 % of adults 50 years of age and older underwent any CRC screening test within the recommended time intervals [12, 13].

There is an ongoing effort to enhance the risk stratification of individuals, increase compliance for initial screening, and streamline the use of surveillance colonoscopy in order to reduce the number of unnecessary tests. Risk scores are commonly used in medicine to quantify a person's risk of developing a disease. Knowledge of individual's risk of CRC could be used to develop risk-tailored strategies to improve the efficiency of screening. While such scores exist for breast and prostate cancers [14, 15], limited data are available for CRC [16–22]. Although prior studies found difference in disease incidence among subjects that were categorized as high risk compared to those categorized as low-risk group, they had low discrimination ability with area under the curve (AUC) of 0.6–0.69. Additional limitations of these studies included incomplete

assessment of colon cancer risk factors [16–21]; restricted age range of subjects [18, 20, 21]; selection bias in studies that evaluated only subjects that were self-referred for screening [18–20]; and lack of controlling for previous CRC screening or family history of CRC [17]. The above-mentioned limitations affected not only model performance but also model generalizability and validity.

The aim of the current study was to develop and validate a risk prediction model for sporadic CRC that incorporates clinical and laboratory data using a large population-representative electronic medical records (EMR) database. If validated, the model could be used to generate patient risk data automatically; such information can then be easily available for physicians and linked with patient directed interventions.

## Methods

### Study Design

We conducted a nested case–control study with incidence density sampling using the health improvement network (THIN), a large EMR database from the UK. Case–control studies with incidence density sampling of controls yield odds ratios (ORs) that are statistically unbiased estimates of incidence-rate ratios (or hazard ratio) from a corresponding cohort study with proportional hazard analysis [23]. The study was approved by the Institutional Review Board at the University of Pennsylvania and by the Scientific Review Committee of THIN.

### Data Source

The THIN database contains comprehensive medical records on approximately ten million patients (5.7 % of the UK population) treated by general practitioners in 570 practices, providing data on exposures and potential confounders important for CRC risk assessment. Registration date is defined as the date when patients were first registered with a practice in THIN, and Vision date is the date that a practice began using in-practice Vision software that collects information for the THIN database [24]. Each medical diagnosis is defined using Read diagnostic codes, the standard coding system used by general practices in the UK. All practices contributing data to THIN are instructed to follow a standardized protocol of entering information. Data quality is monitored through routine analysis of the entered data [25, 26]. The database was shown to be representative of the UK population with excellent quality of information [27]. Cancer rates in THIN, including colon cancer, were shown to be comparable to those reported in cancer registry data [28].

## Study Cohort

All people receiving medical care from 1995 to 2013 from a THIN practitioner were eligible for inclusion. Subjects with a diagnosis of CRC syndromes, familial history of CRC, or IBD were excluded in order to focus on sporadic CRC. Patients without acceptable medical records were excluded (i.e., patients with incomplete documentation or out of sequence date of birth, registration date, date of death, or date of exit from the database). Follow-up started at the later of either the Vision date or 6 months after the date at which the patient registered with the general practitioner [29], and ended on the earliest of CRC diagnosis date, date of death, transferring out of the database, or the end date of the database.

## Case Selection

Cases were defined as all individuals in the cohort with at least one Read code for CRC during the follow-up period that was 50–85 years old at the time of diagnosis. We limited our study population to individuals in this age group since 90 % of people diagnosed with CRC are above the age of 50 [2] and because current guidelines recommend screening in adults beginning at the age of 50 and continuing until age 75, and on an individualized basis in adults between the age of 76 and 85 [6]. Subjects who were diagnosed within the first 6 months after registration were excluded in order to avoid prevalent cases [29]. In addition, in order to predict the risk of early-stage disease and since 20 % of CRC cases in the UK have a distant spread at the time of diagnosis [30] and 75 % of them will die within 2.5 years [2], we excluded subjects with death date within 2.5 years from diagnosis date.

## Selection of Controls

The eligible control pool for each case comprised all individuals without a diagnosis of CRC at the time of sampling and without a history of previous colectomy. Up to four eligible controls were randomly selected and matched with the case on practice site and start date of follow-up. Controls were assigned the same index date as their matched cases.

## Exposures and Covariates

We examined a comprehensive list of potential and known CRC risk factors based on literature review (supplementary index 1). All covariates were measured prior to index date. The risk factors were divided into five categories: anthropomorphic and lifestyle parameters (such as obesity and smoking history), health care utilization (including previous screening for CRC), medical comorbidities (such as diabetes mellitus), medications (such as aspirin and

nonsteroidal anti-inflammatories) as categorical variables of any use before index date, and laboratory results [such as complete blood count (CBC) and inflammatory markers] as continuous variables. For laboratory results, we used both last values within the year before index date and the difference between the last two values before index date in order to evaluate the intra-individual trends.

## Statistical Analysis

The entire study cohort was randomly divided in a 2:1 ratio in order to generate a test and validation sets. The association between each variable and CRC risk was evaluated using a univariate conditional logistic regression analysis to estimate ORs and 95 % confidence intervals (CIs). All variables associated with a  $p$  value  $<0.25$  in the univariate analysis were further evaluated in the multivariate model [31]. Laboratory test-associated variables with  $>67$  % missing data were excluded. We performed a complete case analysis with the remaining variables. Three models were created: a model based only on variables used in previous risk models (reference model); a model based only on laboratory results (laboratory-based model); and a model based on all parameters (combined model). For the multivariate logistic regression in each one of the models, we used backward elimination for variable selection with  $p$  values of  $<0.001$  and  $>0.05$  as inclusion and exclusion criteria, respectively. Additionally, we repeated the multivariate models after testing continuous variables for linearity and correction using fractional polynomials (FP) of second degree with powers  $-2$ ,  $-1$ ,  $-0.5$ ,  $0$ ,  $0.5$ ,  $1$ ,  $2$ ,  $3$ , to improve the fit of the models [32, 33].

The models were tested for collinearity [variance inflation factor (VIF)  $>10$ ] and two-way interactions. The risk for each individual was given as deciles of probability with values from 0 to 1. The calibration of each of the models was evaluated using the McFadden's  $R^2$  goodness-of-fit test with high  $p$  value ( $>0.05$ ) indicating adequately fit of the logistic function and values  $>0.2$  indicating extremely good model fits [34]. The discrimination ability of the models was calculated using the area under the receiver operating curve (ROC). Net reclassification index (NRI) [35] comparing either the combined or the laboratory-based models to the reference model was calculated using the formula:

$$\begin{aligned} &[(\text{net increase of classification for cases}/\text{total number of cases}) \\ &+ (\text{net decrease of classification for controls}/\text{total number} \\ &\text{of controls})] \times 100\%. \end{aligned}$$

The analysis was repeated in the validation set of the data. All calculations were done using STATA 13 (Stata Corp., College Station, TX, USA).

## Results

### Study Cohort and Variables

The study cohort included 22,351 CRC cases that were diagnosed between 1995 and 2013. We excluded 125 cases with family history of CRC; 3194 cases that were diagnosed before the age of 50 or after the age of 85 years old; and 5110 cases that died within 2.5 years of index date and probably had a metastatic disease at diagnosis. Eventually, we were left with 13,879 cases (9299 in the test set and 4580 in the validation set) and 54,109 matched controls (36,199 in the test set and 17,910 in the validation set). Forty-three cases had no matched controls. Characteristics of cases and controls are presented in Table 1. Results for the univariate analysis are presented in Supplementary index 2.

### Reference Model

A reference model based only on variables that were used in previous studies (age, sex, height, obesity, ever smoking, alcohol dependence and previous screening colonoscopy) demonstrated an AUC of 0.58 (95 % CI 0.57–0.59) and low goodness of fit with McFadden’s R2 of 0.03 (Fig. 1). There was no change in results when age was corrected for linearity.

### Laboratory-Based Model

For the model including only laboratory test results, the variables that remained after conducting the backward elimination are presented in Table 2. The AUC for the model based on these parameters was 0.77 (95 % CI 0.76–0.78) with a goodness of fit >0.05 (McFadden’s R2 of 0.23). We excluded creatinine and BUN from the model due to lack of biological plausibility. We repeated the

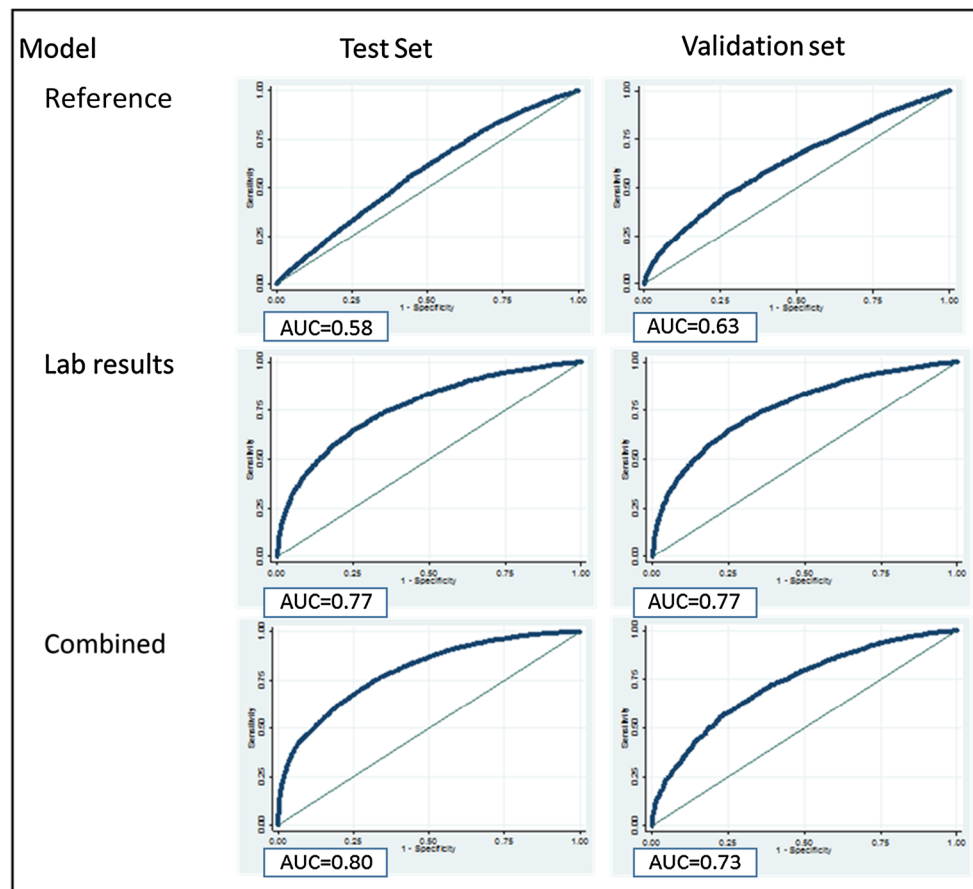
model for hematocrit [PCV (%)], MCV (in fl), lymphocytes (in billion cells/l) and NLR, with complete data available for 16,240 (35.7 %) individuals [4929 (53.0 %) cases and 11,311 (31.2 %) controls]. This model had an AUC of 0.74 (95 % CI 0.73–0.75) with a McFadden’s R2 of 0.16. The AUC improved modestly after correction for linearity (0.76, 95 % CI 0.76–0.77) and a McFadden’s R2 of 0.21. The final equation for this model is presented in supplementary index 3.

### Combined Model

For the combined model including all five groups of variables (anthropomorphic and lifestyle, health care utilization, medical comorbidities, medications, and laboratory results), the variables that remained after conducting the backward elimination are presented in Table 3. The AUC for the model based on these parameters was 0.80 (95 % CI 0.79–0.82) with a McFadden’s R2 of 0.33. We excluded BUN and spironolactone prescriptions due to lack of biological plausibility and antidepressant due to possible confounding by indication. We excluded height due to missing values. Red blood cells and lymphocytes were excluded due to collinearity. From the resulting model age, eosinophil count, and aspirin/NSAIDs, digoxin and recurrent TMP/SMX prescriptions were excluded due to lack of statistical significance. We added to the model platelets and white blood cell count, two additional blood lineages that were significant in the univariate analysis and were suggested as predictors for cancer in previous studies [36–38] and metformin use that was suggested to decrease CRC risk [39]. The final combined model included sex, hemoglobin (in g/dl), MCV, white blood cells (billion cells/l), platelets (billion/l), and NLR as well as previous prescriptions of metformin or other oral hypoglycemic medications. 13,640 (30.0 %) individuals had all laboratory results [4098 cases (44.1 %) and 9542 (26.4 %) controls].

**Table 1** Characteristics of cases and controls

Covariate	Test set			Validation set		
	Cases (9299)	Controls (36,199)	<i>p</i> value	Cases (4580)	Controls (17,910)	<i>p</i> value
Age (±SD)	69.95 (±8.67)	69.66 (±8.62)	<0.0001	69.79 (±8.66)	69.54 (±8.65)	<0.0001
Sex (% males)	5295 (56.94 %)	16,231 (44.84 %)	<0.0001	2584 (56.42 %)	8163 (45.58 %)	<0.0001
Duration of follow-up (years SD)	6.29 (±4.16)	6.30 (±4.16)	NA	6.22 (±4.09)	6.24 (±4.09)	NA
Obesity (BMI >30) (%)	2030 (21.83 %)	7332 (20.25 %)	<0.0001	1060 (23.14 %)	3630 (20.27 %)	<0.0001
Smoking (ever) (%)	4440 (47.75 %)	15,002 (41.44 %)	<0.0001	2195 (47.93 %)	7419 (41.42 %)	<0.0001
Alcohol use (%)	4942 (53.15 %)	17,963 (49.62 %)	<0.0001	2400 (52.40 %)	8713 (48.65 %)	<0.0001
Diabetes mellitus (%)	1183 (12.72 %)	3702 (10.23 %)	<0.0001	645 (14.08 %)	1834 (10.24 %)	<0.0001
Chronic aspirin/NSAIDs use for >1 year (%)	1863 (20.03 %)	7614 (21.03 %)	0.01	883 (19.28 %)	3819 (21.32 %)	0.001



**Fig. 1** Receiver operating curve for the different risk models

The AUC of the model was 0.79 (95 % CI 0.78–0.80) with a McFadden's R<sup>2</sup> of 0.26. The model reached an AUC of 0.80 (95 % CI 0.79–0.81) and a McFadden's R<sup>2</sup> of 0.27 after correction for linearity (Fig. 1). The final equation for this model is presented in supplementary index 3.

Figure 2 and Table 4 present the percent of observed CRC cases in the test set by model's probability deciles (of note, the data describe the percent within the case–control population that had 1:4 ratio between cases and controls).

We further looked at a model that contains only sex and laboratory values and might be easier to use as an automatic application. This model had an AUC of 0.79 (95 % CI 0.78–0.80) and a McFadden's R<sup>2</sup> of 0.26.

### Validation

All models were evaluated in the validation set as well. For the reference model, we had 8210 subjects (36.5 % of the validation population due to lack of height measurements). The AUC was 0.63 (95 % CI 0.61–0.64), and the McFadden's R<sup>2</sup> 0.01. For the model based on laboratory test results, we had 5792 subjects (25.8 %) with all laboratories available. The AUC was 0.77 (95 % CI 0.75–0.78)

similar to the one from the test set with McFadden's R<sup>2</sup> of 0.14. For the combined model, 4946 (22.0 %) had the entire laboratory results. The AUC was 0.73 (95 % CI 0.71–0.74), and the McFadden's R<sup>2</sup> was 0.07. Figure 2 and Table 4 present the percent of observed CRC cases in the validation set by probability deciles of the combined model.

### Net Reclassification Index

We further calculated the NRI for both the test and validation sets using the combined model compared to the model based on variables that were used in previous studies. Detailed reclassification tables are provided (supplementary index 4). The NRI was higher in both sets with values of 60.7 % for the test set (52.8 % for cases and 7.9 % for controls) and 14.7 % for the validation set (−4.9 % for cases and 19.6 % for controls). Since the model based on laboratory values had similar AUC as the overall model with the advantage of additional simplicity as an automatic model, we also calculated the NRI using this model in comparison with the model based on variables that were used in previous studies (supplementary

**Table 2** Variables selected for multivariate model based only on laboratory results

With correction to linearity		Without correction to linearity					
Covariate	Coefficient (95 % CI)	OR (95 % CI)	p value	Covariate	Coefficient (95 % CI)	OR (95 % CI)	p value
Hematocrit (PCV) (FP2 first power -1)	-0.44 [-0.61 to (-0.27)]	0.64 (0.54-0.77)	<0.0001	Hematocrit (PCV)	-0.002 (-0.02 to 0.02)	0.99 (0.98-1.02)	0.8
Hematocrit (PCV) (FP2 second power -1)	-0.13 [-0.17 to (-0.09)]	0.88 (0.84-0.92)	<0.0001				
MCV (FP2 first power 3)	-0.06 [-0.07 to (-0.04)]	0.95 (0.93-0.96)	<0.0001	MCV	-0.1 (-0.11 to (-0.09))	0.9 (0.9-0.91)	<0.0001
MCV (FP2 second power 3)	0.02 (0.02-0.03)	1.02 (1.02-1.03)	<0.0001				
Neutrophil-lymphocyte ratio (FP2 first power -0.5)	-2.11 [-2.92 to (-1.30)]	0.12 (0.05-0.27)	<0.0001	Neutrophil-lymphocyte ratio	0.18 (0.15-0.21)	1.2 (1.16-1.24)	<0.0001
Neutrophil-lymphocyte ratio (FP2 second power -0.5)	-0.32 [-0.53 to (-0.11)]	0.73 (0.59-0.9)	0.003				
Lymphocytes (FP2 first power 0)	0.20 (-0.43 to 0.84)	1.23 (0.65-2.32)	0.531	Lymphocytes	0.09 (0.05-0.12)	1.09 (1.05-1.13)	<0.0001
Lymphocytes (FP2 second power 0.5)	2.86 (0.30-5.43)	17.48 (1.34-227.3607)	0.029				
Creatinine (FP2 first power -2)	-0.0002 (-0.007 to 0.007)	0.99 (0.99-1.00)	0.96	Creatinine	0.01 (0.007-0.01)	1.01 (1.00-1.01)	<0.0001
Creatinine (FP2 second power -2)	1.99 (1.40-2.58)	7.33 (4.06-13.23)	<0.0001				
BUN (FP2 first power 0)	-0.86 [-1.14 to (-0.58)]	0.42 (0.32-0.56)	<0.0001	BUN	-0.15 [-0.18 to (-0.12)]	0.86 (0.83-0.89)	<0.0001
BUN (FP2 second power 2)	-0.24 [-0.45 to (-0.03)]	0.79 (0.64-0.97)	0.02				

index 5). The NRI was 47.6 % for the test set (47.3 % for cases and 0.3 % for controls) and 41.4 % for the validation set (12.2 % for cases and 29.2 % for controls).

**Discussion**

Using a UK population-representative dataset [27], we assessed for the first time a risk prediction model for sporadic CRC based on laboratory test results, mainly CBC and inflammatory markers, and compared it to a reference model based on variables that were previously used in CRC risk models, such as anthropomorphic and lifestyle parameters and medical comorbidities. The reference model had similar low predictive value and goodness of fit as past models with an AUC ranging between 0.58 in the test set and 0.63 in the validation set. However, models based on laboratory test results had high predictive values and discrimination with an AUC of up to 0.74 and 0.80, respectively, and high goodness of fit. The likelihood of a CRC diagnosis was 18 times higher in the highest compared to the lowest risk decile of the combined model (Fig. 2). These results were replicated in the validation set of the study.

A recent systematic review [40] that evaluated previous CRC risk models found weak discriminatory power, with AUCs ranging from 0.6 to 0.69 and large heterogeneity between studies. These models were limited by selection bias as most models used only data from subjects that underwent screening or diagnostic colonoscopies and had no information regarding individuals that were non-compliant with current screening recommendations (up to one-third of the general US population). Furthermore, some of these models focused on specific populations such as physicians, males, or city dwellers [17, 20] that differ in health literacy and the use of health care services from the general population. Thus, previous results lacked generalizability and several studies that tried to confirm the results in different populations showed an even lower discriminating ability [41]. Additionally, these models were also prone to recall bias due to the use of self-report data. Furthermore, numerous risk factors such as family history of CRC, previous screening colonoscopies, use of aspirin/NSAIDs and HRT were not assessed.

The current study had several important advantages. We used a large population-representative EMR that included information both on individuals with and without a history of previous CRC screening. The incidence of CRC in THIN was previously shown to be comparable to the incidence in the entire population of the UK as reported in cancer registry data [26, 28]. The study cohort had a long follow-up with a median of 6.2 years and a maximum of 18 years. By excluding individuals with a history of

**Table 3** Variables selected for multivariate model based on all variables

With correction to linearity			Without correction to linearity				
Covariate	Coefficient (95 % CI)	OR (95 % CI)	<i>p</i> value	Covariate	Coefficient (95 % CI)	OR (95 % CI)	<i>p</i> value
Age (FP2 first power -2)	-321.65 [-605.82 to (-37.48)]	2.0e-140 (7.9e-264 to 5.27e-17)	0.03	Age	0.05 (-0.01 to 0.12)	1.06 (0.99-1.12)	0.08
Age (FP2 second power 1)	-1.37 (-3.14 to 0.40)	0.25 (0.04-1.49)	0.13				
Sex	-0.52 [-0.79 to (-0.24)]	0.6 (0.45-0.78)	<0.0001	Sex	-0.59 [-0.85 to (-0.33)]	0.55 (0.43-0.72)	<0.0001
Height (FP2 first power 3)	0.02 (0.01-0.02)	1.02 (1.01-1.03)	0.002	Height	0.03 (0.02-0.04)	1.03 (1.02-1.05)	<0.0001
Height (FP2 second power 3)	-0.005 [-0.008 to (-0.002)]	0.99 (0.99-1.00)	0.003				
Red blood cells (FP2 first power 0)	-6.89 (-15.63 to 1.85)	0.001 (1.63e-07 to 6.38)	0.12	Red blood cells	0.47 (-0.14 to 1.07)	1.6 (0.87-2.92)	0.13
Red blood cells (FP2 second power 0)	-5.66 [-10.52 to (-0.80)]	0.003 (0.00003-0.45)	0.022				
Hemoglobin (FP2 first power 2)	-8.29 [-11.07 to (-5.51)]	0.0003 (0.00002-0.004)	<0.0001	Hemoglobin	-0.57 [-0.76 to (-0.38)]	0.57 (0.47-0.68)	<0.0001
Hemoglobin (FP2 second power 2)	7.62 (4.60-10.65)	2045.403 (99.37-42,102.85)	<0.0001				
MCV (FP2 first power 3)	0.005 (-0.03 to 0.04)	1.01 (0.97-1.04)	0.76	MCV	-0.02 (-0.06 to 0.01)	0.98 (0.94-1.01)	0.18
MCV (FP2 second power 3)	-0.002 (-0.01 to 0.1)	1.00 (0.99-1.01)	0.74				
Lymphocytes (FP2 first power 0)	0.60 (-0.51 to 1.70)	1.82 (0.60-5.47)	0.29	Lymphocytes	0.1 (0.05-0.16)	1.11 (1.05-1.17)	<0.0001
Lymphocytes (FP2 second power 0.5)	2.53 (-1.82 to 6.87)	12.51 (0.16-964.84)	0.25				
Neutrophil-lymphocyte ratio (FP2 first power -0.5)	-1.81 [-3.38 to (-0.25)]	0.16 (0.03-0.78)	0.02	Neutrophil- lymphocyte ratio	0.16 (0.1-0.23)	1.18 (1.11-1.25)	<0.0001
Neutrophil-lymphocyte ratio (FP2 second power -0.5)	-0.21 [-0.62 to (-0.20)]	0.81 (0.54-1.22)	0.31				
Eosinophils (FP2 first power -2)	0.0005 (-0.0003 to 0.001)	1.00 (0.99-1.00)	0.24	Eosinophils	-0.06 (-0.27 to 0.15)	0.94 (0.76-1.16)	0.57
Eosinophils (FP2 second power -0.5)	-0.04 (-0.20 to 0.12)	0.96 (0.82-1.13)	0.64				
BUN (FP2 first power 0)	-0.75 [-1.21 to (-0.29)]	0.47 (0.30-0.75)	0.001	BUN	-0.15 [-0.19 to (-0.11)]	0.86 (0.83-0.9)	<0.0001
BUN (FP2 second power 2)	-0.30 (-0.64 to 0.04)	0.74 (0.53-1.04)	0.08				

**Table 3** continued

With correction to linearity		Without correction to linearity					
Covariate	Coefficient (95 % CI)	OR (95 % CI)	<i>p</i> value	Covariate	Coefficient (95 % CI)	OR (95 % CI)	<i>p</i> value
Oral hypoglycemic other than metformin	-0.53 [-0.84 to (-0.22)]	0.59 (0.43–0.80)	0.001	Oral hypoglycemic other than metformin	-0.44 [-0.74 to (-0.14)]	0.64 (0.48–0.87)	0.004
Antidepressants	-0.36 [-0.56 to (-0.17)]	0.69 (0.57–0.85)	<0.0001	Antidepressants	-0.35 [-0.54 to (-0.16)]	0.71 (0.58–0.86)	<0.0001
Digoxin	0.18 (-0.27 to 0.63)	1.20 (0.76–1.88)	0.43	Digoxin	0.2 (-0.24 to 0.64)	1.22 (0.78–1.89)	0.38
Spirinolactone	-0.92 [-1.53 to (-0.31)]	0.40 (0.22–0.73)	0.003	Spirinolactone	-0.84 [-1.43 to (-0.26)]	0.43 (0.24–0.77)	0.005
Recurrent TMP/SMX prescriptions	-0.11 (-0.68 to 0.46)	0.90 (0.51–1.59)	0.71	Recurrent TMP/SMX prescriptions	-0.12 (-0.69 to 0.45)	0.89 (0.5–1.57)	0.68
Chronic aspirin/NSAIDs use	-0.13 (-0.41 to 0.15)	0.88 (0.67–1.16)	0.36	Chronic aspirin/NSAIDs use	-0.11 (-0.38 to 0.16)	0.9 (0.68–1.17)	0.43

genetic CRC syndromes, family history of CRC, or IBD and those diagnosed before the age of 50, we were able to focus on sporadic CRC cases, a population that can benefit from better risk stratification. The median age of the study population (69 years) was in the upper age range recommended for screening colonoscopy. However, in contrast to previous models that evaluated only individuals that underwent screening colonoscopy, the current study evaluated all incidence CRC cases. As such, the median age in our study represented the actual median age for CRC diagnosis in the entire population.

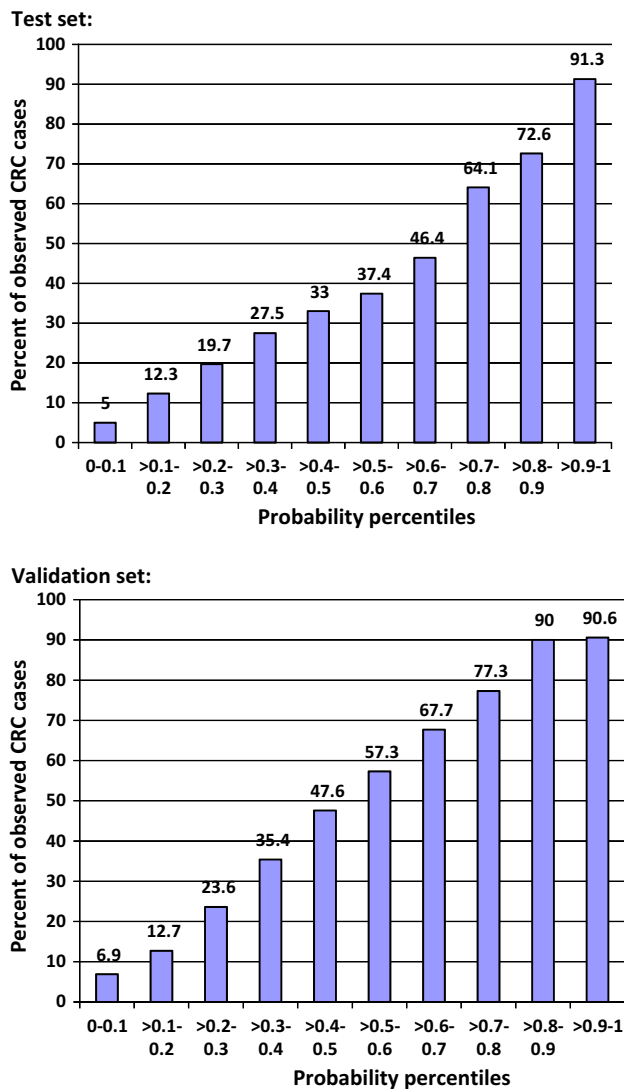
Laboratory parameters are good candidates for automatic EMR-based risk stratification, since after the age of 50 routine blood tests are recommended for other indications (such as lipid profile), at a minimum frequency of every 5 years. In contrast to other known risk factors, such as weight, smoking, or alcohol consumption, laboratory parameters are less prone to information bias, and in contrast to medications there is no possible bias due to lack of compliance. Moreover, the current study evaluated only commonly used laboratory parameters, with results available for at least one-third of the study cohort, and the final model included only variables that are part of the routine blood count and differential. Although changes in the CBC, mainly anemia, decrease in MCV, increase in red cell distribution width, and thrombocytosis were previously described in the literature as features of CRC [38–40], to our knowledge there were no previous studies to date that assessed the value of incorporating these changes in a CRC risk prediction model. Furthermore, most studies evaluated anemia as a dichotomous rather than as a continuous variable. A single study suggested a gradual decline in hemoglobin levels starting 3–4 years before cancer diagnosis [41].

All medical diagnoses, medication prescriptions, and laboratory results were recorded before cancer diagnosis. For laboratory results, we focused only on values from the year prior to diagnosis, in order to evaluate data that were collected in a uniform time window in both cases and controls. In addition, we focused on cases without known metastatic disease at diagnosis, ensuring that the risk factors that were used were actually relevant for early detection.

Despite the large number of variables that were tested, the current study had more than 100 cases per variable, and thus, there was no need to apply penalized regression methods to the analysis and the risk for over-fitting of the model was low. Additionally, we repeated the model with and without correction for linearity for laboratory results, age and height, with no change in results.

The current study had several limitations. The THIN database lacks information regarding some of the known risk factors for CRC such as dietary habits, physical activity, and





**Fig. 2** Observed CRC cases in the test and validation sets by probability percentiles of the combined model

**Table 4** Distribution of the combined model deciles in the test and validation sets

Deciles	Test set (11,935)		Validation set (4946)	
	N (%)	Cases (%)	N (%)	Cases (%)
0–0.1	2209 (18.5 %)	111 (5.0 %)	189 (3.8 %)	13 (6.9 %)
>0.1–0.2	2166 (18.1 %)	266 (12.3 %)	692 (14.0 %)	88 (12.7 %)
>0.2–0.3	1817 (15.2 %)	358 (19.7 %)	1464 (29.6 %)	345 (23.6 %)
>0.3–0.4	1440 (12.1 %)	396 (27.5 %)	1085 (21.9 %)	384 (35.4 %)
>0.4–0.5	1129 (9.5 %)	373 (33.0 %)	651 (13.2 %)	310 (47.6 %)
>0.5–0.6	883 (7.4 %)	331 (37.4 %)	520 (10.5 %)	298 (57.3 %)
>0.6–0.7	700 (5.9 %)	325 (46.4 %)	167 (3.4 %)	113 (67.7 %)
>0.7–0.8	554 (4.6 %)	355 (64.1 %)	75 (1.5 %)	58 (77.3 %)
>0.8–0.9	463 (3.9 %)	336 (72.6 %)	50 (1.0 %)	45 (90.0 %)
>0.9–1.0	574 (4.8 %)	524 (91.3 %)	53 (1.1 %)	48 (90.6 %)

race. However, previous models that used those factors had low discriminatory power with AUC <0.7 [21]. THIN also lacks information regarding tumor location as well as histopathology and staging. Since 20 % of CRC cases in the UK are metastatic at the time of diagnosis [30] and 75 % of the patients die within 2.5 years of diagnosis [2], by excluding all CRC cases who died within 2.5 years of diagnosis, <5 % of our CRC cases would have had metastatic disease. Although there is difference in risk factors and pathogenesis between right- and left-sided tumors and between colon and rectal cancers, those malignancies are diagnosed and treated similarly, thus favouring one model for different disease subgroups. THIN also lacks information regarding the pre-malignant condition, the adenomatous polyp. However, since we evaluated local disease as the study outcome, our model is relevant for detection of early-stage disease.

Despite the large sample size of our study, several laboratory parameters that were previously described as CRC risk factors, such as C-reactive protein and *Helicobacter pylori* infection positivity, were not available for most individuals and were excluded from the multivariate analysis. Performing multiple imputations on variables with large proportion of missing data (approximately 67 %) can be unreliable and introduces bias [42]. Although we were able to demonstrate the importance of laboratory results as predictors of sporadic CRC risk, we did not have sufficient number of individuals with repeated measures during follow-up in order to evaluate intra-individual changes in values. Of note, age despite its known effect on CRC risk was not included in the final model, probably secondary to our research methodology that matched cases and controls on duration of follow-up in order to ascertain equal “opportunity” to develop the disease. Despite this limitation, we were able to show better predictive power compared to previous models.

A possible selection bias could result from the exclusion of 6403 (28.6 %) of CRC cases due to suspected metastatic disease. However, since the aim of the current study was to predict early-stage disease that can lead to clinically meaningful interventions, analyzing individuals with advanced disease could bias such conclusions. An additional selection bias might stem from the fact that only one-third of the study population had values for all the laboratory results that were included in the model and the percent of cases with full laboratory results was higher compared to the percent of controls (40–50 vs. 25–30 %, respectively). Since the current study focused on the year before cancer diagnosis, it is possible that undiagnosed CRC might have influenced some of the laboratory results. However, it is important to note that our objective was not to identify etiological factors for CRC; our objective was to evaluate the power of a combination of biomarkers (causal or non-causal) and clinical factors in predicting early-stage CRC.

Finally, the current work created internal validation set by splitting the data randomly in a 2:1 ratio. We did not perform an external validation of our results. Since the THIN database is representative to the entire UK population, such an internal validation supports the generalizability of the results. No PPV or NPV were calculated for the risk models due to the use of case–control methodology. Category-based NRI was calculated for both test and validation sets and was presented separately for cases and controls as previously described [43, 44].

In summary, we developed and internally validated a CRC risk prediction model that demonstrates superior prediction performance compared to the existing models (with a NRI of more than 40 %). The improved performance resulted from incorporation of routinely available laboratory results that are not susceptible to information bias. Future works will need to perform external validation of the model in diverse populations, for example patients with prior colonoscopy, and evaluate the significance of intra-individual changes in laboratory values on sporadic cancer risk. Such a model can be used for a risk-tailored screening approach that can help determine which patients would benefit most from CRC screening by colonoscopy. Cost effectiveness analysis will be needed in each population to determine risk threshold for different levels of screening. This tool cannot be used to determine such threshold, but can be used to assign individual risk level in order to apply the risk-tailored screening approach once the thresholds are established. The suggested laboratory-based model might represent a shift in the paradigm by which we study CRC risk. Of note, this model does not suggest a need to change current screening guidelines or forgoing screening in individuals at low risk.

**Acknowledgments** Dr. Boursi would like to thank the Djerassi family for supporting his postdoctoral fellowship.

**Funding** The project was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR000003. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

**Compliance with ethical standards**

**Conflict of interest** None.

## References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. 2010;127:2893–2917.
2. American Cancer Society. *Colorectal Cancer Facts & Figures 2015*. Atlanta, GA: American Cancer Society; 2015.
3. Keighley MRB. Gastrointestinal cancers in Europe. *Aliment Pharmacol Ther*. 2003;18:7–30.
4. Burt RW. Colon cancer screening. *Gastroenterology*. 2000;119:837–853.
5. Cannon-Albright LA, Skolnick MH, Bishop DT, Lee RG, Burt RW. Common inheritance of susceptibility to colonic adenomatous polyps and associated colorectal cancer. *N Engl J Med*. 1988;319:533–537.
6. Levin B, Lieberman D, McFarland B, et al. Screening and surveillance for the early detection of CRC and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-society Task Force on CRC, and the American College of Radiology. *Gastroenterology*. 2008;134:1570–1595.
7. Boursi B, Halak A, Umansky M, Galzan L, Guzner-Gur H, Arber N. Colonoscopic screening of an average-risk population for colorectal neoplasia. *Endoscopy*. 2009;41:516–521.
8. Yuhara H, Steinmaus C, Cohen SE, Corley DA, Tei Y, Buffler PA. Is diabetes mellitus an independent risk factor for colon cancer and rectal cancer? *Am J Gastroenterol*. 2011;106:1911–1921.
9. Cook NR, Lee IM, Zhang SM, Moorthy MV, Buring JE. Alternate-day, low-dose aspirin and cancer risk: long-term observational follow-up of a randomized trial. *Ann Intern Med*. 2013;159:77–85.
10. Winawer SJ, Zauber AG, Ho MN, et al. The National Polyp Study Workgroup. Prevention of colorectal cancer by colonoscopic polypectomy. *N Engl J Med*. 1993;329:1977–1983.
11. Zauber AG, Winawer SJ, O'Brien MJ, et al. Colonoscopic polypectomy and long term prevention of colorectal-cancer deaths. *N Engl J Med*. 2012;366:687–696.
12. Laiyemo AO, Pinsky PF, Marcus PM, et al. Utilization and yield of surveillance colonoscopy in the continued follow-up study of the polyp prevention trial. *Clin Gastroenterol Hepatol*. 2009;7:562–567.
13. Sint NJ, de Jonge V, van Baalen O, et al. Optimal resource allocation in colonoscopy: timing of follow-up colonoscopies in relation to adenoma detection rates. *Endoscopy*. 2013;45:545–552.
14. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81:1879–1886.
15. Optenberg SA, Clark JY, Brawer MK, Thompson IM, Stein CR, Friedrichs P. Development of a decision-making tool to predict

- risk of prostate cancer: the cancer of the prostate risk index (CAPRI) test. *Urology*. 1997;50:665–672.
16. Colditz GA, Atwood KA, Emmons K, et al. Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. *Cancer Causes Control*. 2000;11:477–488.
  17. 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;120:257–263.
  18. Kaminski MF, Polkowski M, Kraszewska E, Rupinski M, Butruk E, Regula J. A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. *Gut*. 2014;63:1112–1119.
  19. Yeoh KG, Ho KY, Chiu HM, et al. The Asia-Pacific colorectal screening score: a validated tool that stratifies risk for colorectal advanced neoplasia in asymptomatic Asian subjects. *Gut*. 2011;60:1236–1241.
  20. Wong MC, Lam TY, Tsoi KK, et al. A validated tool to predict colorectal neoplasia and inform screening choice for asymptomatic subjects. *Gut*. 2014;63:1130–1136.
  21. Freedman AN, Slattery ML, Ballard-Barbash R, et al. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *J Clin Oncol*. 2009;27:686–693.
  22. Park Y, Freedman AN, Gail MH, et al. Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. *J Clin Oncol*. 2009;27:694–698.
  23. Lubin JH, Gail MH. Biased selection of controls for case-control analysis of cohort studies. *Biometrics*. 1984;40:63–75.
  24. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiol Drug Saf*. 2009;18:76–83.
  25. Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care*. 2004;12:171–177.
  26. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf*. 2007;16:393–401.
  27. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of the health improvement network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Inform Prim Care*. 2011;19:251–255.
  28. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in the health improvement network. *Pharmacoepidemiol Drug Saf*. 2009;18:730–736.
  29. Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the general practice research database. *Pharmacoepidemiol Drug Saf*. 2005;14:443–451.
  30. Maringe C, Walters S, Rachet B, et al. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000–2007. *Acta Oncol*. 2013;52:919–932.
  31. Mickey J, Greenland S. A study of the impact of confounder-selection criteria on effect estimation. *Am J Epidemiol*. 1989;129:125–137.
  32. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med*. 2007;26:5512–5528.
  33. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Stat Med*. 2013;32:2262–2277.
  34. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, ed. *Frontiers in Econometrics*. New York: Academic Press; 1974:105–142.
  35. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:157–172.
  36. Spell DW, Jones DV Jr, Harper WF, Bessman D. The value of a complete blood count in predicting cancer of the colon. *Cancer Detect Prev*. 2004;28:37–42.
  37. Hamilton W, Lancashire R, Sharp D, Peters TJ, Cheng KK, Marshall T. The importance of anemia in diagnosing colorectal cancer: a case-control study using electronic primary care records. *Br J Cancer*. 2008;98:323–327.
  38. Tchebiner JZ, Nutman A, Boursi B, et al. Diagnostic and prognostic value of thrombocytosis in admitted medical patients. *Am J Med Sci*. 2011;342:395–401.
  39. Goldshtein I, Neeman U, Chodick G, Shalev V. Variations in hemoglobin before colorectal cancer diagnosis. *Eur J Cancer Prev*. 2010;19:342–344.
  40. Ma GK, Ladabaum U. Personalizing colorectal cancer screening: a systematic review of models to predict risk of colorectal neoplasia. *Clin Gastroenterol Hepatol*. 2014;12:1624–1634.
  41. Levitzky BE, Brown CC, Heeren TC, Schroy PC 3rd. Performance of a risk index for advanced proximal colorectal neoplasia among a racially/ethnically diverse patient population (risk index for advanced proximal neoplasia). *Am J Gastroenterol*. 2011;106:1099–1106.
  42. Lee KJ, Carlin JB. Recovery of information from multiple imputation: a simulation study. *Emerg Themes Epidemiol*. 2012; 9:3.
  43. Pencina MJ, Steyerberg EW, D'Agostino RB. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30:11–21.
  44. Pankratz VS, Degnim AC, Frank RD, et al. Model for individualized prediction of breast cancer risk after a benign breast biopsy. *J Clin Oncol*. 2015;33:923–929.