Julia Hippisley-Cox and Carol Coupland

Identifying patients with suspected colorectal cancer in primary care:

derivation and validation of an algorithm

Abstract

Background
Earlier diagnosis of colorectal cancer could help improve survival so better tools are needed to help this.

Aim
To derive and validate an algorithm to quantify the absolute risk of colorectal cancer in patients in primary care with and without symptoms.

Design and setting
Cohort study using data from 375 UK QResearch® general practices for development and 189 for validation.

Method
Included patients were aged 30–84 years, free at baseline from a diagnosis of colorectal cancer and without rectal bleeding, abdominal pain, appetite loss, or weight loss in the previous 12 months. The primary outcome was incident diagnosis of colorectal cancer recorded in the next 2 years. Risk factors examined were age, body mass index, smoking status, alcohol status, deprivation, diabetes, inflammatory bowel disease, family history of gastrointestinal cancer, gastrointestinal polyp, history of another cancer, rectal bleeding, abdominal pain, abdominal distension, appetite loss, weight loss, diarrhoea, constipation, change of bowel habit, tiredness, and anaemia. Cox proportional hazards models were used to develop separate risk equations in males and females. Measures of calibration and discrimination assessed performance in the validation cohort.

Results
There were 4798 incident cases of colorectal cancer from 4.1 million person-years in the derivation cohort. Independent predictors in males and females included family history of gastrointestinal cancer, anaemia, rectal bleeding, abdominal pain, appetite loss, and weight loss. Alcohol consumption and recent change in bowel habit were also predictors in males. On validation, the algorithms explained 65% of the variation in females and 67% in males. The receiver operating curve statistics were 0.89 (females) and 0.91 (males). The D statistic was 2.8 (females) and 2.9 (males). The 10% of patients with the highest predicted risks contained 71% of all colorectal cancers diagnosed over the next 2 years

Conclusion
The algorithm has good discrimination and calibration and could potentially be used to help identify those at highest risk of current colorectal cancer, to facilitate early referral and investigation.

Keywords
colorectal cancer; diagnosis; primary care; qresearch; risk prediction; symptoms.

INTRODUCTION

Colorectal cancer is the second most common cancer in Europe, as well as the second most common cause of cancer death.[1] In the UK, 36 000 patients get colorectal cancer every year and 16 500 die from it. The UK has one of the poorest survival rates for colorectal cancer in Europe,[2] which is thought to be partly related to late presentation, delays in diagnosis, and delays in treatment. The 5-year survival for early-stage colorectal cancer is greater than 90%, compared with 10% for widespread cancer at diagnosis.[3,4] Evidence suggests that increased awareness of symptoms and earlier diagnosis could help improve treatment options and improve 5-year survival.[5] The National Awareness and Early Diagnosis Initiative (NAEDI) in England aims to make the public more aware of the signs and symptoms of cancer, and encourage those with symptoms to seek advice earlier.[6] It has been estimated that such an approach might save 5000 lives, without any new medical advances.[7]

Symptoms such as rectal bleeding, weight loss, appetite loss, diarrhoea, constipation, or abdominal pain might herald an existing condition of colorectal cancer,[8,9] and some of these symptoms, but not all, are included within current National Institute for Health and Clinical Excellence (NICE) guidelines.[10] However, these symptoms are also common and not specific to colorectal cancer, making the identification of patients at risk of suspected cancer a diagnostic challenge. Current guidelines classify many as high risk, while failing to identify a significant number of patients with colorectal cancer who may have early and more curable cancers.[11,12]

Simple approaches based on single 'red flag' symptoms such as rectal bleeding are likely to miss 60% of current colorectal cancers.[13] A better approach to risk stratification of suspected colorectal cancers based on symptom complexes and other risk factors is needed, based on data from primary care, where most patients present.[12,14] A combined algorithm may help clinicians better assess and prioritise patients at high risk of having colorectal cancer for further investigation or referral, while avoiding unnecessary referral of those at low risk. While several studies have derived measures of the positive predictive value (PPV) of individual symptoms,[13] or pairs of symptoms,[8] there is currently no computer-based tool that combines established risk factors such as age, alcohol, and family history with current symptoms, to estimate an individual's absolute risk of colorectal cancer in a primary care setting. Such a tool, developed using data from primary care, could be used to assess and prioritise patients with suspected colorectal cancer in routine clinical practice in primary care.

It was decided to develop and validate a risk-prediction algorithm to estimate an individual's absolute risk of currently having colorectal cancer, incorporating both

J Hippisley-Cox, MD, FRCGP, MRCP, professor of clinical epidemiology and general practice; **C Coupland**, PhD, associate professor in medical statistics, Division of Primary Care, University of Nottingham.

Address for correspondence
Julia Hippisley-Cox, Division of Primary Care, 13th floor, Tower Building, University Park, Nottingham, NG2 7RD.

E-mail: julia.hippisley-cox@nottingham.ac.uk

Submitted: 21 June 2011; **Editor's response:** 28 July 2011; **final acceptance:** 4 August 2011.

This is the full-length article (published online 27 Dec 2011) of an abridged version published in print. Cite this article as: **Br J Gen Pract 2012; DOI: 10.3399/bjgp12X616346**

How this fits in

The UK has one of the poorest survival rates for colorectal cancer in Europe, which is thought to be partly related to late presentation, delays in diagnosis, and delays in treatment. Symptoms that might indicate colorectal cancer are common and not specific to colorectal cancer, making the identification of patients at risk of suspected cancer a diagnostic challenge. Simple approaches based on single 'red flag' symptoms such as rectal bleeding are likely to miss 60% of current colorectal cancers. This study has developed and validated an algorithm that can be used to identify symptomatic patients in primary care with an existing, but as yet undiagnosed, colorectal cancer. The algorithm is based on simple clinical variables that can be ascertained in clinical practice. The algorithm performed well in a separate validation sample, with good discrimination and calibration. It could identify 10% of the population in which over 70% of all new colorectal cancer cases arose over 2 years.

symptoms and other risk factors. The QResearch® primary care database was used to develop the risk prediction models since it contains robust data on many of the relevant exposures and outcomes. It is also representative of the population where such a model is likely to be used and has been used successfully to develop and validate a range of prognostic models for use in primary care.[15–18] Once validated, the prediction models could be made available on the internet for the general public and integrated into clinical computer systems to help systematically identify those at high risk and alert clinicians to those who might benefit most from further assessment or interventions.[15,17]

METHOD

Study design and data source

A prospective cohort study was carried out in a large population of primary care patients from an open cohort study using the QResearch database (version 30). All practices in England and Wales that had been using their EMIS (Egton Medical Information System) computer system for at least a year were included. Two-thirds of practices were randomly allocated to the derivation dataset and the remaining one-third to a validation dataset. An open cohort of patients aged 30–84 years was identified, drawn from patients registered with practices between 1 January 2000 and 30 September 2010. The following were excluded: patients without a postcode-related Townsend score, those with a history of colorectal cancer at baseline, and those with a recorded red flag symptom in the 12 months prior to the study entry date, that is, symptoms of rectal bleeding, loss of appetite, weight loss, or abdominal pain, which might indicate colorectal cancer.

Entry to the cohort was defined as the latest of the study start date (1 January 2000); 12 months after the patient registered with the practice; and for those patients with red flag symptoms, the date of first recorded onset within the study period.

Clinical outcome definition

The study outcome was colorectal cancer, which was defined as incident diagnosis of colorectal cancer during the 2 years after study entry, recorded either on the patient's GP record using the relevant UK diagnostic codes, or on their linked Office for National Statistics (ONS) cause-of-death record using the relevant International Classification of Diseases (ICD)-9 codes (153 or 154) or ICD-10 diagnostic codes (C18–C21). A 2-year period was used, since this represents the period of time during which existing cancers are likely to become clinically manifest.[13] Patients without the study outcome were were censored at the earliest of the date of death, date of leaving the practice study end date, or 2 years of follow-up.

Predictor variables

Established predictor variables were examined, focusing on those that are likely to be recorded in the patient's electronic record and that the patient is likely to know.[8,13] Red flag symptoms were also included, such as rectal bleeding, appetite loss, weight loss, and abdominal pain, and other symptoms that might herald a diagnosis of colorectal cancer such as constipation and diarrhoea. Separate analyses were carried out in males and females. The predictor variables were:

- currently consulting a GP with first onset of rectal bleeding (yes/no);
- currently consulting a GP with first onset of loss of appetite (yes/no);
- currently consulting a GP with first onset of weight-loss symptom (yes/no);
- currently consulting a GP with first onset of abdominal pain (yes/no);
- recently consulted a GP with first onset of any of:

- abdominal distension in past 12 months (yes/no);
- constipation in past 12 months (yes/no);
- diarrhoea in past 12 months (yes/no);
- change in bowel habit in past 12 months (yes/no);
- tiredness in past 12 months (yes/no);

- age at baseline (continuous, ranging from 30 to 84 years);
- body mass index (continuous);
- alcohol status (non-drinker; trivial [<1 unit/day]; light [1–2 units/day]; moderate/heavy [≥3 units/day]);
- smoking status (non-smoker; ex; light [1–9 cigarettes/day]; moderate [10–19 cigarettes/day]; heavy smoker [≥20 cigarettes/day];
- Townsend deprivation score, derived from patients' postcodes (continuous);
- family history of gastrointestinal cancer (yes/no);
- previous diagnosis of cancer apart from colorectal cancer;
- inflammatory bowel disease (Crohn's disease, ulcerative colitis, coeliac disease);
- previous history of gastrointestinal polyp;
- diabetes (type1/type2/no diabetes); and
- anaemia, defined as recorded haemoglobin <11 g/dl in the past 12 months (yes/no).

Derivation and validation of the models

The risk-prediction algorithm was developed and validated using established methods.[15–21] Multiple imputation was used to replace missing values for body mass index, and alcohol and smoking status, and these values were used in the main analyses.[22–25] Five imputations were carried out. Cox's proportional hazards models was used to estimate the coefficients for each risk factor for males and females separately, using robust variance estimates to allow for the clustering of patients within general practices. Rubin's rules were used to combine the results across the imputed datasets.[26] Fractional polynomials were used to model non-linear risk relationships with continuous variables.[27] A full model was fitted initially, and variables were retained if they had a hazard ratio of <0.80 or >1.20 (for binary variables) and were statistically significant at the 0.01 level. Interactions between predictor variables and age were examined and included in the final models if they were statistically significant at the 0.01 level.

The regression coefficients for each variable from the final model were used as weights, which were combined with the baseline survivor function evaluated at 2 years, to derive absolute risk equations.[28] The baseline survivor function was estimated, based on zero values of centred continuous variables, with all binary predictor values set to zero, using the methods implemented in STATA.

The risk equations for males and females obtained from the derivation cohort were then applied to the validation cohort and measures of discrimination were calculated. R^2 (a measure of variation explained in the time to diagnosis of colorectal cancer),[29] the D statistic (a measure of discrimination where higher values indicate better discrimination),[30] and the area under the receiver operating characteristic (ROC) curve at 2 years were calculated. Calibration was assessed by comparing the mean predicted risks at 2 years with the observed risk by tenth of predicted risk. The observed risk was obtained using the Kaplan–Meier estimate evaluated at 2 years.

The validation cohort was used to define the thresholds for the 1%, 5%, and 10% of patients at highest estimated risk of colorectal cancer at 2 years. Sensitivity, specificity, and positive and negative predictive values were calculated using these thresholds, restricting the analyses to patients who had the outcome within 2 years or had at least 2 years of follow-up.

All the available data on the database were used to maximise the power and also the generalisability of the results. STATA (version 11) was used for all analyses.

RESULTS

Overall study population

Overall, 564 QResearch practices in England and Wales met the study inclusion criteria, of which 375 were randomly assigned to the derivation dataset, with the remainder assigned to a validation cohort. A total of 2 538 615 patients aged 30–84 years were identified in the derivation cohort, and 124 458 patients (4.9%) without a recorded Townsend deprivation score were excluded; 5506 (0.2%) patients with a history of colorectal cancer were also excluded, as well as a further 57 599 patients (2.3%) with at least one red flag symptom recorded in the 12 months prior to entry to the study at baseline, leaving 2 351 052 patients for analysis

A total of 1 342 329 patients aged

30–84 years were identified in the validation cohort; of these, 70 847 patients (5.3%) without a recorded Townsend score were excluded, as well as 2908 (0.2%) with a history of colorectal cancer, and 31 973 (0.2%) with at least one red flag symptom recorded in the 12 months prior to study entry, leaving 1 236 601 patients for analysis.

The baseline characteristics of each cohort were very similar, as shown in Table 1. As in previous studies,[15,17,31] the patterns of missing data supported the use of multiple imputation to replace missing values for alcohol and smoking status and body mass index (not shown, available from the authors).

Table 1. Baseline characteristics of patients in the derivation and validation cohorts; patients are free of a diagnosis of colorectal cancer at baseline. Figures are *n* (%) unless otherwise specified

Characteristic	Derivation cohort (*n* = 2 351 052)	Validation cohort (*n* = 1 236 601)
Female	1 172 670 (49.9)	616 361 (49.8)
Male	1 178 382 (50.1)	620 240 (50.2)
Mean age (SD), years	50.1 (15.0)	50.1 (14.9)
Mean Townsend score (SD)[a]	−0.3 (3.4)	−0.2 (3.6)
BMI recorded prior to study entry	1 865 822 (79.4)	1 003 783 (81.2)
Mean BMI (SD), kg/m^2	26.4 (4.6)	26.4 (4.7)
Alcohol status, *n* (%)		
None	510 179 (21.7)	275 152 (22.3)
Trivial (<1 unit/day)	656 450 (27.9)	355 654 (28.8)
Light (1–2 units/day)	492 318 (20.9)	257 381 (20.8)
Moderate or heavy (≥3 units/day)	175 953 (7.5)	93 075 (7.5)
Alcohol status not recorded	516 152 (22.0)	255 339 (20.6)
Smoking status		
Non-smoker	1 192 200 (50.7)	623 599 (50.4)
Ex-smoker	425 933 (18.1)	228 748 (18.5)
Current smoker, amount not recorded	71 363 (3.0)	39 196 (3.2)
Light smoker (<10/day)	147 852 (6.3)	79 729 (6.4)
Moderate smoker (10–19/day)	179 727 (7.6)	95 657 (7.7)
Heavy smoker (≥20/day)	133 865 (5.7)	73 501 (5.9)
Smoking status not recorded	200 112 (8.5)	96 171 (7.8)
Medical history		
Family history of gastrointestinal cancer	29 483 (1.3)	17 672 (1.4)
Prior cancer apart from colorectal cancer	49 331 (2.1)	26 169 (2.1)
Coeliac disease	3682 (0.2)	1869 (0.2)
Ulcerative colitis	9678 (0.4)	5183 (0.4)
Crohn's disease	4978 (0.2)	2591 (0.2)
Type 1 diabetes	7230 (0.3)	3966 (0.3)
Type 2 diabetes	79 010 (3.4)	42 032 (3.4)
Prior colorectal polyp	2504 (0.1)	1247 (0.1)
Haemoglobin <11 g/dl in last year	31 330 (1.3)	16 985 (1.4)
Haemoglobin recorded in last year	405 071 (17.2)	218 098 (17.6)
Symptoms		
Current rectal bleeding	52 453 (2.2)	29 118 (2.4)
Current abdominal pain	224 880 (9.6)	125 816 (10.2)
Current appetite loss	9959 (0.4)	5358 (0.4)
Current weight loss	25 113 (1.1)	14 065 (1.1)
Recent abdominal distension	1500 (0.1)	812 (0.1)
Recent change in bowel habit	3153 (0.1)	1821 (0.1)
Recent diarrhoea	22 451 (1.0)	12 288 (1.0)
Recent constipation	15 072 (0.6)	8458 (0.7)
Recent tiredness	22 521 (1.0)	12 620 (1.0)

[a]Townsend score is a deprivation score derived from patients' postcodes, which ranges between −6 (most affluent) and +11 (most deprived). BMI = body mass index. SD = standard deviation.

Incidence rates for red flag symptoms

Overall, in the derivation cohort, 52 453 patients were identified with incident rectal bleeding, 9959 with appetite loss, 25 113 with weight loss, and 224 880 with abdominal pain. Table 2 shows the age–sex incidence rates of each symptom. Apart from abdominal pain, the incidence rates were similar in males and females and increased with age. Abdominal pain was more common in females and tended to decrease with age in females and increase with age in males. The incidence of rectal bleeding was very similar to published rates from similar populations.[13]

Incidence rates of colorectal cancer

Overall in the derivation cohort, during the 2-year follow-up, a total of 4798 incident cases of colorectal cancer, arising from 4 110 382 person-years of observation were identified, giving a crude rate of 117 per 100 000 person-years. The incidence rate of colorectal cancer was higher among males than females and rose steeply with age. Of the 4798 incident cases, 2908 (60.6%) were colon cancer and 1890 (39.4%) rectal cancer. There were 4297 cases (89.6%) identified using the GP record and an additional 501 (10.4%) identified solely from the linked death record.

In the validation cohort, 2603 incident cases of colorectal cancer arising from 2 163 167 person-years of observation were identified, giving a crude rate of 120 per 100 000 person-years. Of these, 1562 (60.0%) were colon cancer and 1041 (40.1%) were rectal cancer. There were 2326 cases (89.4%) identified using the GP record, and an additional 277 (10.6%) solely from the linked death record.

Predictor variables

Table 3 shows the predictor variables selected for the final model for females and males. The predictors for both males and females included: age, family history of gastrointestinal cancer, anaemia, rectal bleeding, abdominal pain, appetite loss, and weight loss. Alcohol status and recent change in bowel habit were significant predictors in males but not in females.

The risk of colorectal cancer was elevated among males with a family history of gastrointestinal cancer (1.5-fold higher risk), anaemia (3.3-fold higher risk), rectal bleeding (27-fold higher at the mean age in males), abdominal pain (6.8-fold higher at

Table 2. Incidence rates of appetite loss, weight loss, rectal bleeding, and abdominal pain per 100 000 person-years in the derivation cohort and for age

Symptom and age range, years	Incidence rate (95% CI)	
	Females	Males
Appetite loss		
<35	66.5 (59.1 to 74.8)	33.8 (28.7 to 39.7)
35–44	73.8 (70 to 77.9)	42.1 (39.3 to 45.1)
45–54	61.4 (57.8 to 65.3)	39.9 (37 to 43)
55–64	58.3 (54.6 to 62.3)	46.8 (43.5 to 50.4)
65–74	97.8 (92.2 to 104)	88.3 (82.8 to 94.2)
75–84	275 (265 to 285)	224 (214 to 235)
Weight loss		
<35	113 (103 to 124)	58.8 (52 to 66.5)
35–44	131 (126 to 137)	74 (70.2 to 77.9)
45–54	146 (140 to 151)	104 (99.6 to 109)
55–64	183 (176 to 190)	168 (162 to 175)
65–74	294 (284 to 304)	281 (271 to 292)
75–84	594 (579 to 609)	593 (575 to 611)
Rectal bleeding		
<35	284 (269 to 301)	283 (268 to 300)
35–44	314 (306 to 322)	328 (320 to 337)
45–54	378 (368 to 387)	374 (365 to 383)
55–64	466 (455 to 477)	440 (429 to 451)
65–74	510 (496 to 523)	543 (528 to 557)
75–84	605 (590 to 620)	601 (583 to 619)
Abdominal pain		
<35	2691 (2639 to 2744)	1096 (1064 to 1128)
35–44	2445 (2420 to 2470)	1159 (1143 to 1175)
45–54	2187 (2163 to 2212)	1261 (1243 to 1278)
55–64	2103 (2078 to 2129)	1422 (1402 to 1442)
65–74	2133 (2104 to 2163)	1744 (1718 to 1772)
75–84	2061 (2031 to 2091)	1936 (1902 to 1971)

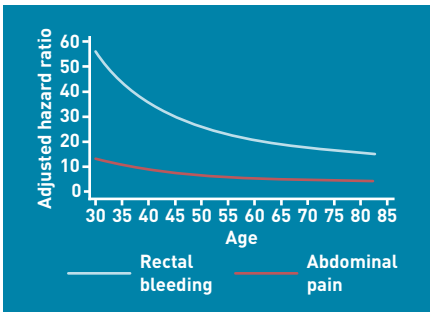
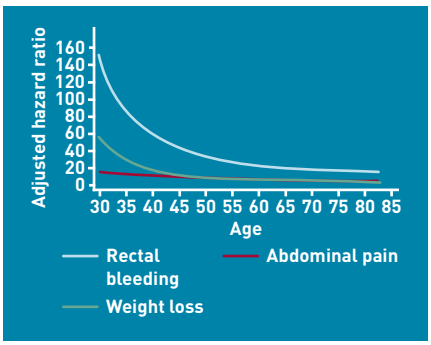


Figure 1. Adjusted hazard ratios for symptoms in males by age.

Figure 2. Adjusted hazard ratios for symptoms in females by age.



the mean age), appetite loss (2.2-fold higher), weight loss (4.1-fold higher), or change in bowel habit (2.3-fold higher). There were significant interactions between age and two variables (rectal bleeding and abdominal pain) for males, as shown in Figure 1. The graph indicates that the hazard ratios for both risk factors were higher among younger patients. The risk among males recorded as drinking 1–2 units of alcohol/day was 1.2-fold higher than in non-drinkers and the risk among those recorded as drinking ≥3 units/day was 1.4-fold higher. The other variables examined were not independent risk factors in males, so were not included in the final model.

The magnitudes of the hazard ratios for females were generally similar to those found for males, as shown in Table 3. As shown in Figure 2, there were significant interactions between age and three variables in females (rectal bleeding, abdominal pain, and weight loss). The risk associated with these three symptoms tended to be proportionately greater among younger females than in older females.

Validation

The validation statistics in Table 4 showed that the algorithms explained 65% of the variation in time to diagnosis in females and 67% of the variation in males. The D statistic was 2.8 for females and 2.9 for males. The ROC statistics were 0.89 for females and 0.91 for males.

Figure 3 shows the mean predicted scores and the observed risks at 2 years within each tenth of predicted risk, in order to assess the calibration of the model in the validation cohort. Overall, the model was well calibrated. There was close correspondence between predicted and observed 2-year risks within each model tenth for males and females, with a small degree of over-prediction in the highest tenth.

Individual risk assessment and thresholds

One potential use for this algorithm is within consultations with individual patients, particularly if they present with new onset of an alarm symptom such as rectal bleeding, abdominal pain, loss of weight, or loss of appetite. Some clinical examples are shown in Box 1. The algorithm could also be used for systematic risk stratification for a population of patients aged 30–84 years. Software implementing the algorithm could calculate the risk of a patient having an existing, but as yet undiagnosed, colorectal cancer, based on information already recorded in the patient's electronic health record. Patients at highest risk could be identified for further clinical assessment or investigation, such as colonoscopy or barium enema.

The 90th centile defined a high-risk group with a 2-year risk score of >0.5 % (Table 5). There were 1838 new cases of colorectal cancer within this group, out of 2603 new cases identified in the validation cohort, which accounted for 71% of all new cases of colorectal cancer (sensitivity). The PPV with this threshold was 2.1%. Alternatively, using a threshold based on the top 1% of risk (that is, a risk score >5.3%) had a sensitivity of 24.6% and a PPV of 8.1%. In contrast, the PPV of rectal bleeding alone was 3.9% and the sensitivity was 32.3%. In other words, an approach based only on a single symptom of rectal bleeding is likely to miss 70% of cases of colorectal cancer.

DISCUSSION

Summary

This research has developed and validated an algorithm designed to quantify the absolute risk of having existing, but as yet undiagnosed, colorectal cancer, based on a

Table 3. Adjusted hazard ratios (95% CI) for the final model for colorectal cancer for males and females in the derivation cohort

	Adjusted[a] hazard ratios for females (95% CI)	Adjusted[a] hazard ratios for males (95% CI)
Alcohol status		
Non-drinker	NS	1
Trivial drinker	NS	1.07 (0.949 to 1.20)
Light drinker	NS	1.20 (1.06 to 1.35)
Moderate/heavy drinker	NS	1.43 (1.25 to 1.63)
History and investigations		
Family history of gastrointestinal cancer[b]	1.39 (1.02 to 1.89)	1.52 (1.12 to 2.07)
Haemoglobin <11 g/dl in last year[b]	3.26 (2.84 to 3.74)	3.33 (2.86 to 3.87)
Current symptoms		
Current rectal bleeding[b]	32.3 (27.7 to 37.6)[c]	27.0 (23.5 to 31.1)[c]
Current abdominal pain[b]	6.90 (5.91 to 8.06)[c]	6.78 (5.76 to 7.97)[c]
Current appetite loss[b]	2.43 (1.70 to 3.47)	2.15 (1.53 to 3.03)
Current weight loss[b]	7.70 (5.32 to 11.1)[c]	4.07 (3.42 to 4.85)
Change in bowel habit in previous year[b]	NS	2.25 (1.47 to 3.46)

[a]Hazard ratios adjusted for all other terms in the table and for age. [b]Compared with a person without this characteristic. [c]Interaction term, at mean age. The models also included fractional polynomial terms for age, which were age^{-1} for females and age$^{-0.5}$ for males. The model for females also included interactions for the age term with rectal bleeding, abdominal pain, and weight loss. The model for males also included interaction for the age term with rectal bleeding and abdominal pain. NS = not significant.

Table 4. Validation statistics for the risk-prediction algorithm in the validation cohort

Statistic	Mean (95% CI)
Females	
R^2 statistic[a] (%)	64.8 (63.2 to 66.3)
D statistic[b]	2.78 (2.68 to 2.87)
ROC statistic[c]	0.89 (0.88 to 0.90)
Males	
R^2 statistic[a] (%)	66.7 (65.3 to 68.0)
D statistic[b]	2.90 (2.81 to 2.98)
ROC statistic[c]	0.906 (0.899 to 0.913)

[a]R^2 statistic shows explained variation in time to diagnosis of colorectal cancer — higher values indicate more variation is explained. [b]D statistic is a measure of discrimination — higher values indicate better discrimination. [c]ROC statistic is a measure of discrimination — higher values indicate better discrimination.

combination of symptoms and patient characteristics that the patient is likely to know or that can be easily ascertained in a primary care setting. The algorithm is based on seven simple clinical variables in females: age; family history of gastrointestinal cancer; anaemia; rectal bleeding; abdominal pain; appetite loss; and weight loss. The algorithm for males is similar but also includes alcohol use and change in bowel habit. The algorithm performed well in a separate validation sample, with good discrimination and calibration. This algorithm has been developed using data recorded within consultations in primary care and is likely to perform best when used in this setting. While it could be used outside the consultation setting (for example, by patients using a web calculator), caution is needed in the interpretation of the results regarding symptoms that have not been severe enough to prompt the patient to consult their GP.

Strengths and limitations

Key strengths of this study include its large size and the duration of follow-up. The study has a representative population, which increases the generalisability of the results. The study design has minimised the risk of a number of potential biases that could have otherwise affected the results. For example, it used prospectively recorded data from an existing database, which is not subject to selection, recall, and responder bias, since all eligible patients were included and symptoms and other risk factors were recorded before the diagnosis of colorectal cancer.

UK general practices have good levels of accuracy and completeness in recording clinical diagnoses and prescribed medications.[32] The authors consider this study has good face validity, since it has been conducted in the setting where the majority of patients in the UK are assessed, treated, and followed-up. The study has developed algorithms in patients from one group of practices and validated them in patients from separate practices who are representative of the patients likely to be considered for referral and treatment. This means it has been possible to demonstrate that the algorithm is likely to work outside the setting where it was developed.

Limitations include the lack of formally adjudicated outcomes, since the diagnoses were based on information recorded in the electronic record rather than collected as part of a clinical trial. Nonetheless, previous studies have validated similar outcomes and exposures using questionnaire data, and found levels of completeness and accuracy in similar GP databases to be good.[33,34]

Table 5. Comparison of strategies to identify patients at risk of having a diagnosis of colorectal cancer in the next 2 years, based on the validation cohort

Criteria	Risk threshold %	True negative[a]	False negative[b]	False positive[c]	True positive[d]	Sensitivity, %	Specificity, %	Positive predictive value (%)	Negative predictive value (%)
Rectal bleeding alone	n/a	1 204 833	1762	28 111	841	32.3	97.7	2.9	99.9
Abdominal pain alone	n/a	1 108 552	1758	124 392	845	32.5	89.9	0.7	99.8
Appetite loss alone	n/a	1 227 674	2557	5270	46	1.8	99.6	0.9	99.8
Weight loss alone	n/a	1 219 043	2497	13 901	106	4.1	98.9	0.8	99.8
Change in bowel habit alone[a]	n/a	617 486	1402	742	21	1.5	99.9	2.8	99.8
Anaemia alone	n/a	1 216 368	2356	16 576	247	9.5	98.7	1.5	99.8
Top 10% risk score	0.5	1 111 261	765	121 683	1838	70.6	90.1	1.5	99.9
Top 5% risk score	1.2	1 172 641	1134	60 303	1469	56.4	95.1	2.4	99.9
Top 1% risk score	5.2	1 221 229	1963	11 715	640	24.6	99.0	5.2	99.8

n/a = not applicable. [a]Criterion not met does not have disease. [b]Criterion not met does have disease. [c]Criterion met does not have disease. [d]Criterion met does have disease. [e]Males only.

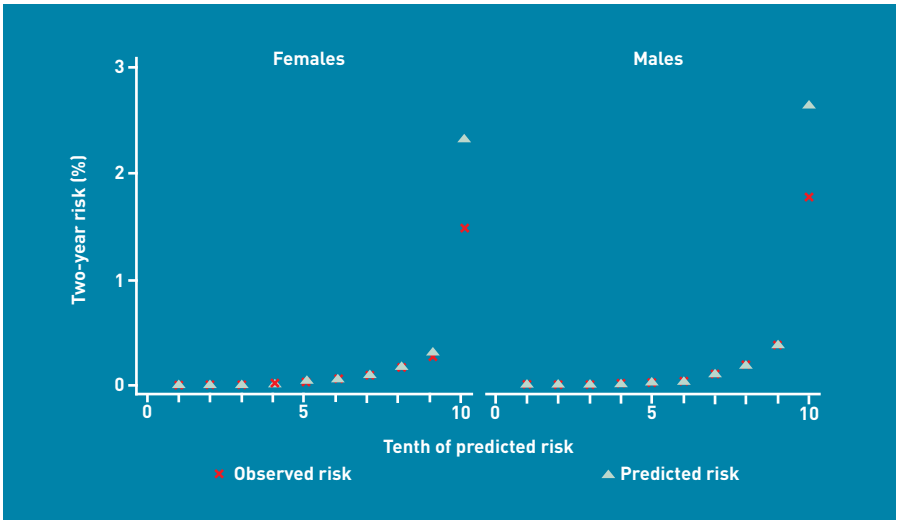


Figure 3. Mean predicted risk and observed risk of colorectal cancer at 2 years by tenth of predicted risk, applying the risk-prediction scores to the validation cohort.

There could be information bias and missing data, since not all patients with symptoms will attend their GP, and in those who do, not all symptoms will be reported or recorded. The effect of this information or recording bias could be to overinflate the hazard ratios if they relate to more severe symptoms, or underestimate the hazard ratios if patients with the symptoms do not have them recorded. The study has included 'change in bowel habit' as a predictor variable, as it tends to be included in national guidelines. However, it is possible that GPs document the words 'change in bowel habit' when they have a suspicion of cancer already. Nonetheless, the study analysis confirmed it as an independent predictor for males but not females and it is therefore included in the final model for males.

The study database has linked cause of death from the UK ONS, and the study is therefore likely to have picked up the majority of cases of colorectal cancer. This should help to minimise ascertainment bias and increase confidence in the reliability of the results. This is further supported by the incidence rates for symptoms and colorectal cancer in the study population, which are close to other published data with comparable proportions of colon and rectal cancer.[1]

While the validation cohort is derived from practices using the same clinical computer system (EMIS), they were physically discrete. Also, since this computer system is used in over half of UK general practices, the study results are likely to generalise well. A separate independent validation study using another GP database is planned and has not been included in the present study, so that it can be undertaken and published by an independent team.

Comparison with existing literature

While decisions currently made by doctors regarding which patients to investigate or refer tend to be based on many factors, previous studies and guidelines tend to focus on the predictive value of individual symptoms. The present study builds on previous work,[8,13,35] by providing an algorithm that can give a measure of absolute risk, taking account of age, symptoms, and family history, which can be integrated into clinical computer systems. It is reassuring that the risk ratios, sensitivities, and PPVs associated with individual symptoms found in this study are comparable to those reported elsewhere.[8,13,14] For example, Jones *et al* reported PPVs for colorectal cancer in the presence of rectal bleeding of 2.0% for females and 2.4% for males, and sensitivities of 25% and 33% respectively.[13] In the present study, the sensitivity of rectal bleeding as a single symptom in males and females combined was 33%, and the PPV was 2.9%.

Implications for practice and research

The algorithm can identify the 10% of the population in which approximately 70% of all new colorectal cancer cases are likely to be diagnosed over the next 2 years. Following external validation, this new algorithm could potentially be used to identify those at highest risk of having an existing colorectal cancer, to facilitate early referral and investigation and so help identify patients with colorectal cancer earlier, and potentially improve prognosis.

The authors recognise that use of the algorithm within a primary care consultation could have a major potential effect on referrals and use of investigations (such as colonoscopy), depending on how the algorithm is used and which thresholds are selected. The risks and benefits of decisions at various thresholds require

Box 1. Clinical examples

- A 60-year-old male who is a non-drinker with a positive family history of gastrointestinal cancer, anaemia, and a recent change in bowel habit has an estimated risk of colorectal cancer of 1.5%. If he also has loss of appetite, the estimated risk is 3.1%. If he also has rectal bleeding in addition to these symptoms, his estimated risk of colorectal cancer is 48.6%.
- A 45-year-old male who drinks 3 or more units of alcohol a day, who has a recent change in bowel habit, weight loss, and abdominal pain has an estimated risk of colorectal cancer of 2.4%. If he also has anaemia, his estimated risk of colorectal cancer is 7.9%. If he also has a positive family history of gastrointestinal cancer, his estimated risk is 11.8%
- A 70-year-old female with rectal bleeding and anaemia has an estimated risk of colorectal cancer of 10.5%. If she also has abdominal pain, her estimated risk of colorectal cancer is 42.5%.
- A 35-year-old female with abdominal pain and anaemia has a 0.1% estimated risk of colorectal cancer. If she also has loss of appetite and weight loss, her estimated risk of colorectal cancer is 5.3%.

further cost-effectiveness modelling, which is outside the scope of the present study.

The colorectal cancer algorithm is intended to help early diagnosis of an existing cancer, rather than to identify patients at high risk of a future cancer for prevention. As such, it differs from other validated algorithms already developed using the QResearch database and integrated into GP clinical computer systems. Examples include QRISK2®, which identifies patients with a high 10-year risk of developing cardiovascular disease,[15,20] and the QDScore®, which identifies patients with a high 10-year risk of developing type 2 diabetes.[17,19] A third example is the QFracture® score, which identifies people with a high 10-year risk of hip or osteoporotic fracture.[16,36] An equivalent algorithm could be developed and used for identifying patients at high risk of colorectal cancer for prevention and screening, although this would be a different algorithm from the one presented in this paper. It would be modelled over a longer period of time (for example, 10-year or lifetime risk of colorectal cancer), and is likely to include additional risk factors that operate over a longer period of time.

This study has developed and validated an algorithm that can be used to help identify symptomatic patients in primary care at high risk of having an existing, but as yet undiagnosed, colorectal cancer. This potentially offers an alternative approach to that offered in NICE guidelines to help improve the early diagnosis of colorectal cancer.

Funding

This work was undertaken by ClinRisk Ltd. There was no external funding.

Ethics committee

All QResearch® studies are independently reviewed in accordance with the QResearch® agreement with Trent Multi-Centre Ethics Committee (UK).

Provenance

Freely submitted; externally peer reviewed.

Web calculator

Here is a simple web calculator to implement the QCancer® (colorectal) algorithm, which is publically available alongside the paper and open source software (http://www.qcancer.org/colorectal).

Competing interests

Julia Hippisley-Cox is professor of clinical epidemiology at the University of Nottingham and co-director of QResearch® — a not-for-profit organisation which is a joint partnership between the University of Nottingham and EMIS (leading commercial supplier of IT for 60% of general practices in the UK). Julia Hippisley-Cox is also a paid director of ClinRisk Ltd, which produces software to ensure the reliable and updatable implementation of clinical risk algorithms within clinical computer systems to help improve patient care. Carol Coupland is associate professor of medical statistics at the University of Nottingham and a paid consultant statistician for ClinRisk Ltd. This work and any views expressed within it are solely those of the co-authors and not of any affiliated bodies or organisations.

Acknowledgements

We acknowledge the contribution of EMIS practices who contribute to QResearch® and EMIS for expertise in establishing, developing and supporting the database. The algorithms presented in this paper will be released as Open Source Software under the GNU lesser GPL v3.

Discuss this article

Contribute and read comments about this article on the Discussion Forum: **http://www.rcgp.org.uk/bjgp-discuss**

REFERENCES

1. Ferlay J, Autier P, Boniol M, *et al*. Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol* 2007; **18(3):** 581–592.
2. Berrino F, De Angelis R, Sant M, *et al*. Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995–99: results of the EUROCARE-4 study. *Lancet Oncol* 2007; **8(9):** 773–783.
3. O'Connell JB, Maggard MA, Ko CY. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *J Natl Cancer Inst* 2004; **96(19):** 1420–1425.
4. Davila RE, Rajan E, Baron TH, *et al*. ASGE guideline: colorectal cancer screening and surveillance. *Gastrointest Endosc* 2006; **63(4):** 546–557.
5. Thomson CS, Forman D. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EUROCARE results? *Br J Cancer* 2009; **101(suppl 2):** S102–109.
6. Richards MA. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br J Cancer* 2009; **101(suppl 2):** S1–4.
7. Department of Health. *The Cancer Reform Strategy*. London: Department of Health, 2007.
8. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. *Br J Cancer* 2009; **101(suppl 2):** S80–S86.
9. Majumdar SR, Fletcher RH, Evans AT. How does colorectal cancer present? symptoms, duration, and clues to location. *Am J Gastroenterol* 1999; **94(10):** 3039–3045.
10. National Institute for Health and Clinical Excellence. *Referral guidelines for suspected cancer*. London: National Institute for Health and Clinical Excellence, 2005.
11. Selvachandran SN, Hodder RJ, Ballal MS, *et al*. Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. *Lancet* 2002; **360(9329):** 278–283.
12. Jones R, Rubin G, Hungin P. Is the two week rule for cancer referrals working? *BMJ* 2001; **322(7302):** 1555–1556.
13. Jones R, Latinovic R, Charlton J, Gulliford MC. Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database. *BMJ* 2007; **334(7602):** 1040.
14. Jellema P, van der Windt DA, Bruinvels DJ, *et al*. Value of symptoms and additional diagnostic tests for colorectal cancer in primary care: systematic review and meta-analysis. *BMJ* 2010; 340: c1269.
15. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; **336(7659):** 1475–1482.
16. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 2009; **339:** b4229.
17. Hippisley-Cox J, Coupland C, Robson J, *et al*. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009; **338:** b880.
18. Hippisley-Cox J, Coupland C. Predicting the risk of chronic kidney disease in men and women in england and Wales: prospective derivation and external validation of the QKidney® Scores. *BMC Fam Pract* 2010; **11(1):** 49.
19. Collins GS, Altman DG. External validation of the QDScore for predicting the 10-year risk of developing type 2 diabetes. *Diabet Med* 2011; **28(5):** 599–607.
20. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010; **340:** c2442.
21. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009; **339:** b2584.
22. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods* 2002; **7(2):** 147–177.
23. Group TAM. Academic medicine: problems and solutions. *BMJ* 1989; **298:** 573–579.
24. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Epidemiol Community Health* 2007; **60(9):** 979.
25. Moons KGM, Donders RART, Stijnen T, Harrell FJ. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006; **59(10):** 1092.
26. Rubin DB. *Multiple imputation for non-response in surveys*. New York: John Wiley, 1987.
27. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999; **28(5):** 964–974.
28. Hosmer D, Lemeshow S. *Applied logistic regression*. New York, NY: John Wiley & Sons Inc., 1989.
29. Royston P. Explained variation for survival models. *Stata J* 2006; **6:** 1–14.
30. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23(5):** 723–748.
31. Hippisley-Cox J, Coupland C. Predicting the risk of osteoporotic fracture in England and Wales: prospective derivation and validation of QFracture scores. *BMJ* 2009; **339:** b4229.
32. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991; **302(6779):** 766–768.
33. Herrett E, Thomas SL, Schoonen WM, *et al*. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010; **69(1):** 4–14.
34. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010; **60(572):** e128–136.
35. Marshall T, Lancashire R, Sharp D, *et al*. The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. *Gut* 2011; **60(9):** 1242–1248.
36. Collins GS, Mallett S, Altman DG. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. *BMJ* 2011; **342:** d3651.