ORIGINAL RESEARCH

ColoRectal Cancer Predicted Risk Online (CRC-PRO) Calculator Using Data from the Multi-Ethnic Cohort Study

Brian J. Wells, MD, PhD, Michael W. Kattan, PhD, Gregory S. Cooper, MD, Leila Jackson, PhD, MPH, and Siran Koroukian, PhD

Background: **Better risk predictions for colorectal cancer (CRC) could improve prevention strategies by allowing clinicians to more accurately identify high-risk individuals. The National Cancer Institute's CRC risk calculator was created by Freedman et al using case control data.**

Methods: **An online risk calculator was created using data from the Multi-Ethnic Cohort Study, which followed >180,000 patients for the development of CRC for up to 11.5 years through linkage with cancer registries. Forward stepwise regression tuned to the c statistic was used to select the most important variables for use in separate Cox survival models for men and women. Model accuracy was assessed using 10-fold cross-validation.**

Results: **Patients in the cohort experienced 2762 incident cases of CRC. The final model for men contained age, ethnicity, pack-years of smoking, alcoholic drinks per day, body mass index, years of education, regular use of aspirin, family history of colon cancer, regular use of multivitamins, ounces of red meat intake per day, history of diabetes, and hours of moderate physical activity per day. The final model for women included age, ethnicity, years of education, use of estrogen, history of diabetes, pack-years of smoking, family history of colon cancer, regular use of multivitamins, body mass index, regular use of nonsteroidal anti-inflammatory drugs, and alcoholic drinks per day. The calculator demonstrated good accuracy with a cross-validated c statistic of 0.681 in men and 0.679 in women, and it seems to be well calibrated graphically. An electronic version of the calculator is available at http://rcalc.ccf.org.**

Conclusion: **This calculator seems to be accurate, is user friendly, and has been internally validated in a diverse population. (J Am Board Fam Med 2014;27:42–55.)**

Keywords: **Colorectal Cancer, Medical Decision Making, Prevention and Control, Risk**

Current colorectal cancer (CRC) risk prediction strategies seem to be effective at reducing deaths from colorectal cancer,[1] but they have certain limitations, including cost of the procedures, risk of rare but serious complications such as bowel perforation, and poor compliance. National screening guidelines suggest considering earlier screening for some high-risk patients (eg, patients with a family history of CRC) but otherwise promote a one size fits all philosophy.[2] The dichotomization of patients into low- and high-risk categories based on a single variable (eg, age ≥50) by the guidelines is

This article was externally peer reviewed.

Submitted 18 January 2013; revised 5 July 2013; accepted 10 July 2013.

From the Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH (BJW, MWK); University Hospitals Case Medical Center, Cleveland OH (GSC); Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH (LJ, SK).

Funding: Funding for the Multi-Ethnic Cohort Study was provided by grants 5R37 CA054281 and 5P01 CA033619 from the National Cancer Institute. Dr. Koroukian was supported in part by the Clinical and Translational Science Collaborative of Cleveland, UL1TR000439 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Conflict of interest: none declared.

Corresponding author: Brian J. Wells, MD, PhD, Cleveland Clinic, 9500 Euclid Ave., Cleveland, OH 44195 (E-mail: wellsb@ccf.org).

unfortunate since this strategy results in a loss of information and does not take into account other factors associated with CRC risk. Additional established risk factors for CRC include family history of CRC, personal history of adenomatous colorectal polyps, black race, obesity, lack of physical exercise, and inflammatory bowel disease (IBD).[3] Other risk factors that are less established include history of diabetes, smoking, excessive alcohol intake, a diet high in red meat, and preference for eating well-done meat.[3–7] Several possible protective factors for CRC include nonsteroidal anti-inflammatory medications, multivitamins, and hormone replacement therapy.[3] Supplementation with vitamin E, calcium, vitamin D, or fiber was thought to decrease risk based on observational studies but was proved ineffective when tested in randomized controlled trials.[8–10] Previous research suggests substantial differences between men and women in terms of CRC risk. Women may be exposed to exogenous estrogen in the form of oral contraceptives or hormone replacement therapy. Elevated body mass index seems to confer different effects on CRC risk in men than in women.[11,12] In addition, at least one previous study found that cigarette smoking was an important risk factor for men but not for women.[13]

There are 2 major CRC risk calculators currently described in the literature: the Harvard Cancer Risk Index[14] and a model by Freedman et al[13] and validated by Park et al[15] for the National Cancer Institute (NCI). The Harvard model only predicts the risk of colon cancer, which limits its use clinically since the screening and diagnostic testing for rectal cancer is essentially the same as it is for colon cancer. In addition, both the Harvard and NCI models were created using data from case control studies, which are prone to significant bias. Case control studies cannot definitively determine temporality and they are particularly prone to recall bias when the outcome being studied is cancer.[16] This may make the calculators perform poorly when applied to healthy patients in the future.

Accurate risk prediction estimates might improve the efficiency of screening by targeting high-risk patients with earlier or more frequent screening and by either delaying or foregoing screening in very low-risk individuals. There are limited colonoscopy resources, and each additional procedure carries costs and the potential for complications.[17] In addition, there may be a role for chemopreventive medications in some patients. Nonsteroidal anti-inflammatory medications have been shown to prevent adenomatous polyps but in general have not been found to be cost-effective for CRC prevention in patients at average risk.[18,19] Improved chemoprevention with fewer side effects may be possible with the combination of sulindac and difluoromethylornithine,[20] which may be cost-effective if reserved for high-risk individuals.

The Multiethnic Cohort Study (MEC) provides a potential data source for creating a new CRC calculator. The MEC is a large, prospective survey of a diverse population of residents from California and Hawaii who are >45 years old. Patients were followed for up to 11.5 years until death, the development of CRC, or until December 31, 2004. Details of the survey have been described elsewhere.[21] The MEC is an excellent source of data because it contains detailed information about exposure and has a large proportion of racial/ethnic minorities. In addition, the ascertainment of CRC was performed by linking with cancer registries in California and Hawaii, states that have state-wide, population-based cancer registries. Mortality was ascertained with linkages to death certificate files in Hawaii and California. Additional state cancer registries and the National Death Index were used to obtain outcomes for patients known to have moved out of state. Patients were actively followed through an annual newsletter that helped to maintain current addresses, and they were periodically contacted directly for follow-up questionnaires. Loss to follow-up was <1%.

The Cancer Risk Prediction Models Workshop held in 2004 by the NCI in Washington, DC, recommended that risk prediction models incorporate more minority patients and encouraged new risk models to be simple and user-friendly.[22] The goal of this project was to create an easy to use CRC calculator using the ethnically diverse population of patients available in the MEC. The idea was not to create a model for patients to use in the hope of changing behavior (which would require assumptions regarding cause and effect), but rather to create a model that could quickly provide a clinician with an accurate estimate of the patient's risk. The hope was to perform a head-to-head comparison of this new calculator with the existing NCI calculator. There were, unfortunately, differences in variable definitions between the MEC and the NCI datasets that made it difficult to accurately calculate the NCI predicted risk in the MEC co-

Table 1. Differences between Variable Definitions in the Freedman (aka National Cancer Institute) Calculator and the Multiethnic Cohort Study

Variable	National Cancer Institute Definition	Multiethnic Cohort Study
Estrogen	Asks about estrogen use in the past 2 years. Also asks about menopausal status and allows different estrogen effects depending on menopausal status	Does not specifically ask about estrogen use in the past 2 years
Vegetable intake	"In the past 30 days, about how many servings per week of vegetables or leafy green salads did you eat?" (cups)	Vegetables quantified according to grams per day based on detailed food frequency questionnaire
Family history of colorectal cancer	Number of first-degree relatives with a history of cancer of the colon or rectum	Family history of colon cancer in first-degree relative (yes or no)
Activity	Defined according to the number of hours of activity in the past week, but also requires detailed information about how many months the participant was active in the past year	Only asked about current activity level
Nonsteroidal anti-inflammatory drugs and aspirin	Regular use defined as 3 times per week	Regular use was defined as twice per week

hort. Differences in variable definitions between the NCI and the MEC are depicted in Table 1.

Methods

The calculator resulting from this study was created using data from the MEC described above. Individuals with a history of CRC or adenomatous polyps were excluded from the analysis, resulting in a final sample size of 180,630 patients, of whom 2762 developed CRC. The following risk factors were measured at baseline in the MEC and were included as candidate variables in our final model: age (continuous); sex (male vs female); personal history of cancer (dichotomous); body mass index (continuous); regular aspirin use (currently, previously, or no); family history of colon cancer (dichotomous); estrogen use (currently, previously, or no); daily alcohol intake (continuous); regular multivitamin usage (currently, previously, or no); hours of moderate or strenuous activity per day (continuous); primary race/ethnicity (black, Hawaiian, Japanese, Latino, or white); diabetes (dichotomous); years of education (continuous); pack-years of smoking (continuous); regular use of nonsteroidal anti-inflammatory drugs (NSAIDS) (currently, previously, or no); intake (in ounces) of red meat per day (continuous); and preference for well-done meat (dichotomous). Continuous variables were modeled using restricted cubic splines, which use calculus to describe the nonlinear relationships between the predictor variables and the outcome.[23] Categorization of continuous variables is frequently done but results in a loss of information and decreased prediction accuracy.[24,25] Categorization is no longer necessary with computer-based calculator interfaces.

Detailed dietary questions were not included since randomized trials for fiber and vitamin supplementation proved ineffective at preventing adenomatous colorectal polyps and/or CRC. In addition, it was important to keep the calculator simple by avoiding detailed dietary questions. It is unfortunate that the initial MEC survey did not contain information regarding IBD disease, sigmoidoscopy, or colonoscopy; as a result, these risk factors could not be included.

Because of the potential differences in risk between men and women, we decided to create separate risk calculators for men and women to keep the models simple while allowing the greatest flexibility in the selection of variables without increasing the work required by the end user. The decision to create separate models for men and women also obviated the need to include other interactions in the model since there were no clear interactions described in the literature that did not involve sex. As recommended by Harrell et al,[26] *a posteriori* interactions were not explored because of the possibility of spurious findings and unnecessary model complexity.

Missing data were imputed using the "mice" package for R in which all the predictor variables were used in regression equations to impute the missing values without knowledge of the outcome.[27] Table 2 displays the missing data for each of the variables. Models for predicting the risk of

Table 2A. Descriptive Statistics for Men by Colorectal Cancer Outcome in the Multiethnic Cohort Study (n = 80,062)

	Developed Colorectal Cancer		
Characteristics	No (n = 78,576)	Yes (n = 1,486)	Patients with Missing Data
Mean age, years (SD)*	59.8 (8.9)	64.2 (7.8)	0 (0.0)
Race/ethnicity*			0 (0.0)
Black	11,202 (14.3)	231 (15.5)	
Hawaiian	5,565 (7.1)	105 (7.1)	
Japanese	22,247 (28.3)	558 (37.6)	
Latino	19,879 (25.3)	301 (20.3)	
White	19,683 (25)	291 (19.6)	
Mean pack-years smoking (SD)*	13.9 (16.6)	17 (18.4)	3,088 (3.9)
Alcohol, mean drinks/day (SD)*†	1.0 (2.3)	1.3 (2.6)	0 (0.0)
Mean years of education (SD)*	13.2 (3.4)	13 (3.1)	884 (1.1)
Family history of colon cancer*	6,030 (7.7)	156 (10.5)	11,015 (13.8)
Mean body mass index (SD)	26.6 (4.1)	26.6 (4.2)	753 (0.9)
Regular use of aspirin‡			7,439 (9.3)
No	46,001 (58.5)	901 (60.6)	
Yes, not currently	13,795 (17.6)	255 (17.2)	
Yes, currently	18,780 (23.9)	330 (22.2)	
Regular use of multivitamins*§	37,395 (47.6)	627 (42.2)	4,189 (5.2)
Diabetes*	9,749 (12.4)	214 (14.4)	0 (0.0)
Mean hours of moderate activity per day (SD)*	1.3 (1.5)	1.2 (1.4)	3,590 (4.5)
History of cancer	7,038 (9.0)	130 (8.7)	1 (<0.1)
Preference for well-done meat	34,122 (43.4)	625 (42.1)	2,327 (2.9)
Mean intake of red meat per day, oz (SD)*	2.6 (2.1)	2.5 (1.9)	0 (0.0)
Regular use of NSAIDs*†			10,214 (12.8)
No	57,852 (73.6)	1,146 (77.1)	
Yes, not currently	12,924 (16.4)	207 (13.9)	
Yes, currently	7,800 (9.9)	133 (9)	

Data are n (%) unless otherwise indicated. The Multiethnic Cohort study enrolled an ethnically diverse mix of residents from Hawaii and California between 1993 and 1996.

*Statistically significant in univariate analysis ($P < .05$).

†One alcoholic drink is defined as 1 oz of alcohol, which is approximately equivalent to one 12 oz beer, one 4-oz glass of wine, or 1 shot of liquor.

‡At least twice a week for 1 month or longer.

§At least once per week for the past year.

NSAIDS, nonsteroidal anti-inflammatory drugs; SD, standard deviation.

CRC in men and women were fit using Cox regression from the time of the initial survey until the development of CRC or death.

Variable selection was performed using a modified version of forward stepwise regression that was tuned to Harrell's c-statistic,[28] which will be referred to herein as the forward stepwise c-statistic. In a previous study involving 100 random cross-validations of 4 separate datasets, the forward stepwise c-statistic was the most likely to produce the most accurate model (highest c-statistic) when compared with traditional forward or backward stepwise regression and Harrell's model approximation.[29] The forward stepwise c-statistic was determined by simply calculating the apparent c-statistic for each possible 1-variable model, selecting the most accurate variable, and then sequentially adding additional variables to the model by repeating this process until the c-statistic no longer increased (or the full model is reached).

The regression equations used to create the models also were used to construct a free online version of the calculator using the Cleveland Clinic Risk Calculator Constructor (http://makercalc.ccf.org). The models were internally validated using 10-fold cross-validation to assess both discrimination (c-statistic) and calibration

Table 2B. Descriptive Statistics for Women by Colorectal Cancer Outcome (n = 100,568)

	Developed Colorectal Cancer		
Characteristics	No (n = 99,292)	Yes (n = 1,276)	Patients with Missing Data
Mean age, years (SD)*	59.5 (8.8)	64 (7.9)	0 (0.0)
Race/ethnicity*			0 (0.0)
Black	19,694 (19.8)	347 (27.2)	
Hawaiian	7,464 (7.5)	77 (6)	
Japanese	26,500 (26.7)	413 (32.4)	
Latino	21,730 (21.9)	194 (15.2)	
White	23,904 (24.1)	245 (19.2)	
Mean pack-years smoking (SD)*	6.6 (12.1)	7.4 (12.8)	3,739 (3.7)
Alcohol intake (mean drinks/day [SD])*†	0.3 (1.1)	0.4 (1.6)	0 (0.0)
Mean years of education (SD)	13.0 (3.3)	12.8 (2.9)	1,255 (1.2)
Family history of colon cancer*	9,165 (9.2)	174 (13.6)	12,707 (12.6)
Mean body mass index (SD)	26.4 (5.5)	26.6 (5.7)	2,346 (2.3)
Regular use of aspirin‡			4,971 (4.9)
No	61,593 (62)	806 (63.2)	
Yes, not currently	19,193 (19.3)	234 (18.3)	
Yes, currently	18,506 (18.6)	236 (18.5)	
Regular use of multivitamins*§	53,596 (54)	625 (49)	2,649 (2.6)
Diabetes*	10,910 (11)	188 (14.7)	0 (0.0)
Mean hours of moderate activity/day (SD)	1.1 (1.2)	1 (1.2)	2,420 (2.4)
History of cancer	11,576 (11.7)	151 (11.8)	0 (0.0)
Preference for well-done meat*	53,082 (53.5)	739 (57.9)	1,287 (1.3)
Mean intake of red meat per day, oz (SD)*	1.7 (1.6)	1.6 (1.4)	0 (0.0)
Regular use of NSAIDs*‡			6,488 (6.2)
No	62,132 (62.6)	849 (66.5)	
Yes, not currently	21,284 (21.4)	266 (20.8)	
Yes, currently	15,876 (16)	161 (12.6)	
Estrogen use*‖			3,277 (3.1)
No	53,754 (54.1)	709 (55.6)	
Yes, not currently	17,752 (17.9)	278 (21.8)	
Yes, currently	27,786 (28)	289 (22.6)	

Data are n (%) unless otherwise indicated. The Multiethnic Cohort study enrolled an ethnically diverse mix of residents from Hawaii and California between 1993 and 1996.
*Statistically significant in univariate analysis ($P < .05$).
†One alcoholic drink is defined as 1 oz of alcohol, which is approximately equivalent to one 12 oz beer, one 4 oz glass of wine, or one shot of liquor.
‡At least twice a week for 1 month or longer.
§At least once per week for the past year.
‖Estrogen use was defined as female hormones administered by pill, injection, or patch for menopause or other reasons.
NSAID, nonsteroidal anti-inflammatory drug; SD, standard deviation.

(assessed graphically). Statistical calculations were performed using R version 2.10 with the Design library (available at http://www.r-project.org).

Results

Participants in the MEC had a mean age of 59.7 years at baseline, and there were 2762 incident cases of CRC. Detailed descriptive statistics according to sex and CRC outcome for the variables included in the final models are shown in Table 2A (men) and B (women). In univariate analyses among both men and women, the following factors were significantly associated with risk of CRC: age, race, smoking, alcohol intake, family history of colon cancer, multivitamin use, diabetes, and regular use of NSAIDs. Years of education and hours of activity per day were associated with CRC in unadjusted analyses among men only, whereas intake (in ounces) of

Table 3A. Variable Impact on the Apparent C-Statistic for the Colorectal Cancer Model in Men

Variable	C-Statistic	Change in C-Statistic	Variables (n)
No Model	0.5	—	0
Age	0.663280	0.163280	1
Race/ethnicity	0.672989	0.009710	2
Pack-years of smoking	0.678214	0.005224	3
Alcoholic drinks per day	0.681442	0.003229	4
Body mass index	0.684263	0.002821	5
Years of education	0.686596	0.002333	6
Regular use of aspirin	0.688405	0.001809	7
Family history of colon cancer	0.689931	0.001526	8
Regular use of multivitamins	0.691143	0.001212	9
Red meat intake (oz) per day	0.691765	0.000622	10
History of diabetes	0.692365	0.000600	11
Moderate physical activity per day (hours)	0.692879	0.000513	12
History of cancer*	0.693176	0.000297	13
Regular use of NSAIDs*	0.693397	0.000221	14
Preference for well-done meat*	0.693538	0.000141	15 (Full)

The outcome was colorectal cancer.
*The variable was excluded from the final model because of the relatively small amount of additional prediction accuracy that it could have contributed to the model. The final model, which included 12 variables, had an apparent c-statistic of 0.6929, which was within 0.001 of the accuracy associated with the c-statistic of the full model, which was 0.6935.
NSAIDS, nonsteroidal anti-inflammatory drugs.

red meat per day, preference for well-done meat, and use of estrogen were associated with CRC in women only.

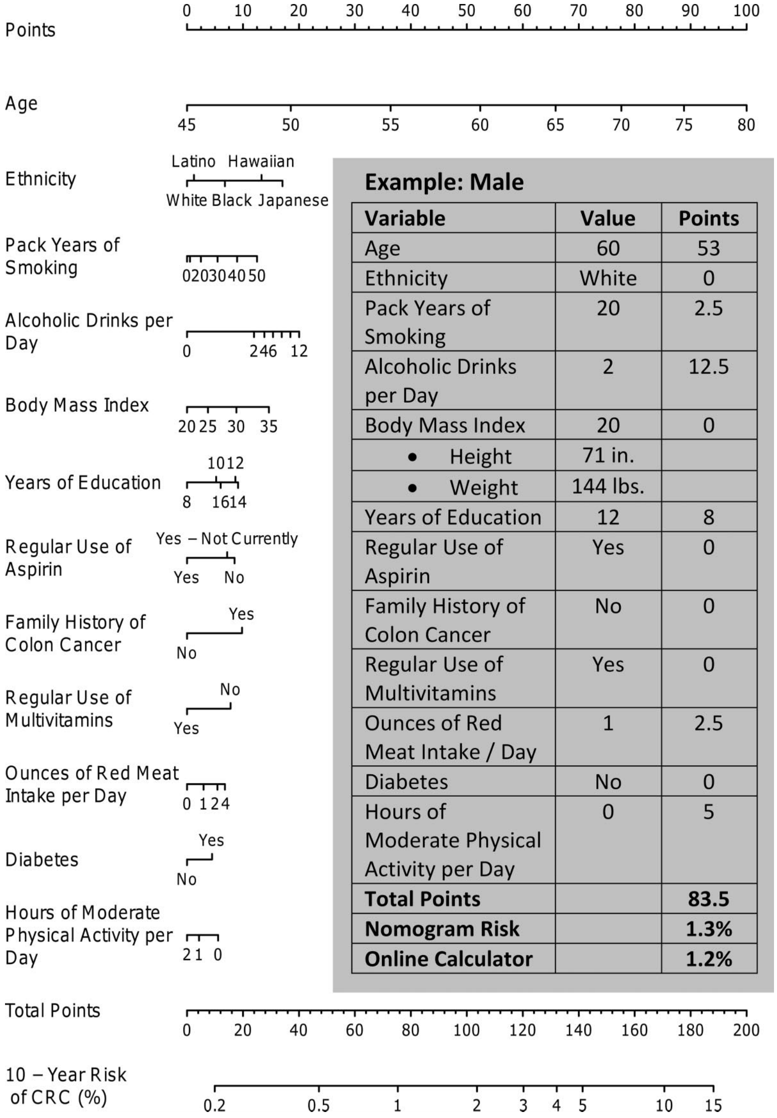
Table 3A and B shows the absolute contribution that each additional variable made on the c-statistic for men and women, respectively. Age had a sub-

Table 3B. Variable Impact on the Apparent C-Statistic for the Colorectal Cancer Model in Women

Variable	C-Statistic	Change in C-Statistic	Variables (n)
No Model	0.5	—	0
Age	0.657655	0.157655	1
Race/ethnicity	0.667666	0.010010	2
Years of education	0.670909	0.003243	3
Use of estrogen	0.674048	0.003139	4
History of diabetes	0.676648	0.002600	5
Pack-years of smoking	0.679123	0.002476	6
Family history of colon cancer	0.680963	0.001839	7
Regular use of multivitamins	0.682497	0.001534	8
Body mass index	0.683796	0.001300	9
Regular use of NSAIDs	0.684893	0.001097	10
Alcoholic drinks per day	0.685869	0.000976	11
Preference for well-done meat*	0.686268	0.000399	12
Moderate physical activity per day (hours)*	0.686525	0.000257	13
Regular use of aspirin*	0.686790	0.000266	14
Red meat intake per day (oz)*	0.686869	0.000078	15
History of cancer*	0.686865	-0.000003	16 (Full)

The outcome was colorectal cancer.
*The variable was excluded from the final model because of the relatively small amount of additional prediction accuracy that it could have contributed to the model. The final model, which included 11 variables, had an apparent c-statistic of 0.6859, which was within 0.001 of the accuracy associated with the c-statistic of the full model, which was 0.6869.
NSAIDS, nonsteroidal anti-inflammatory drugs.

Figure 1A. Nomogram for predicting colorectal cancer risk in men. Instructions: Draw a perpendicular line from the patient's age to the "points" axis and record the value. Repeat this process for the remaining variables and tally. The 10-year risk of colorectal cancer (CRC) is identified where a line drawn straight down from the "total points" axis intersects the "10-year risk of CRC (%)." Please note that the "years of education" variable has a *U*-shaped relationship with the 10-year risk of CRC. That is, the lowest risk of CRC occurs at 8 years and increases as you move along the top of the axis from left to right until reaching the highest risk at 14 years, and then it decreases along the bottom of the axis as you move to the left from 14 to 16 years.



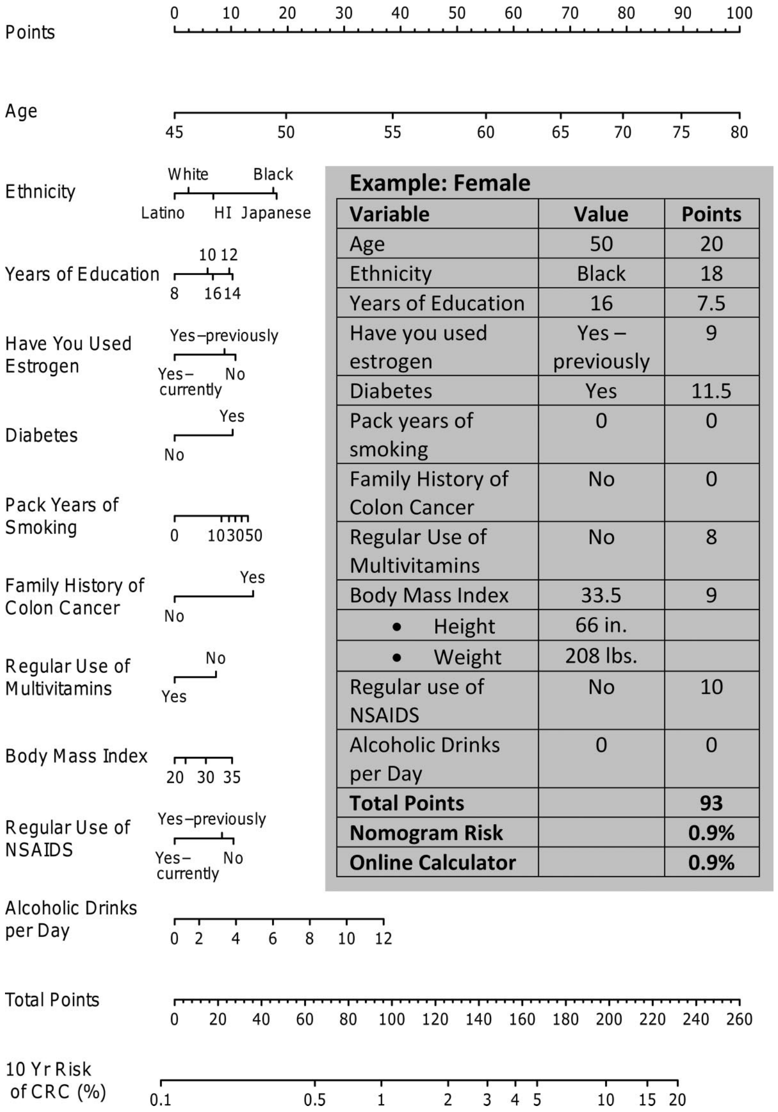
Example: Male

Variable	Value	Points
Age	60	53
Ethnicity	White	0
Pack Years of Smoking	20	2.5
Alcoholic Drinks per Day	2	12.5
Body Mass Index	20	0
• Height	71 in.	
• Weight	144 lbs.	
Years of Education	12	8
Regular Use of Aspirin	Yes	0
Family History of Colon Cancer	No	0
Regular Use of Multivitamins	Yes	0
Ounces of Red Meat Intake / Day	1	2.5
Diabetes	No	0
Hours of Moderate Physical Activity per Day	0	5
Total Points		**83.5**
Nomogram Risk		**1.3%**
Online Calculator		**1.2%**

stantially stronger effect on the accuracy of the model than any of the other variables for both men and women. Race/ethnicity was the second most important variable for both sexes, and the relative importance of the additional variables were different between the sexes. The regular use of NSAIDs was an important predictor for women but not for men, whereas a personal history of cancer was not an important predictor variable for either sex. The final models contained 12 variables and 11 variables for men and women, respectively. These sized models were chosen because the c-statistic was within 0.001 of the maximal c-statistic that could be achieved with a full model.

Figure 1A and B shows the paper-based nomograms for calculating risk of CRC in men and women, respectively. The nomograms provide a visual method for quickly identifying the relationship between predictor variables and outcome and provide a general idea of the relative impact that

Figure 1B. Nomogram for predicting colorectal cancer risk in women. NSAIDS, nonsteroidal anti-inflammatory drugs.



Example: Female

Variable	Value	Points
Age	50	20
Ethnicity	Black	18
Years of Education	16	7.5
Have you used estrogen	Yes – previously	9
Diabetes	Yes	11.5
Pack years of smoking	0	0
Family History of Colon Cancer	No	0
Regular Use of Multivitamins	No	8
Body Mass Index	33.5	9
• Height	66 in.	
• Weight	208 lbs.	
Regular use of NSAIDS	No	10
Alcoholic Drinks per Day	0	0
Total Points		**93**
Nomogram Risk		**0.9%**
Online Calculator		**0.9%**

different values of the variable of interest might have on risk. As you move to the right along any variable axis, the risk of CRC increases and the length of the axis shows the maximum contribution that any variable may have on overall risk. Age had the highest potential effect on risk of CRC for both sexes. The relationship between variables and risk of CRC seems to be consistent with the existing literature, except that the years of education variable has a j-shaped relationship for both men and women. Table 4 displays the multiple regression models using hazard ratios. Continuous variables were modeled using restricted cubic splines, and therefore a consistent hazard ratio for each incremental increase in a continuous variable cannot be displayed. Therefore, Table 4 shows the hazards associated with the 75th percentile versus the 25th percentile for these variables. Individuals at the 75th percentile for age (67 years in both men and women) were approximately 3 times more likely to be diagnosed with CRC than someone aged 52 years (point estimate of 3.03 in men and 2.81 in women). Coefficients for the final models are displayed in Table 5.

The median predicted 10-year risk of CRC was 1.0% in women and 1.6% in men. The percentage of men with a 10-year risk <1% was 29.5%, whereas >48% of women had a 10-year risk <1%. In contrast, there were 2600 men (3.2%) and 206 women (0.2%) who had a predicted 10-year risk >5%. A preponderance of men with a risk of >5% were Japanese (66.6%), whereas 74.8% of the women with a risk >5% were black.

Table 4A. Adjusted Multiple Regression Analysis for Colorectal Cancer in Men

		Adjusted Hazard Ratio	
Variable	Comparison Groups*	Point Estimate	95% CI
Age (years)	67 vs 52	3.03	2.69–3.41
Diabetes	Yes vs no	1.12	0.96–1.30
Regular multivitamin use	Yes vs no	0.83	0.74–0.92
Family history of colon cancer	Yes vs no	1.27	1.08–1.51
Education (years)	16 vs 12	0.94	0.85–1.03
Race/ethnicity	Black vs white	1.18	0.99–1.42
	Hawaiian vs white	1.39	1.10–1.75
	Japanese vs white	1.52	1.31–1.77
	Latino vs white	1.03	0.86–1.23
Body mass index (kg/m^2)	28.7 vs 23.8	1.12	1.04–1.21
Alcoholic drinks per day (n)	1.15 vs 0	1.26	1.13–1.41
Moderate activity per day (hours)	1.6 vs 0.4	0.92	0.83–1.02
Regular aspirin use	Not currently vs no	0.97	0.84–1.11
	Yes vs no	0.81	0.71–0.92
Pack-years smoking (n)	19.8 vs 0	1.06	0.93–1.21

*Continuous variables were modeled using restricted cubic splines with 3 knots. Comparison groups for the continuous variables are based on the 75th percentile versus the 25th percentile.
CI, confidence interval.

The 10-fold cross-validation revealed bias-adjusted c-statistics of 0.681 (95% confidence interval [CI], 0.669–0.694) and 0.679 (95% CI, 0.665–0.692) in men and women, respectively. Calibration curves, which suggest that the models are well calibrated, are shown in Figure 2A and B for men and women, respectively. An online version of the CRC-PRO calculator that provides calculations for men and women is available on the Cleveland Clinic calculator site (http://rcalc.ccf.org) under the heading "Colorectal Cancer."

Table 4B. Adjusted Multiple Regression Analysis for Colorectal Cancer in Women

		Adjusted Hazard Ratio	
Variable	Comparison Groups*	Point Estimate	95% CI
Age (years)	67 vs 52	2.81	2.48–3.18
Diabetes	Yes vs no	1.26	1.08–1.48
Regular use of multivitamins	Yes vs no	0.85	0.76–0.95
Family history of colon cancer	Yes vs no	1.37	1.17–1.61
Education (years)	14 vs 12	1.01	0.97–1.06
Race/ethnicity	Black vs white	1.41	1.18–1.67
	Hawaiian vs white	1.10	0.85–1.44
	Japanese vs white	1.43	1.20–1.70
	Latino vs white	0.95	0.77–1.17
Body mass index (kg/m^2)	29.3 vs 22.5	1.10	1.00–1.21
Alcoholic drinks per day (n)	0.1 vs 0	0.99	0.93–1.06
Regular use of NSAIDs	Not currently vs no	0.95	0.83–1.10
	Yes vs no	0.79	0.66–0.94
Pack-years of smoking (n)	8.9 vs 0	1.20	1.05–1.38
Estrogen use	Not currently vs no	0.96	0.83–1.10
	Yes vs no	0.78	0.68–0.90

*Continuous variables were modeled using restricted cubic splines with 3 knots. Comparison groups for the continuous variables are based on the 75th percentile versus the 25th percentile.
CI, confidence interval; NSAIDS, nonsteroidal anti-inflammatory drugs.

Table 5A. Coefficients for the Model Predicting Colorectal Cancer in Men

Variables	Coefficient	Standard Error
Age*		
Linear component	0.0917	0.0106
Nonlinear component	−0.0234	0.0106
Diabetes	0.1102	0.0758
Regular use of multivitamins	−0.1918	0.0532
Family history of colon cancer	0.2425	0.0850
Years of education*		
Linear component	0.0721	0.0181
Nonlinear component	−0.0734	0.0193
Race/ethnicity		
Hawaiian	0.1609	0.1208
Japanese	0.2535	0.0820
Latino	−0.1366	0.0925
White	−0.1673	0.0921
Body mass index*		
Linear component	0.0180	0.0162
Nonlinear component	0.0090	0.0194
Alcoholic drinks per day*		
Linear component	0.2838	0.0752
Nonlinear component	−1.7375	0.5356
Hours of moderate activity per day*		
Linear component	−0.0907	0.0702
Nonlinear component	0.1103	0.1496
Regular aspirin use		
Yes, not currently	−0.0323	0.0719
Yes	−0.2096	0.0655
Pack-years of smoking*		
Linear component	0.0002	0.0059
Nonlinear component	0.0177	0.0170

*Continuous variables were modeled using restricted cubic splines with 3 knots. Therefore, continuous variables contain one coefficient for the linear component and a nonlinear component.

Table 5B. Coefficients for the Model Predicting Colorectal Cancer in Women

Variable	Coefficient	Standard Error
Age*		
Linear component	0.0900	0.0112
Nonlinear component	−0.0276	0.0114
Diabetes	0.2333	0.0821
Regular use of multivitamins	−0.1665	0.0568
Family history of colon cancer	0.3159	0.0820
Years of education*		
Linear component	0.0744	0.0204
Nonlinear component	−0.0757	0.0212
Race/ethnicity		
Hawaiian	−0.2410	0.1284
Japanese	0.0140	0.0844
Latino	−0.3967	0.0977
White	−0.3406	0.0892
Body mass index		
Linear component	0.0075	0.0146
Nonlinear component	0.0121	0.0177
Alcoholic drinks per day		
Linear component	−0.0886	0.3046
Nonlinear component	0.4092	0.7886
Regular use of NSAIDs		
Yes, not currently	−0.0464	0.0731
Yes	−0.2370	0.0887
Pack-years of smoking†		
Linear component	0.0627	0.0498
Nonlinear component 1	−1.8515	2.1240
Nonlinear component 2	2.2999	2.6795
Estrogen use		
Yes, not currently	−0.0443	0.0719
Yes	−0.2450	0.0724

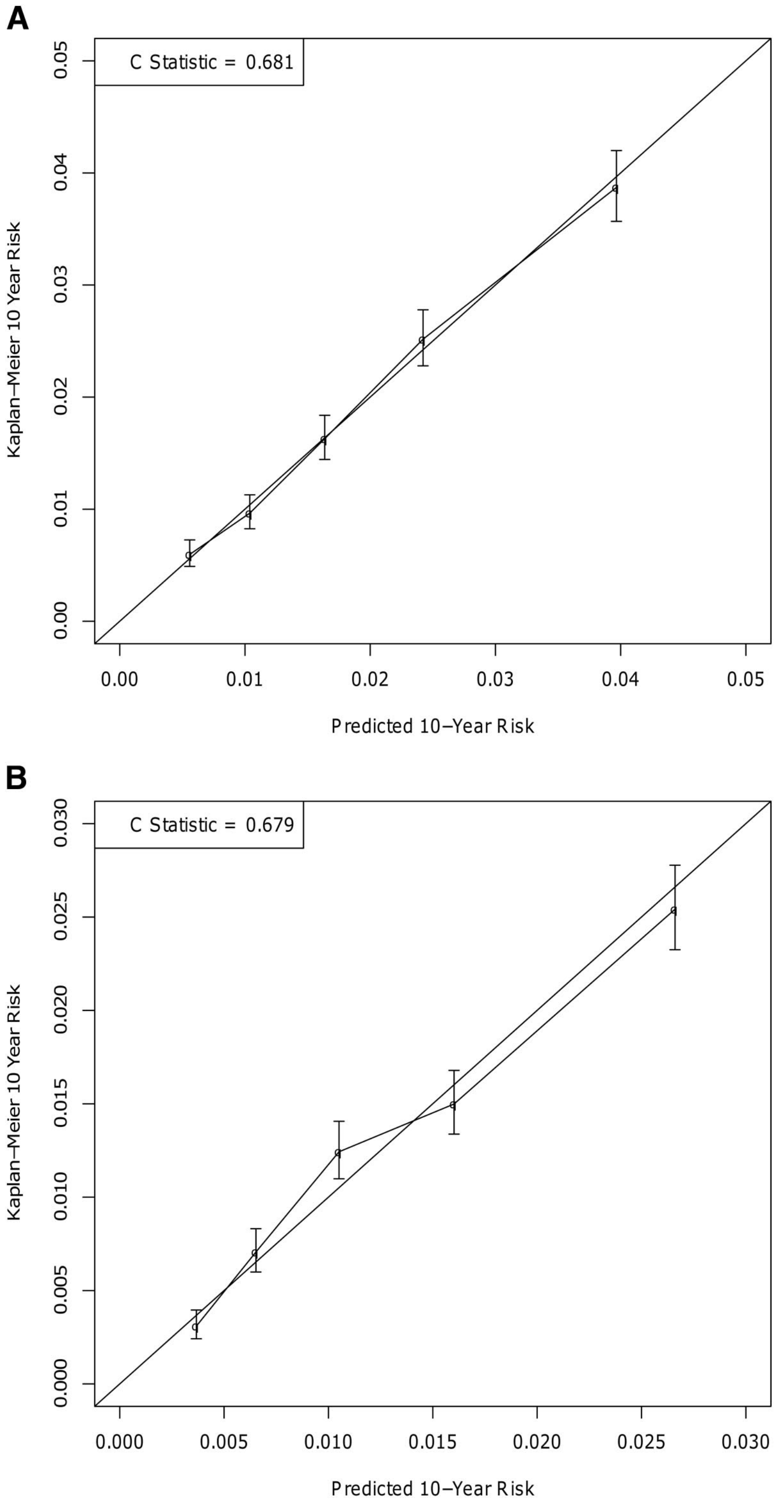
*Continuous variables were modeled using restricted cubic splines with three knots. Therefore, continuous variables contain one coefficient for the linear component (displayed first) and a non-linear component (displayed in the next row and denoted with a single quotation mark).

†Unable to obtain a natural spline using 3 knots and therefore restricted cubic splines with 4 knots were used for this variable. NSAIDS, nonsteroidal anti-inflammatory drugs.

Discussion

The user-friendly CRC risk calculator created in this project seems to have good prediction accuracy and to be well calibrated in the 10-fold cross-validation. There is always a trade-off between model complexity and prediction accuracy. The models in this study were limited to 11 or 12 variables because of the very meager gains in accuracy beyond this model size. The full models for both men and women had apparent c-statistics that were <0.001 better than the final chosen models, which is unlikely to provide any meaningful benefit in clinical practice. It was surprising to see the dominance of age, in terms of prediction accuracy, for the models of both men and women. A single-variable model containing only age had bias-corrected c-statistics of 0.657 (95% CI, 0.628–0.685) and 0.662 (95% CI, 0.636–0.688) in women and men, respectively. At first glance, the ability of age to predict CRC this well by itself may argue for the current practice of creating screening guidelines based largely on age alone. However, there may still be individual patients of the same age whose risk of CRC could

Figure 2. Calibration curve for the prediction in men (A) and women (B). The calibration curves were created by plotting the mean predicted risk of colorectal cancer in each quintile of risk on the *x*-axis against the corresponding Kaplan-Meier (K-M) estimated incidence in the same quintile. Error bars reflect the 95% confidence interval around the K-M estimate. The risks are displayed as probabilities. Perfect calibration would fall directly on the 45-degree line.



vary greatly. For example, a 50-year-old woman with the lowest risk possible according to the female nomogram in Figure 1B would have a 10-year risk of CRC of approximately 0.2%, whereas a 50-year-old at the highest risk, according to the extremes on the nomogram, would have a risk that is >10 times higher (ie, >2%).

It is hoped that future cost-effectiveness analyses and clinical guidelines involving CRC screening will focus on absolute risks based on sophisticated risk models rather than simple patient categories. The American Diabetes Association recently adopted this type of strategy with its recommendation that aspirin for the prevention of cardiovascular disease among patients with type 2 diabetes should be based on absolute risk of cardiovascular disease.[30] The current joint guidelines for CRC screening recommend screening for all average-risk individuals starting at age 50.[2] The 10-year probability of CRC for both men and women of all races at age 50 is approximately 0.646%.[31] A look at the MEC data shows that 81.9% of men but only 57.0% of women between 50 to 55 years of age without a history of intestinal polyps have a 10-year probability of CRC that is ≥0.646%. A cost-effectiveness analysis that uses individual patient risk probabilities could help to better delineate the exact risk threshold at which screening becomes cost-effective.

Some researchers have suggested that absolute cutoffs for c-statistics can reflect the quality of a model (eg, a c-statistic >0.7 is considered to be good). In reality, any model with a c-statistic >0.5 may have clinical utility. The current model is more accurate than predicting risk with age alone and, as shown above, could result in a significant change in practice. However, the relatively low c-statistic does seem to indicate that important risk factors are conspicuously missing.

The nomograms revealed relationships between predictor variables and CRC that are largely consistent with the existing literature. It is also somewhat artificial to evaluate a single variable in the nomogram. In the real world, a change to one independent variable (ie, a patient characteristic) is almost always associated with changes to other variables. The goal of this calculator was not to evaluate the relationship between individual variables and CRC risk with the aim of changing behaviors, but rather to develop the most accurate overall risk prediction possible for tailoring screening and prevention strategies.

Because of differences in the datasets and variable definitions, it was, unfortunately, not possible to compare the accuracy of the models in this study head-to-head with the NCI calculator. In theory, the predictions obtained from this calculator should be more accurate than the NCI model because of the large sample size, the prospective data analysis, and the avoidance of the categorization of continuous variables. We hope that future analyses can directly compare the calculators using another dataset.

As mentioned previously, the MEC does not contain information about IBD, which is a known CRC risk factor. The estimated prevalence of Crohn's disease and ulcerative colitis in North America ranges from 37 to 246 and 26 to 199 cases per 100,000, respectively.[32] In practice, patients with IBD should undergo more aggressive screening and prevention strategies, as reflected in a separate set of guidelines that have been established for these patients.[33] It would have been ideal to have excluded patients with IBD from the creation of the risk calculator. Ignorance of IBD status could have artificially reduced the accuracy of the calculated prediction, but it is unlikely that this had a tremendous effect since the expected number of patients with IBD in the MEC should only be a few hundred.

Also mentioned previously, the initial survey of the MEC did not ask about a history of large-bowel endoscopy. Although this question was asked of a significant portion of the original patients in a follow-up survey, we did not feel it was appropriate to include this variable since the information was not known at baseline.

The large proportion of patients in the MEC who had a predicted 10-year risk of CRC <1% suggests that it may become reasonable clinically to delay or forgo colonoscopy in some very low-risk patients. This may be especially true for women since 1475 of the women in the study with a predicted risk <1% were at least 65 years old. Some experts have already suggested that a once-per-lifetime colonoscopic screening may be cost-effective in low risk patients.[34]

In contrast, there were more than 2806 patients who had a 10 year risk ≥5%. In comparison, a meta-analysis of ulcerative colitis found a 2% cumulative incidence of CRC 10 years after diagnosis.[35] This raises the possibility that some of these high-risk patients might benefit from more aggressive preventive measures. NSAIDS have clearly been shown to reduce the risk of adenomatous polyps in randomized trials[36–38] and seem to reduce the risk of CRC in at least one post-trial analysis.[37] A couple of decision analyses have found that NSAIDS would save lives but would not be cost-effective in average-risk patients.[18,19,38] A combination regimen comprising a low-dose NSAID (sulindac) combined with a low-dose chemotherapeutic drug (difluoromethylornithine) has been successful at preventing recurrent adenomatous polyps with few side effects.[20,39–41] It is possible that this regimen could prove to be a cost-effective adjunct to current prevention strategies in some patients, especially given the extremely poor compliance with current CRC screening guidelines.

It is hoped that the calculator created in this study can help to improve the assessment of CRC risk in individual patients and encourage the use of absolute risk thresholds for decision making. Cost-effectiveness analyses may be needed to determine the exact effect that this improved prediction may have on clinical care.

The authors would like to acknowledge the assistance of Stephanie Kocian in editing the manuscript.

References

1. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149:659–69.
2. Levin B, Lieberman DA, McFarland B, et al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. CA Cancer J Clin 2008;58:130–60.
3. Ahnen DJ, Macrae FA. Colorectal cancer: Epidemiology, risk factors, and protective factors. In: Basow DS, ed. UpToDate. Version 41.0. Waltham, MA; UpToDate; 2013.
4. Sinha R, Chow WH, Kulldorff M, et al. Well-done, grilled red meat increases the risk of colorectal adenomas. Cancer Res 1999;59:4320–4.
5. Hu FB, Manson JE, Liu S, et al. Prospective study of adult onset diabetes mellitus (type 2) and risk of colorectal cancer in women. J Natl Cancer Inst 1999;91:542–7.
6. Larsson SC, Orsini N, Wolk A. Diabetes mellitus and risk of colorectal cancer: a meta-analysis. J Natl Cancer Inst 2005;97:1679–87.
7. Larsson SC, Giovannucci E, Wolk A. Diabetes and colorectal cancer incidence in the cohort of Swedish men. Diabetes Care 2005;28:1805–7.
8. Alberts DS, Martinez ME, Roe DJ, et al. Lack of effect of a high-fiber cereal supplement on the recurrence of colorectal adenomas. Phoenix Colon Cancer Prevention Physicians' Network. N Engl J Med 2000;342:1156–62.
9. Bjelakovic G, Nagorni A, Nikolova D, Simonetti RG, Bjelakovic M, Gluud C. Meta-analysis: antioxidant supplements for primary and secondary prevention of colorectal adenoma. Aliment Pharmacol Ther 2006;24:281–91.
10. Wactawski-Wende J, Kotchen JM, Anderson GL, et al. Calcium plus vitamin D supplementation and the risk of colorectal cancer. N Engl J Med 2006;354:684–96.
11. Giovannucci E. Metabolic syndrome, hyperinsulinemia, and colon cancer: a review. Am J Clin Nutr 2007;86:s836–42.
12. Pischon T, Lahmann PH, Boeing H, 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;98:920–31.
13. 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;27:686–93.
14. 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;11:477–88.
15. 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;27:694–8.
16. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol 1990;43:87–91.
17. Goodwin JS, Singh A, Reddy N, Riall TS, Kuo YF. Overuse of screening colonoscopy in the Medicare population. Arch Intern Med 2011;171:1335–43.
18. Ladabaum U, Chopra CL, Huang G, Scheiman JM, Chernew ME, Fendrick AM. Aspirin as an adjunct to screening for prevention of sporadic colorectal cancer. A cost-effectiveness analysis. Ann Intern Med 2001;135:769–81.
19. Ladabaum U, Scheiman JM, Fendrick AM. Potential effect of cyclooxygenase-2-specific inhibitors on the prevention of colorectal cancer: a cost-effectiveness analysis. Am J Med 2003;114:546–54.
20. Meyskens FL Jr, McLaren CE, Pelot D, et al. Difluoromethylornithine plus sulindac for the prevention of sporadic colorectal adenomas: a randomized placebo-controlled, double-blind trial. Cancer Prev Res (Phila Pa) 2008;1:32–8.
21. Kolonel LN, Henderson BE, Hankin JH, et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol 2000;151:346–57.
22. Freedman AN, Seminara D, Gail MH, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. J Natl Cancer Inst 2005;97:715–23.
23. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989;8:551–61.

24. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41.

25. Kattan MW, Zelefsky MJ, Kupelian PA, Scardino PT, Fuks Z, Leibel SA. Pretreatment nomogram for predicting the outcome of three-dimensional conformal radiotherapy in prostate cancer. J Clin Oncol 2000;18:3352–9.

26. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.

27. Van Burren S, Oudshoorn CGM. MICE: multivariate imputation by chained equations. R package version 2.3. Available from: http://www.stefvanbuuren.nl/mi/MICE.html. Accessed November 16, 2013.

28. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247:2543–6.

29. Wells BJ, Yu C, Koroukian S, Kattan MW. Comparison of variable selection methods for the generation of parsimonious prediction models for use in clinical practice. Presented at the 33rd Annual Meeting of the Society for Medical Decision Making. Available from: http://smdm.confex.com/smdm/2011ch/webprogram/Paper6541.html. Accessed December 10, 2012.

30. Statements P. Standards of medical care in diabetes—2012. Diabetes Care 2012;35(Suppl 1):S11–63.

31. Surveillance, Epidemiology, and End Results Program. Fast stats: an interactive tool for access to SEER cancer statistics. Washington, DC: National Cancer Institute. Available from: http://seer.cancer.gov/faststats. Accessed November 19, 2013.

32. Loftus EV Jr. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology 2004;126:1504–17.

33. Itzkowitz SH, Present DH; Crohn's and Colitis Foundation of America Colon Cancer in IBD Study Group. Consensus conference: colorectal cancer screening and surveillance in inflammatory bowel disease. Inflamm Bowel Dis 2005;11:314–21.

34. Ness RM, Holmes AM, Klein R, Dittus R. Cost-utility of one-time colonoscopic screening for colorectal cancer at various ages. Am J Gastroenterol 2000;95:1800–11.

35. Eaden JA, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut 2001;48:526–35.

36. Asano TK, McLeod RS. Non steroidal anti-inflammatory drugs (NSAID) and aspirin for preventing colorectal adenomas and carcinomas. Cochrane Database Syst Rev 2004;(2):CD004079.

37. Flossmann E, Rothwell PM; British Doctors Aspirin Trial and the UK-TIA Aspirin Trial. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. Lancet 2007;369:1603–13.

38. Rostom A, Dubé C, Lewin G, et al. Nonsteroidal anti-inflammatory drugs and cyclooxygenase-2 inhibitors for primary prevention of colorectal cancer: a systematic review prepared for the U.S. Preventive Services Task Force. Ann Intern Med 2007;146:376–89.

39. Loprinzi CL, Messing EM, O'Fallon JR, et al. Toxicity evaluation of difluoromethylornithine: doses for chemoprevention trials. Cancer Epidemiol Biomarkers Prev 1996;5:371–4.

40. Meyskens FL Jr, Emerson SS, Pelot D, et al. Dose de-escalation chemoprevention trial of alpha-difluoromethylornithine in patients with colon polyps. J Natl Cancer Inst 1994;86:1122–30.

41. Meyskens FL Jr, Gerner EW, Emerson S, et al. Effect of alpha-difluoromethylornithine on rectal mucosal levels of polyamines in a randomized, double-blinded trial for colon cancer prevention. J Natl Cancer Inst 1998;90:1212–8.