FULL PAPER

BJC

British Journal of Cancer (2017) 116, 944–950 | doi: 10.1038/bjc.2017.53

Keywords: colon cancer; electronic medical records; laboratory analytes; risk prediction; screening; colonoscopy

Predicting the presence of colon cancer in members of a health maintenance organisation by evaluating analytes from standard laboratory records

Ran Goshen[1], Barak Mizrahi[2], Pini Akiva[2], Yaron Kinar[2], Eran Choman[1], Varda Shalev[3], Victoria Sopik[4], Revital Kariv[5,6] and Steven A Narod*[,4,7]

[1]Medial Early Sign, Kfar Malal, Israel; [2]Medial Research, Kfar Malal, Israel; [3]Institute of Health Research and Innovation, Maccabi Healthcare Services, Tel-Aviv, Israel; [4]Familial Breast Cancer Research Unit, Women's College Research Institute, Women's College Hospital, 76 Grenville Street, 6th Floor, Suite 6420, Toronto, ON, Canada M5S 1B2; [5]Department of Gastroenterology, Souraski Medical Centre, Tel-Aviv, Israel; [6]Faculty of Medicine, Tel Aviv University, Tel-Aviv, Israel and [7]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Background: A valid risk prediction model for colorectal cancer (CRC) could be used to identify individuals in the population who would most benefit from CRC screening. We evaluated the potential for information derived from a panel of blood tests to predict a diagnosis of CRC from 1 month to 3 years in the future.

Methods: We abstracted information on 1755 CRC cases and 54 730 matched cancer-free controls who had one or more blood tests recorded in the electronic records of Maccabi Health Services (MHS) during the period 30–180 days before diagnosis. A scoring model (CRC score) was constructed using the study subjects' blood test results. We calculated the odds ratio for being diagnosed with CRC after the date of blood draw, according to CRC score and time from blood draw.

Results: The odds ratio for having CRC detected within 6 months for those with a score of four or greater (*vs* three or less) was 7.3 (95% CI: 6.3–8.5) for men and was 7.8 (95% CI: 6.7–9.1) for women.

Conclusions: Information taken from routine blood tests can be used to predict the risk of being diagnosed with CRC in the near future.

In certain circumstances, data abstracted from electronic medical records can be used to predict an impending clinical condition (Goldstein *et al*, 2016). The prescreening triage paradigm we propose here is applicable to diseases where early diagnosis is relevant to halt the progression from subclinical to frank disease (Benson *et al*, 2008; Byrd *et al*, 2014; Siu, 2015; Mamtani *et al*, 2016). For colorectal cancer (CRC), early diagnosis has been shown to be beneficial and screening is routinely recommended to adults (Benson *et al*, 2008). Colonoscopy is effective but is costly and many consider colonoscopy to be inefficient as a population-wide screening tool (Frazier *et al*, 2000). In Israel, all adults are offered screening for occult blood loss using the faecal immunochemical testing (FIT) annually (von Karsa *et al*, 2013). However, adherence to FIT screening and compliance with colonoscopy after a positive FIT test are suboptimal. Approximately 56% of eligible Israeli residents reported having had a colonoscopy in the past 10 years or

*Correspondence: Dr SA Narod; E-mail: steven.narod@wchospital.ca

Received 10 November 2016; revised 7 February 2017; accepted 8 February 2017; published online 2 March 2017

an FIT test in the past year (Klabunde *et al*, 2015). Adherence may be enhanced if individuals can be stratified into a high-risk group; that is, to identify men and women who have a higher than average risk of CRC for screening colonoscopy.

We have previously described MeScore, a machine learning-based algorithm, using historical complete blood counts, sex and age, and documented its performance in identifying patients with CRC (Kinar *et al*, 2016). In the current study, we evaluate the potential for using the results of all lab tests, including complete blood counts, obtained in the course of routine clinical care and recorded in an electronic health records database. We analysed the medical record database of a large HMO in Israel and linked these electronic records with the Israeli Cancer Registry.

MATERIALS AND METHODS

Study population. Macabbi Health Services (MHS) is the second largest HMO in Israel with approximately two million enrollees. We abstracted information on the patient population from MHS electronic medical record database, including demographic data (sex and date of birth) and the results of all blood tests conducted from January 2001 until the end of 2011.

Cases were MHS enrollees who were diagnosed with CRC between 40 and 75 years of age between 2002 and 2011. To be included in this study, cases had to have had at least one blood test recorded in MHS electronic medical records before the date of diagnosis (index date). Individuals with a history of CRC or with another form of cancer before 2002 were excluded from the analysis. The diagnosis of CRC among MHS patients was made through linkage to the Israeli Cancer Registry using a unique national identifier.

Controls were selected from among all individuals in MHS registry who did not have any cancer, according to the Israeli Cancer Registry. Control subjects were matched to cases based on sex and year of birth. Controls were required to have had at least one blood test before the index date of the matched case. All controls were iteratively matched with cases, resulting in a fixed ratio of 45 controls per case (Table 1).

Data were taken from the results of the blood tests extracted from the patients' electronic medical records. All blood tests were analysed at a single national lab. The blood tests were ordered for a variety of reasons, including routine health examination and for specific clinical indications, but we did not have details regarding the specific indication. We included all blood analytes in our initial analyses, provided that the test had been conducted on a minimum of 10% of the controls. For the purposes of this study, the index blood test was defined as one that was carried out from 1 to 6 months before the date of diagnosis of CRC in the cases or the index date for controls. In the event that more than one blood test was carried out during this interval, the most recent one was considered the index test. The complete set of analytes is listed in Table 2. For each of the analytes, we compared the mean value of the cases and controls and assessed the significance of the difference using a *t*-test. We then considered the distribution of each test parameter in the control group, arranging the controls into five quintiles of equal size. We compared the distribution of cases and controls by quintile and then calculated the odds ratio for the high-risk quintile, relative to the reference quintile. The reference quintile was either the top or the bottom quintile and was chosen *post hoc* as the one which was associated with an odds ratio above unity.

Table 1. Characteristics of cases and controls

	Cases	Controls
Number of patients	2294	102 775
Age at index blood test	62.2	61.5
Males	1226	55 000
Females	1068	47 775
Year of birth	1943.8	1944.4
Age at colon cancer (years)	62.6	
Localised	568	
Regional	1020	
Distant	118	
Other	588	
Date of diagnosis of colon cancer	2006.5	

Generating a logistic regression model to choose the lab tests that contribute to CRC prediction (feature selection). To reduce the number of analytes into a small reference panel of parameters that carried the most information, for each individual we assigned a value of one for those in the high-risk category or zero for those in the low-risk category or with a missing value. To account for correlation between analytes within a functional group (as defined in Table 2) we applied a logistic regression model to the parameters within that group and then flagged those analytes that were statistically significant. From this set (all groups combined) we then generated a final regression model that included only those lab tests that showed significant difference ($P<0.0001$) and for which the odds ratio exceeded 1.5. There were nine variables in the canonical sets for men and women. We created a CRC score for each individual, which ranged from zero to nine, based on the number of risk factors. We then compared the distribution of the CRC scores for cases and controls for men and for women and we generated odds ratios for each risk score, using the reference level of zero.

The CRC risk score was developed using cancer cases diagnosed at all stages using values from the blood test most recently acquired before diagnosis. We wished to evaluate the potential for using the CRC risk score to identify CRC cases diagnosed at an early (i.e., treatable) stage as well as to predict the development of CRC in the future. First, we repeated the analysis, restricting the cases to early-stage (localised) CRC. Second, we generated the distribution of the risk score for analytes for lab tests taken at various time windows before the index date to see to what extent the risk score predicted the risk of developing CRC up to 3 years in the future.

Evaluating sensitivity and specificity. We evaluated the overall performance of the risk score for predicting CRC in the entire MHS population. We assigned a risk score for each individual in the MHS database. We then calculated the performance of the score for different age groups and the score cutoffs for unmatched population (all controls).

RESULTS

There were 102 775 subjects in the electronic medical record of MHS (Table 1). Of these, 2294 (2.2%) were diagnosed with CRC between the ages of 40 and 75 years in the period of diagnosis 2002 to 2011. Of these, 1755 cases (77%) had one or more blood test recorded during the period 30–180 days before diagnosis. Of the 1755 eligible CRCs, 716 were distal, 367 were proximal and 539 were rectal cancers (unknown 133). Of the 1755 CRCs, 450 were localised (26%), 774 were regional (44%) and 93 had distant metastases (5.3%) (438 other or stage unknown). Each case was matched to 45 controls.

We studied 30 analytes in five groups. The cases and controls were compared for all 30 analytes using a *t*-test (Table 2). We then selected those analytes that were most helpful in discriminating between cases and controls, by generating odds ratios using the

Table 2. All analytes

Blood test	Cases Mean	Controls Mean	RR (high-risk quintile *vs* low-risk quintile)	*P*-value (*t*-test)
Females				
Haematology				
Basophils	0.03	0.03	1.19 (1.02–1.48)	0.0003
Eosinophils	0.21	0.18	2.03 (1.58–2.79)	<0.0001
Haemoglobin	11.8	13.02	5.69 (4.31–7.97)	<0.0001
Lymphocytes	2.25	2.21	1.37 (1.06–1.78)	0.43
Mean corpuscular volume	84.50	88.64	3.52 (2.84–4.39)	<0.0001
Monocytes	0.56	0.51	1.99 (1.63–2.65)	<0.0001
Mean platelet volume	10.78	11.06	2.33 (1.72–3.26)	<0.0001
Neutrophils	4.33	3.70	3.02 (2.42–4.17)	<0.0001
Platelets	301	254	3.87 (3.09–5.21)	<0.0001
Red blood cell count	4.39	4.48	1.97 (1.51–2.61)	<0.0001
Red blood cell distribution width	14.81	13.71	4.54 (3.58–6.26)	<0.0001
White blood cell count	7.46	6.65	2.17 (1.66–3.02)	<0.0001
Liver function				
Alkaline phosphatase	92.37	76.39	2.43 (1.75–3.33)	<0.0001
Alanine aminotransferase	18.94	21.91	3.12 (2.45–4.2)	<0.0001
Aspartate aminotransferase	21.23	22.64	2.45 (1.82–3.2)	0.011
Bilirubin	0.52	0.60	2.31 (1.71–3.25)	<0.0001
γ-Glutamyl transferase	58.76	38.35	1.42 (1.04–2.15)	0.10
Metabolic				
Albumin	4.08	4.22	2.62 (1.91–3.44)	<0.0001
Cholesterol	201	210	1.77 (1.29–2.39)	<0.0001
Glucose	110	109	1.2 (1.03–1.74)	0.28
Haemoglobin A1c	6.57	6.65	1.06 (1–1.75)	0.16
Lactic acid dehydrogenase	375	351	1.32 (1.04–2.57)	0.17
Low-density lipoprotein	121	126	1.4 (1.09–1.86)	0.00082
Protein	7.11	7.28	2.12 (1.54–3.08)	<0.0001
Prothrombin time (%)	15.00	15.47	4.58 (2.97–8)	0.44
Partial thromboplastin time	25.97	27.19	1.75 (1.06–4.13)	0.00016
Other				
Fe – iron	46.84	77.22	16 (9.14–33.4)	<0.0001
Ferritin	46.24	69.62	4.59 (3.06–7.73)	<0.0001
Vitamins				
Folic acid	4.33	4.43	1.91 (1.28–2.96)	0.37
Vitamin B12	430	426	1.07 (1–1.54)	0.40
Males				
Haematology				
Basophils	0.03	0.03	1.4 (1.14–1.75)	0.0017
Eosinophils	0.25	0.22	1.62 (1.29–2.04)	<0.0001
Haemoglobin	13.3	14.43	3.77 (2.66–5.08)	<0.0001
Lymphocytes	2.13	2.21	1.17 (1.01–1.53)	0.026
Mean corpuscular volume	85.7	88.90	3.44 (2.7–4.87)	<0.0001
Monocytes	0.68	0.61	2.11 (1.74–2.8)	<0.0001
Mean platelet volume	10.78	11.07	2.33 (1.8–2.93)	<0.0001
Neutrophils	4.69	4.13	2.29 (1.73–2.96)	<0.0001
Platelets	261	222	3.78 (2.95–4.88)	<0.0001
Red blood cell count	4.76	4.87	1.75 (1.45–2.24)	<0.0001
Red blood cell distribution width	14.26	13.61	2.87 (2.23–3.78)	<0.0001
White blood cell count	7.79	7.20	2.31 (1.87–3.05)	<0.0001
Liver function				
Alkaline phosphatase	86	73.36	2.76 (2.21–3.88)	<0.0001
Alanine aminotransferase	20.97	25.61	3.61 (2.69–4.72)	<0.0001
Aspartate aminotransferase	21.19	23.77	2.75 (2.17–3.88)	<0.0001
Bilirubin	0.63	0.74	2.71 (1.9–3.72)	<0.0001
γ-Glutamyl transferase	59.73	45.92	1.29 (1.02–2.03)	0.026
Metabolic				
Albumin	4.19	4.30	1.97 (1.5–2.55)	<0.0001
Cholesterol	184	189.48	1.46 (1.1–1.9)	0.00038
Glucose	112	114	1.13 (1–1.43)	0.046
Haemoglobin A1c	6.76	6.74	1.07 (1–1.67)	0.40
Lactic acid dehydrogenase	332	341	1.64 (1.05–3.57)	0.24
Low-density lipoprotein	113	116	1.28 (1.02–1.55)	0.044
Protein	7.20	7.30	2.62 (1.81–3.58)	0.00034
Prothrombin time (%)	21.32	22.63	3 (2.07–4.67)	0.33
Partial thromboplastin time	27.85	28.55	1.61 (1.04–3.17)	0.079
Other				
Fe – iron	54.68	84.35	5.38 (3.36–9.08)	<0.0001
Ferritin	73.53	124.84	4.95 (3.22–8.45)	<0.0001
Vitamins				
Folic acid	4.42	4.46	1.53 (1.06–2.52)	0.44
Vitamin B12	381	393	1.39 (1.02–2.5)	0.16

Abbreviation: RR = risk ratio.

logistic regression model, based on the distribution of cases and controls in the highest risk quintile, *vs* the bottom quintile.

To generate the final model, we incorporated nine analytes. The nine analytes were selected from the various groupings based on the odds ratios (>1.5) and the corresponding *P*-values (<0.0001). For each of the analytes, we estimated the odds ratio for an individual in the high-risk quintile, compared with all other individuals. The (mutually adjusted) odds ratios associated with the high-risk quintile for the analytes in the canonical panel (*vs* all others) ranged from 1.7 (for low protein, females) to 6.0 (for low iron, females) (Table 3).

Finally, a risk score was generated for each individual, which was based on the total number of risk factors that an individual carried (Table 4). Among controls, the mean risk score was 1.23 for men and was 1.25 for women. Among cases, the mean risk score was 2.62 for men and was 2.86 for women. Among men, the odds ratios associated with risk scores above zero ranged from 1.5 for those with a risk score of one to 233 for those with a risk score of eight or nine. Among women, the odds ratios associated with risk scores above zero ranged from 1.5 for those with a risk score of one to 76 for those with a risk score of eight or nine. The odds ratio for a risk score of four or greater for men (*vs* three or less) was 7.3 (95% CI: 6.3–8.5). The odds ratio for a risk score of four or more for women (*vs* three or less) was 7.8 (95% CI: 6.7–9.1). For localised cancer, the odds ratio of a risk score of four or greater for men (*vs* three or less) was 6.1 (95% CI: 4.6–8.2) and the odds ratio for a risk score of four or more for women (*vs* three or less) was 5.0 (Supplementary Table 1).

It is hoped that the test could be used to identify cancers well in advance of clinical symptoms. We analysed the predictive value (odds ratio) for risk scores based on blood values taken in advance of the date of diagnosis (Supplementary Table 2). Using a cutoff of 4+ for women, the odds ratio was 4.95 (95% CI: 4.1–6.0) for the period 90–180 days before diagnosis, was 3.2 (95% CI: 2.7–3.9) for 180 days to 1 year, was 2.3 (95% CI: 1.9–2.8) for 1–2 years and was 1.3 (95% CI: 1.0–1.7) for 2–3 years (Table 5). Using a cutoff of 4+ for men, the odds ratio was 5.1 (95% CI: 4.2–6.2) for the period 90–180 days before diagnosis, was 3.5 (95% CI: 2.8–4.2) for 180 days to 1 year, was 2.1 (95% CI: 1.7–2.6) for 1–2 years and was 1.3 (95% CI: 1.0–1.6) for 2–3 years.

Table 3. Final model (logistic regression)

Blood test	High-risk quintile	RR (95% CI)	95% CI	*P*-value
Females				
Haematology				
Haemoglobin	5	3.83	3.38–4.46	<0.0001
Mean corpuscular volume	5	3.04	2.7–3.54	<0.0001
Neutrophils	1	2.03	1.82–2.35	<0.0001
Platelets	1	2.95	2.56–3.35	<0.0001
Red blood cell distribution width	1	3.14	2.81–3.66	<0.0001
Liver function				
Alanine aminotransferase	5	1.83	1.6–2.14	<0.0001
Metabolic				
Protein	5	1.71	1.42–2.01	<0.0001
Other				
Fe – iron	5	6.01	5.08–7.08	<0.0001
Ferritin	5	5.04	4.47–6.26	<0.0001
Males				
Haematology				
Haemoglobin	5	3.06	2.76–3.52	<0.0001
Mean corpuscular volume	5	2.98	2.58–3.42	<0.0001
Monocytes	1	1.85	1.6–2.12	<0.0001
Platelets	1	2.84	2.5–3.27	<0.0001
Liver function				
Alkaline phosphatase	1	2.10	1.79–2.34	<0.0001
Alanine aminotransferase	5	2.09	1.82–2.36	<0.0001
Aspartate aminotransferase	5	1.77	1.52–2.04	<0.0001
Other				
Fe – iron	5	5.44	4.47–6.16	<0.0001
Ferritin	5	5.75	4.74–6.9	<0.0001

Abbreviations: CI = confidence interval; RR = risk ratio.

To estimate the sensitivity of the high-risk score at a specificity of 95%, we used the entire MHS database (Table 6). Using a cutoff level of risk score of 4+ for men, the sensitivity was 31%, the specificity was 95% and the positive predictive value of an abnormal score was 7.3%. For women, using a cutoff of 5+, the sensitivity was 24%, the specificity was 95% and the positive predictive value was 4.2%.

DISCUSSION

We show that information abstracted from the electronic medical records of a large health-care provider can be adapted to identify a

Table 4. Distribution of risk scores for cases and controls with associated odds ratios

Score	Controls	Cases	% of Cases	Odds ratio	95% CI
Females					
0	9643	127	16	1.00	0.78–1.28
1	7628	146	18	1.45	1.14–1.85
2	4590	137	17	2.27	1.78–2.89
3	2470	98	12	3.01	2.31–3.94
4	1129	111	14	7.47	5.74–9.7
5	513	76	9	11.25	8.35–15.15
6	193	79	10	31.08	22.68–42.58
7	59	31	4	39.89	24.97–63.75
8/9	14	14	1	75.93	35.46–1626
Total	26 239	819	100		
Males					
0	10 259	147	16	1.00	0.79–1.26
1	8302	180	19	1.51	1.21–1.89
2	5438	164	18	2.10	1.68–2.64
3	2834	154	16	3.79	3.01–4.77
4	1080	113	12	7.30	5.67–9.41
5	398	90	10	15.78	11.92–20.9
6	135	48	5	24.81	17.18–35.83
7	42	30	3	49.85	30.36–81.86
8/9	3	10	1	232.6	63.37–854.2
Total	28 491	936	100		

Abbreviation: CI = confidence interval.

Table 5. Odds ratios for colorectal cancer associated with a risk score of four or more (*vs* three or less), by time period before diagnosis

Time period before diagnosis	Odds ratio for CRC (risk score 4–9 *vs* 0–3)	95% CI
Females		
30–180 days	7.8	6.7–9.1
90–180 days	5.0	4.1–6.0
180 days to 1 year	3.2	2.7–3.9
1–2 years	2.3	1.9–2.8
2–3 years	1.3	1.0–1.7
Males		
30–180 days	7.3	6.3–8.5
90–180 days	5.1	4.2–6.2
180 days to 1 year	3.5	2.8–4.2
1–2 years	2.1	1.7–2.6
2–3 years	1.3	1.0–1.6

Abbreviation: CI = confidence interval.

Table 6. Characteristics of screening test

Cutoff	False negatives	True Negatives	False positives	True positives	Total	Specificity	Sensitivity
Females							
All ages							
⩾6	695	90 533	1654	124	93 006	98.2%	15%
⩾5	619	87 681	4506	200	93 006	95.1%	24%
⩾4	508	82 099	10 088	311	93 006	89.1%	38%
40–50 years							
⩾6	58	40 747	1152	13	41 970	97.3%	18%
⩾5	47	38 885	3014	24	41 970	92.8%	34%
⩾4	37	35 387	6512	34	41 970	84.5%	48%
50–60 years							
⩾6	190	25 521	292	35	26 038	98.9%	16%
⩾5	171	25 022	791	54	26 038	96.9%	24%
⩾4	139	23 979	1834	86	26 038	92.9%	38%
60–70 years							
⩾6	297	17 274	128	35	17 734	99.3%	11%
⩾5	270	16 978	424	62	17 734	97.6%	19%
⩾4	231	16 353	1049	101	17 734	94.0%	30%
70–75 years							
⩾6	150	6991	82	41	7264	98.8%	21%
⩾5	131	6796	277	60	7264	96.1%	31%
⩾4	101	6380	693	90	7264	90.2%	47%
Males							
All ages							
⩾6	848	71 285	314	88	72 535	99.6%	9%
⩾5	758	70 454	1145	178	72 535	98.4%	19%
⩾4	645	67 923	3676	291	72 535	94.9%	31%
40–50 years							
⩾6	72	30 532	102	14	30 720	99.7%	16%
⩾5	68	30 292	342	18	30 720	98.9%	21%
⩾4	58	29 322	1312	28	30 720	95.7%	33%
50–60 years							
⩾6	199	21 325	81	11	21 616	99.6%	5%
⩾5	179	21 092	314	31	21 616	98.5%	15%
⩾4	154	20 409	997	56	21 616	95.3%	27%
60–70 years							
⩾6	346	13 959	80	36	14 421	99.4%	9%
⩾5	305	13 716	323	77	14 421	97.7%	20%
⩾4	267	13 142	897	115	14 421	93.6%	30%
70–75 years							
⩾6	231	5469	51	27	5778	99.1%	10%
⩾5	206	5354	166	52	5778	97.0%	20%
⩾4	166	5050	470	92	5778	91.5%	36%

sub-population at risk of having CRC up to 2 years before diagnosis. We generated a CRC score, which varied from zero to nine, based on laboratory values that were recorded for a minimum of 10% of the subjects. As the risk score increased, the odds ratio for a cancer being diagnosed in the near term (30–180 days post-test) increased greatly. A score of four or more was associated with an odds ratio of 7.3 for men and of 7.8 for women, respectively. For subjects with the highest risk scores, the odds ratios exceeded 50. At a specificity of 95%, the sensitivity of a CRC score was 31% for men and was 24% for women. Among those in the top 5% of the CRC score, the positive predictive value for colon cancer was 7.3% for men and 4.2% for women.

In the MHS health plan, 2294 of 379 701 individuals, aged 40–75 years (36%) were diagnosed with CRC from 2002 to 2011 (Table 1). That is, the annual incidence of CRC was 61 per 100 000 per year – in contrast to the incidence of 600 per 100 000 per year among those individuals with a CRC score of five or more.

There are limitations to our study. This is a large observational study and the data were collected for clinical care and monitoring health-care services and was not designed for the purposes of evaluating screening. As a consequence, the data on the laboratory values and the CRC diagnoses were less detailed than we would have liked. The blood samples were taken at various time intervals before cancer diagnosis and blood collection was not standardised. The levels for many of the analytes vary over time and each analyses were based on a single test result per subject. Furthermore, the tests were ordered by the treating physicians for various clinical reasons, largely unrelated to CRC. For example, the mean number of analytes for which information was available was 6.5 for the male cases and was 6.2 for the male controls (out of nine) and the mean number of analytes was 6.7 for the female cases and was 6.4 for the female controls (out of nine). Ideally, we would have had a complete data set for each case and each control. However, it is a testament to the strength of this approach that we were able to generate significant and large risk ratios despite a high level of missing values. Further, the goal of our initiative is to provide a prescreening tool to health-care institutions that can be implemented immediately using existing medical records and at little cost. We did not incorporate other factors in the model such as age, family history of cancer, BMI, or aspirin use, previous colonoscopies or adenomatous polyps.

Our study is similar to a previous study of Boursi *et al* (2016), which was conducted using a UK-based data set (THIN). In that study, which included 13 879 colon cancer cases and 54 109 matched controls, the risk model contained several of the haematologic variables we have included here, including red blood cell counts and neutrophils. In that study, the odds ratio for individuals in the highest *vs* the lowest risk decile was 18. In comparison, in our study, for individuals in the highest decile, the odds ratio was ∼28. For both the studies, the discriminatory power greatly exceeded that of previous models based on demographic factors and other risk factors (Ma and Ladabaum, 2014; Tao *et al*, 2014; Imperiale *et al*, 2015).

The specific components of the CRC score fall into three principal categories: iron deficiency, inflammation and thrombosis, and liver function. Those individuals in the lowest quintile of haemoglobin level (<13.5 for men and <12.2 for women) were at increased risk for the development of CRC, presumably due to occult colonic bleeding, but the magnitude of the associations using haemoglobin alone (odds ratios three to four) were much less than those generated using the CRC score. There were also very strong and independent associations with low levels of serum iron and ferritin, which also contributed to risk beyond that of haemoglobin and red cell morphology. A high platelet count (top quintile *vs* others) was associated with a three-fold increased risk of CRC. Platelets have been implicated in CRC carcinogenesis and progression in several studies, although the exact mechanism by which platelets contribute to CRC incidence and metastasis is not known (Guillem-Llobat *et al*, 2014; Seretis *et al*, 2015). Possible mechanisms include transport of cancer cells through intravasation and extravasation in and out of vessels and promoting cancer cell aggregation (Stegner *et al*, 2014; Seretis *et al*, 2015; Patrignani and Patrono, 2016). Platelets also induce epithelial–mesenchymal transition of cancer cells, thereby enhancing their metastatic ability (Patrignani and Patrono, 2016). Importantly, there is now compelling evidence that daily low-dose aspirin is an effective chemoprevention agent against CRC (Cole *et al*, 2009; Rothwell *et al*, 2011; Cao *et al*, 2016). It is proposed that the protective effect of aspirin works through the inhibition of platelet activation at GI mucosal lesions during the early stages of colorectal carcinogenesis (Thun *et al*, 2012; Di Francesco *et al*, 2015; Patrignani and Patrono, 2016). We hope to replicate the findings of the current study using other populations and to conduct studies to confirm the potential of our preliminary findings in terms of facilitating the early diagnosis of cancer.

We were able to predict cancers up to 2 years in advance of the clinical diagnosis using the CRC score, but after 2 years the discriminant ability fell off. One would expect that a 2-year advance in the date of diagnosis would be of clinical utility, and it is encouraging that the CRC score effectively predicted the presence of localised cancer; it is expected that the majority of these cases are curable (Gatta *et al*, 2000). The risk score was performed optimally for the time window from 30 to 180 days before diagnosis, but it was also predictive of cancer 2 years in the future; this lends support to the hope that this can be a useful adjunct to the early detection of CRC. Also, screening colonoscopy is not routinely recommended to men or women before age 50 years; in the age group 40–50 years, those with a CRC score of six or greater had a high risk of CRC that was comparable or greater to much older individuals (Table 5) and it is rational to consider that screening colonoscopy may be extended to this small high-risk subgroup.

The data presented here support the principle that information contained in the laboratory record of a large health services organisation can be used to predict the risk of CRC. Within these large databases, information that is relevant to clinical care may not be apparent using clinical judgment alone but might be exploited using electronic medical records and statistical methodologies. Valuable information may be obtained by interpreting multiple analytes that may fall within the clinical norms. In this study, we use simplified statistical tools designed into an innovative scoring model to document the performance of a CRC score. Through the use of machine learning on a wide set of analytes, we hope to significantly improve test performance.

The future use of machine learning is expected to allow us to handle blood test values in a continuous (rather than discrete) way, to provide different weights to each test, and to combine the test values in a nonlinear and complex manner.

It is to be determined if the reporting of the scoring to the treating physician will improve adherence and compliance with current screening guidelines.

ACKNOWLEDGEMENTS

This work was supported by Medial Research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Benson VS, Patnick J, Davies AK (2008) Colorectal cancer screening: a comparison of 35 initiatives in 17 countries. *Int J Cancer* **122**: 1357–1367.

Boursi B, Mamtani R, Hwang WT, Haynes K, Yang YX (2016) A risk prediction model for sporadic CRC based on routine lab results. *Dig Dis Sci* **61**: 2076–2086.

Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF (2014) Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int J Med Inform* **83**: 983–992.

Cao Y, Nishihara R, Wu K, Wang M, Ogino S, Willett WC, Spiegelman D, Fuchs CS, Giovannucci EL, Chan AT (2016) Population-wide impact of long-term use of aspirin and the risk for cancer. *JAMA Oncol* **2**: 762–769.

Cole BF, Logan RF, Halabi S, Benamouzig R, Sandler RS, Grainge MJ, Chaussade S, Baron JA (2009) Aspirin for the chemoprevention of colorectal adenomas: meta-analysis of the randomized trials. *J Natl Cancer Inst* **101**: 256–266.

Di Francesco L, López Contreras LA, Sacco A, Patrignani P (2015) New insights into the mechanism of action of aspirin in the prevention of colorectal neoplasia. *Curr Pharm Des* **21**: 5116–5126.

Frazier AL, Colditz GA, Fuchs CS, Kuntz KM (2000) Cost-effectiveness of screening for colorectal cancer in the general population. *JAMA* **284**: 1954–1961.

Gatta G, Capocaccia R, Sant M, Bell CM, Coebergh JW, Damhuis RA, Faivre J, Martinez-Garcia C, Pawlega J, Ponz de Leon M, Pottier D, Raverdy N, Williams EM, Berrino F (2000) Understanding variations in survival for colorectal cancer in Europe: a EUROCARE high resolution study. *Gut* **47**: 533–538.

Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP (2016) Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* **24**: 198–208.

Guillem-Llobat P, Dovizio M, Alberti S, Bruno A, Patrignani P (2014) Platelets, cyclooxygenases, and colon cancer. *Semin Oncol* **41**: 385–396.

Imperiale TF, Monahan PO, Stump TE, Glowinski EA, Ransohoff DF (2015) Derivation and validation of a scoring system to stratify risk for advanced colorectal neoplasia in asymptomatic adults: a cross-sectional study. *Ann Intern Med* **163**: 339–346.

Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, Chodick G, Shalev V (2016) Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *J Am Med Inform Assoc* **23**: 879–890.

Klabunde C, Blom J, Bulliard JL, Garcia M, Hagoel L, Mai V, Patnick J, Rozjabek H, Senore C, Törnberg S (2015) Participation rates for organized colorectal cancer screening programmes: an international comparison. *J Med Screen* **22**: 119–126.

Ma GK, Ladabaum U (2014) Personalizing colorectal cancer screening: a systematic review of models to predict risk of colorectal neoplasia. *Clin Gastroenterol Hepatol* **12**: 1624–1634.

Mamtani M, Kulkarni H, Wong G, Weir JM, Barlow CK, Dyer TD, Almasy L, Mahaney MC, Comuzzie AG, Glahn DC, Magliano DJ, Zimmet P, Shaw J, Williams-Blangero S, Duggirala R, Blangero J, Meikle PJ, Curran JE (2016) Lipidomic risk score independently and cost-effectively predicts risk of future type 2 diabetes: results from diverse cohorts. *Lipids Health Dis* **15**: 67.

Patrignani P, Patrono C (2016) Aspirin and Cancer. *J Am Coll Cardiol* **68**: 967–976.

Rothwell PM, Fowkes FG, Belch JF, Ogawa H, Warlow CP, Meade TW (2011) Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *Lancet* **377**: 31–41.

Seretis C, Youssef H, Chapman M (2015) Hypercoagulation in colorectal cancer: what can platelet indices tell us? *Platelets* **26**: 114–118.

Siu AL (2015) Screening for high blood pressure in adults: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* **163**: 778–786.

Stegner D, Dütting S, Nieswandt B (2014) Mechanistic explanation for platelet contribution to cancer metastasis. *Thromb Res* **133**(Suppl 2): S149–S157.

Tao S, Hoffmeister M, Brenner H (2014) Development and validation of a scoring system to identify individuals at high risk for advanced colorectal neoplasms who should undergo colonoscopy screening. *Clin Gastroenterol Hepatol* **12**: 478–485.

Thun MJ, Jacobs EJ, Patrono C (2012) The role of aspirin in cancer prevention. *Nat Rev Clin Oncol* **9**: 259–267.

von Karsa L, Patnick J, Segnan N, Atkin W, Halloran S, Lansdorp-Vogelaar I, Malila N, Minozzi S, Moss S, Quirke P, Steele RJ, Vieth M, Aabakken L, Altenhofen L, Ancelle-Park R, Antoljak N, Anttila A, Armaroli P, Arrossi S, Austoker J, Banzi R, Bellisario C, Blom J, Brenner H, Bretthauer M, Camargo Cancela M, Costamagna G, Cuzick J, Dai M, Daniel J, Dekker E, Delicata N, Ducarroz S, Erfkamp H, Espinàs JA, Faivre J, Faulds Wood L, Flugelman A, Frkovic-Grazio S, Geller B, Giordano L, Grazzini G, Green J, Hamashima C, Herrmann C, Hewitson P, Hoff G, Holten I, Jover R, Kaminski MF, Kuipers EJ, Kurtinaitis J, Lambert R, Launoy G, Lee W, Leicester R, Leja M, Lieberman D, Lignini T, Lucas E, Lynge E, Mádai S, Marinho J, Maučec Zakotnik J, Minoli G, Monk C, Morais A, Muwonge R, Nadel M, Neamtiu L, Peris Tuser M, Pignone M, Pox C, Primic-Zakelj M, Psaila J, Rabeneck L, Ransohoff D, Rasmussen M, Regula J, Ren J, Rennert G, Rey J, Riddell RH, Risio M, Rodrigues V, Saito H, Sauvaget C, Scharpantgen A, Schmiegel W, Senore C, Siddiqi M, Sighoko D, Smith R, Smith S, Suchanek S, Suonio E, Tong W, Törnberg S, Van Cutsem E, Vignatelli L, Villain P, Voti L, Watanabe H, Watson J, Winawer S, Young G, Zaksas V, Zappa M, Valori R (2013) European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. *Endoscopy* **45**: 51–59.

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)