CrossMark

ORIGINAL ARTICLE

Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data

Mark C. Hornbrook[1] · Ran Goshen[2] · Eran Choman[2] · Maureen O'Keeffe-Rosetti[1] · Yaron Kinar[2,3] · Elizabeth G. Liles[1] · Kristal C. Rust[1,4]

Abstract

Background Machine learning tools identify patients with blood counts indicating greater likelihood of colorectal cancer and warranting colonoscopy referral.

Aims To validate a machine learning colorectal cancer detection model on a US community-based insured adult population.

Methods Eligible colorectal cancer cases (439 females, 461 males) with complete blood counts before diagnosis were identified from Kaiser Permanente Northwest Region's Tumor Registry. Control patients ($n = 9108$) were randomly selected from KPNW's population who had no cancers, received at ≥ 1 blood count, had continuous enrollment from 180 days prior to the blood count through 24 months after the count, and were aged 40–89. For each control, one blood count was randomly selected as the pseudo-colorectal cancer diagnosis date for matching to cases, and assigned a "calendar year" based on the count date. For each calendar year, 18 controls were randomly selected to match the general enrollment's 10-year age groups and lengths of continuous enrollment. Prediction performance was evaluated by area under the curve, specificity, and odds ratios.

Results Area under the receiver operating characteristics curve for detecting colorectal cancer was 0.80 ± 0.01. At 99% specificity, the odds ratio for association of a high-risk detection score with colorectal cancer was 34.7 (95% CI 28.9–40.4). The detection model had the highest accuracy in identifying right-sided colorectal cancers.

Conclusions ColonFlag® identifies individuals with ten-fold higher risk of undiagnosed colorectal cancer at curable stages (0/I/II), flags colorectal tumors 180–360 days prior to usual clinical diagnosis, and is more accurate at identifying right-sided (compared to left-sided) colorectal cancers.

✉ Mark C. Hornbrook
  mark.c.hornbrook@gmail.com

  Ran Goshen
  ran@earlysign.com

  Eran Choman
  eran@earlysign.com

  Maureen O'Keeffe-Rosetti
  maureen.rosetti@kpchr.org

  Yaron Kinar
  yaron@medial-research.com

  Elizabeth G. Liles
  Elizabeth.G.Liles@kp.org

  Kristal C. Rust
  kristal.c.rust@kp.org

[1] Kaiser Permanente Center for Health Research, 3800 North Interstate Avenue, Portland, OR 97227-1110, USA

[2] Medial EarlySign Inc., 11 HaZait St., Kfar Malal, Israel

[3] Medial Research, Inc., 11 HaZait St., Kfar Malal, Israel

[4] Kaiser Sunnyside Medical Center, LL Nursing Administration, 10180 SE Sunnyside Road, Clackamas, OR 97015, USA

Background and Aims

An estimated 134,492 new cases of colorectal cancer (CRC), evenly distributed among men and women, were diagnosed in 2016 in the USA, and 49,190 persons died from CRC in 2016—26,020 males and 23,170 females [1].

CRC is the fourth most prevalent cancer in the USA (after prostate, breast, and lung) and the second-highest cause of cancer-related deaths (after lung) [1].

Research has demonstrated that screening for CRC improves survival and reduces mortality rate. The American Cancer Society, American College of Physicians, American Gastroenterology Association, European Union Health Program, and United States Preventive Services Task Force (USPTF) clinical guidelines, based on scientific evidence, recommend CRC screening for all individuals over the age of 50 even if no additional risk factors are present [2–5]. The USPTF rates CRC screening at age 50 through 75 years as "Grade A—likely net benefit is substantial" [6, 7]. In addition, specialty guidelines recommended earlier or more frequent CRC screening for members of high-risk population groups—hereditary CRC syndromes, individuals with first-degree relatives with CRC, and patients suffering from inflammatory bowel disease [2–5]. Research has demonstrated that compared with no endoscopic screening, receipt of a screening colonoscopy is associated with a 67% reduction in the risk of death from any colorectal cancer [adjusted odds ratio (aOR) = 0.33, 95% confidence interval (CI) 0.21–0.52] [8]. By cancer location, screening colonoscopy is associated with a 65% reduction in risk of death for right colon cancers (aOR = 0.35, CI 0.18–0.65) and a 75% reduction for left colon/rectal cancers (aOR = 0.25, CI 0.12–0.53) [8].

The USPTF rates screening for CRC in older adults aged 76–85 years as "Grade C—likely net benefit is small" [6, 7], and recommends that adults in this age group who have never been screened for colorectal cancer would be more likely to benefit, and that screening would be most appropriate among older adults who: (1) are healthy enough to undergo treatment if colorectal cancer is detected; (2) do not have comorbid conditions that would significantly limit their life expectancy [6, 7].

Identifying patient populations who appear to be at increased risk of CRC and for whom further investigation is especially encouraged is a significant priority for managing the overall CRC burden in a defined population. Risk assessment tools are commonly used to estimate a patient's relative risk of disease on the basis of well-established biological, behavioral, and/or demographic factors. Those persons in the highest risk categories are expected to be particularly motivated to comply with disease screening recommendations [9]. Analysis of common clinical parameters (i.e., demographic data and common laboratory tests, such as blood counts) could enable more efficient screening of large populations.

Previous reports have shown that analysis of complete blood counts (CBCs) can help identify patients at high risk of CRC [10]. Unexplained anemia is a major predictor of CRC in the elderly [11] and, together with hemorrhoids, is the most common cause for delay in CRC diagnosis [11–14]. Blood loss is present in 60% of CRC cases, and a daily loss of as little as 3 mL in the stool can cause iron deficiency anemia [15]. However, as only 18% of CRC cases had anemia more than a year before diagnosis [16], a significant proportion of the population is not anemic [17]. The fecal occult blood test detects only current bleeding, while in CRC, blood loss is commonly intermittent [10]. It seems logical that tests designed to detect intermittent blood loss should improve the sensitivity of screening for colonic malignancies. It has been reported that 88% of CRC patients had at least one blood abnormality [10]. As such, attempts to predict CRC from the CBC are under active research. Previous publications have shown that red blood cell distribution width (RDW) had 84% sensitivity and 88% specificity for right-sided CRC cases; no improved sensitivity in combination with red cell distribution width (RDW), hemoglobin (Hgb), and mean corpuscular volume (MCV) was documented [10]. Goldshtein et al. [18] have shown that a minor decrement in the levels of blood Hgb may signify the early development of CRC. Recognition of a change in Hgb levels over time, rather than the most current value alone, has been shown to improve detection of CRC [19].

Kinar et al. [20, 21] developed a novel method for identifying individuals at increased risk of having CRC through empirically derived detection models of their blood counts, age, and sex. Based on machine learning methods—decision trees and cross-validation techniques—this method, called ColonFlag®, enabled generation and evaluation of data-driven detection models [20, 21]. ColonFlag® was developed using data from Maccabi Healthcare Services (MHS) and the Israel National Cancer Register (INCR) [22]. This statistical detection model was validated using the UK's The Health Information Network (THIN) database—an anonymized UK primary care database derived from General Practitioners in the UK, which is broadly representative of the UK population in terms of sex, age, and major condition prevalence [23]. Individuals in the highest one percentile of ColonFlag® scores faced a 20-fold higher risk of being diagnosed with CRC in the subsequent 12–18 month period [20, 21]. This performance of ColonFlag® on Israeli and UK patient populations suggests that it should be estimated for other defined populations.

The primary aims of this study are to develop and evaluate the performance of the ColonFlag® on a US insured population, and assess the applicability of the score in different subgroups and various scenarios of use and in comparison with clinical indicators that warrant referral to colonoscopy due to a higher likelihood of CRC.

Methods

Ethics Approval

This research was approved by the Kaiser Permanente Northwest Region (KPNW) Human Subjects Protection Program (institutional review board), which granted waivers of informed consent because this study involved analyses of retrospective data where all patient information was anonymized and de-identified prior to transfer to Medial Early Sign (MES) for analysis. The KPNW IRB and KPNW HIPAA compliance officer approved this file transfer process and sharing of these data. Unique study-specific patient numbers were included on these files: links to patient names, medical record numbers, and other identifying information were not disclosed.

Setting

KPNW is a prepaid integrated healthcare system with an electronic medical record system and represents a logical test bed for this statistical detection modeling effort. KPNW patients are covered for all medical care ordered or referred by their KPNW physicians. Patients are not covered for services sought on their own initiative from non-plan sources, with the exception of true emergency care. This economic incentive assures high capture of comprehensive healthcare utilization and laboratory test result data on KPNW members. This setting enabled identifying all KPNW patients diagnosed with CRC as well as all KPNW patients who had no evidence of CRC during their effective enrollment periods. KPNW medical information systems also included tumor registry data on nearly all diagnosed cancers among KPNW members, as well as laboratory test results for all members. The KPNW Tumor Registry coordinates with the Oregon and Washington State Tumor Registries to identify KPNW members who received their cancer diagnoses outside of the KPNW system.

Disease Detection Modeling Paradigm

The overall purpose of this CRC detection model is to detect adults who are likely at an elevated risk of having CRC based on their demographics, and previous laboratory test results in a population of adults with comprehensive medical history data available. From this empirical healthcare data resource, we extracted equivalent data on hypothesized CRC predictors for large samples of patients with and without CRC. We employed statistical detection modeling techniques to derive models that predicted the likelihood of individuals having or developing CRC. Our purpose was to identify an enriched sample of patients for whom having a colonoscopy was a high clinical priority. The model's parameters could be adjusted to achieve desired combinations of true positives, true negatives, false positives, and false negatives, given the available resources to recruit targeted patients, perform their colonoscopies, and treat identified precancerous lesions and cancers.

Study Population Selection and Matching

The colorectal cancer cases were selected from the KP Tumor Registry using the following selection criteria: (1) diagnosed with colorectal cancer—International Classification of Diseases-Oncology (ICD-O) sites C18.0–C18.9, C19.9, and C20.9; (2) had one or more CBCs within 6 months of the CRC diagnosis date; (3) had at least 180 days of continuous KPNW enrollment prior to CRC diagnosis date (enrollment gaps of 90 days or less were considered continuous enrollment); (4) CRC patients with any cancer diagnosis prior to the CRC diagnosis date were excluded; and (5) CRC patients with other cancers diagnosed on the same date as the CRC diagnosis date were flagged so that this variable was available to the detection modeling effort.

Control cases were selected from the KPNW membership using the following criteria: (1) received at least one outpatient CBC between 2000 and 2013; (2) age between age 40 and 89 years at time of at least one CBC; (3) no history of cancer diagnoses in the KPNW Tumor Registry or electronic medical record systems; (4) were continuously enrolled in KPNW from 180 days prior to CBC date through 24 months after the CBC date (30 months of cancer-free continuous enrollment, with gaps of up to 3 months patched); (5) because potential controls could have more than one CBC in their study eligibility period, one CBC for each control case was randomly selected to assign a pseudo-diagnosis date for purposes of matching to CRC cases; (6) for each calendar year, 18 control cases were randomly selected for each CRC case diagnosed in that calendar year, matching on the general enrollment population's 10-year age groups (up to 80–89 years) and lengths of continuous enrollment (0.5–5 years, more than 5 years up to 10 years; and more than 10 years prior to diagnosis or pseudo-diagnosis date); (7) controls for 2013 cases were selected from 2012 potential controls and matched on the 2012 general population's distribution of 10-year age groups and length of continuous enrollment; and (8) random matching was repeated until 18 controls per case were identified or three iterations were completed. A random sample of 900 KPNW adults with CRC (and having at least one prior CBC) who were at least 40 years of age at the time of disease onset, and a random sample of 16,195 healthy KPNW controls were created.

Data Needs for Disease Detection Modeling

For the calculating the CRC detection score, the model requires, at minimum, gender, year of birth, and at least one CBC, which includes at least one of the following combinations of findings: {RBC, Hgb, Hct}, {RBC, Hct, MCH}, {RBC, MCH, MCHC}, {Hgb, Hct, MCH}, {Hgb, MCH, MCHC}, or {Hct, MCH, MCHC}. When the minimum required information is not available, no score is produced, and an error message is returned for the specific patient's record. If available, multiple CBCs for each patient can be put into the model, and the algorithm will compute an optimized likelihood of CRC. The ColonFlag® algorithm performs the following main functions: (1) batch processing of patient data input files; (2) validation of input files for valid data structure, logic, and conformity with model requirements; (3) calculation of a predictive score for each patient; and (4) creation of an output file containing a CRC risk score for each patient.

Data Extraction

Data were extracted on all colonoscopy procedures performed on cases from 2000 through the CRC diagnosis date, and on all controls through 2013. Note that KPNW members are not reimbursed for laboratory tests not prescribed by a KPNW physician. Colonoscopies with tissue removal were linked to the respective pathology reports. All CBC and serum ferritin results from 1998 through 2013 for cases and controls were extracted. Data were also extracted on patient demographics, deaths, all inpatient and outpatient diagnoses, tumors, colonoscopies performed, flexible/rigid sigmoidoscopies performed, FOBT and FIT test results, enrollment history, hospitalizations, body mass index, and tobacco consumption.

Data Transfer

Limited data files with study-specific case identifying numbers were created by content areas and unit of observation for all cases and controls. These files were transferred by KPNW to Medial EarlySign, Inc. (MES) via secure encrypted Web transfer.

Data Quality Check

Manual re-abstraction of tumor and medical record data was conducted for stratified random samples of 10 study cases and 10 control cases each for selected patient characteristics—demographics, laboratory results, cancer registry, procedures, and diagnosis information. MES identified cases and controls with a screening colonoscopy in 2006 or later at age 50 or older from their version of the study data files. Screening colonoscopies had to have the reason for referral as a family history of CRC or a patient-requested colonoscopy. Twice the numbers needed were sampled in order to allow replacements for colonoscopies that were diagnostic instead of screening procedures. Eligibility for the subgroups was based on red-blood-cells-related parameters and ferritin results and if a pathology report existed for the colonoscopy. The order of priority for subgroup assignment was: First, microcytic anemia—30 subjects with $MCV < 82$ fL and $RDW > 15$ and $Hgb < 11$ for women and <12 for men; second, low ferritin—30 subjects ≤ 20 ng/mL; third, low Hgb—30 subjects <11 for women and <12 for men; and last, no findings—10 subjects where no biopsy was taken and no pathology report existed. An MES investigator conducted an in-person blinded re-abstraction of the medical record data for these 100 cases with a KPNW medical record technician. The MES investigator read the research case number to the KPNW medical record abstractor, who, in turn, read the selected variable values from the medical record back to the MES investigator. The result was 100% in agreement on all abstracted variables for all 100 cases between the MES version of the study data files and the original KPNW medical records.

Detection Model Development

MES performed diagnostics on the data supplied by KPNW. Missing data were verified with KPNW. The majority of questions from MES staff required explanations of allowable ranges and acceptable patterns across multiple variables. The entire KPNW sample—both cases and controls—were used for our analysis to test the ColonFlag® detection algorithm.

Results

Size and Demographics of US HMO Study Samples

A total of 17,095 patients were included in this analysis. The CRC sample included 900 patients—439 females and 461 males (Table 1). The CRC-free control sample included 16,195 patients—9108 females and 7087 males. Overall, female CRC patients were 10.8 years older than the female control sample, and male CRC patients were 9.8 years older than the male control sample. The requirement of being cancer-free may account, at least in part, for the younger age distribution of controls.

Table 1 Size and demographics of study sample

Group	Female		Male		Total	
	N	Age (mean, SD)	N	Age (mean, SD)	N	Age (mean, SD)
Colorectal cancer cases	439	68.6, 12.8	461	66.9, 11.5	900	67.7, 12.1
Controls	9108	57.8, 12.1	7087	57.1, 10.9	16,195	57.5, 11.6
Total	9547	58.3, 12.3	7548	57.7, 11.2	17,095	58.0, 11.8

Kaiser Permanente Northwest, Portland, Oregon, USA

Performance of CRC Detection Model

Sensitivity$_{Hgb}$ refers to CRC cases only and is the rate of CRC's identified by low Hgb levels alone out of the total CRC cases, based on available CBC's in two adjacent time windows—0–180, and 181–360 days before the date of CRC diagnosis. Sensitivity$_{ColonFlag}^®$ is the rate of CRCs identified by ColonFlag$^®$ out of the total CRC cases, based on available CBC's in two time windows—0–180 versus 181–360 days before the date of CRC diagnosis. It should be noted that the cutoff was calculated according to the specificity level of the Hgb group. For the 0–180-day window, Sensitivity$_{ColonFlag}^®$ was 34% and 36% higher for the 50–75- and 40–89-year-old age groups' Sensitivity$_{Hgb}$, respectively (Table 2). In the 181–360-day window, Sensitivity$_{ColonFlag}^®$ was 47% and 84% higher than Sensitivity$_{Hgb}$ for the 50–75- and 40–89-year-old CRC age groups, respectively.

Our CRC detection model had an area under the receiver operating characteristics (AUROC) curve of 0.81 for women and 0.79 for men, respectively (Table 3). The model's odds ratios for women were higher than for men at various high specificity levels ranging from 90 to 99%.

The ROC curve for ColonFlag$^®$ applied to KPNW data is shown in Fig. 1 (AUC = 0.81, both genders combined). For comparison, the AUROC curve for a detection algorithm using only age has an AUROC of 0.73). The ROC curve ColonFlag$^®$ applied to the MHS (Israel) data had the best performance (AUROC = 0.87) and applied to the NHS data the second best (AUROC = 0.85).

The predicted relative risks generated by the CRC detection model were 12.1 and 16.7 for in situ and Stage I, respectively, at 99% specificity (Table 4). The predicted relative risks of CRC from the CRC detection model were

54.1 and 57.3 for Stage II and Stage III, respectively, and 40.4 for Stage IV.

Our CRC detection model performed best in detecting CRC tumors in the cecum and ascending colon, and less well detecting tumors in the transverse colon, and worst for detecting tumors in the sigmoid colon and rectum. The odds ratio of the CRC detection model for detecting tumors in the cecum was 93.4 at the 99% specificity level, as compared to an OR of 10.2 for detecting tumors in the rectum (Table 5). At the 95% and 90% specificity levels, the ORs for detecting tumors in the ascending colon were higher than for the cecum—40.3 at 95% and 28.0 at 90%—versus 5.4 and 4.9 for the rectum, respectively.

Odds ratios for detecting CRC declined over longer time intervals after the CBC tests were performed. Odds ratios for detecting CRC in patients aged 40–89 years at the 99% specificity level were 34.7 for the 0–180-day window after the CBC versus 20.4 for the 181–365-day window (Table 6).

Bleeding in the bowel can result from conditions other than cancer. The ORs for selected non-cancerous bowel conditions that can cause internal bleeding are shown in Table 7 by specificity levels. While the ORs were much lower for these conditions compared to CRC, these data reveal that detection models may have applicability for passive screening of defined populations for some of these conditions, such as angiodysplasia/angioectasia.

Discussion

An algorithm-based analysis of medical information that includes a CBC had higher sensitivity for detecting CRC cases compared to Hgb alone within 6 months and

Table 2 Sensitivity of the ColonFlag$^®$ detection model by age group and time window

Time window (days after complete blood count test)	Age group	Sensitivity$_{Hgb}$	Sensitivity$_{ColonFlag}^®$	Improvement (%)
0–180 days	50–75	25.7	34.5	34
	40–89	29.4	39.9	36
181–360 days	50–75	12.8	18.8	47
	40–89	14.9	27.4	84

Kaiser Permanente Northwest, Portland, Oregon, USA

Table 3 Area under the receiver operating characteristics curve and odds ratios for ColonFlag® by gender and specificity levels

Gender	AUROC (95% CI)	Odds ratio by specificity level (95% confidence interval)			
		99%	97%	95%	90%
Female	0.81	39.6	22.0	16.8	11.0
	(0.78–0.83)	(30.3–50.6)	(18.0–26.8)	(13.7–20.2)	(8.9–13.4)
Male	0.79	31.5	17.4	12.8	8.5
	(0.77–0.81)	(24.3–39.5)	(13.8–21.5)	(10.2–15.9)	(6.9–10.5)
Combined	0.80	34.7	19.7	14.6	10.0
	(0.79–0.82)	(28.9–40.4)	(17.0–22.6)	(12.5–16.6)	(8.6–11.5)

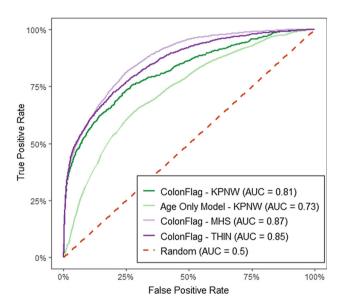Kaiser Permanente Northwest, Portland, Oregon, USA

Fig. 1 Receiver operating characteristics curves, 0–180 days prior to colorectal cancer diagnosis, ages 40–89 years

Table 4 ColonFlag® odds ratios of colorectal cancer by stage for various specificity levels, ages 40–89 years

Colorectal cancer stage	N	Odds ratio by specificity level			
		99%	97%	95%	90%
In situ	24	12.1	6.8	6.2	4.9
I	61	16.7	7.8	6.2	4.6
II	215	54.1	29.1	21.9	15.1
III	185	57.3	29.2	21.1	12.6
IV	120	40.4	26.3	18.9	13.4

Kaiser Permanente Northwest, Portland, Oregon, USA

Table 5 ColonFlag® odds ratios of colorectal cancer by tumor location for various specificity levels, ages 40–89 years

Colorectal cancer location	N	Odds ratio by specificity level			
		99%	97%	95%	90%
Cecum	181	93.4	60.5	40.3	28.0
Ascending	107	90.0	56.4	51.1	32.9
Transverse	44	51.1	32.3	20.8	15.8
Sigmoid	171	13.8	7.2	6.8	4.9
Rectum	182	10.2	6.1	5.4	4.9

Kaiser Permanente Northwest, Portland, Oregon, USA

the first US-based study of the ColonFlag® early CRC detection model. Previous validations have been performed on members of the MHS in Israel and on a British National Health Services population [20, 21]. Performance of the ColonFlag® CRC detection model with the KPNW validation data is similar to these previous foreign studies; the algorithm-based analysis performed best in detecting CRC tumors in the cecum and ascending colon. Furthermore, we demonstrated the model's significant advantage over a model based on age only.

The overall compliance rate of CRC screening in the USA is still considered suboptimal [19, 24–26]. About one-third of eligible adults in the USA have never been screened for CRC [27]. Offering choice in CRC screening strategies may increase screening uptake [28]. In the USA, CRC screening is promoted through the dissemination of guidelines and media campaigns, although some organized programs are run through health plans and local health departments [25]. CRC screening rates of adults aged 50–75 years reported by the CDC's Behavioral Risk Factor Surveillance System in 2010 have reached 60% [26]. The National Colorectal Cancer Roundtable is a coalition of organizations—healthcare systems, government agencies, health insurers, universities, medical schools, scientific organizations, professional health organizations, health care providers, individuals, etc.—that have pledged to

6–12 months after the CBC tests. The algorithm-based analysis had higher sensitivity for identifying CRC cases diagnosed in the first 6 months, as compared to 6–12 months before CRC diagnosis, and for detecting CRC cases among the 40–89-year-old CRC population age range compared to the 50–75-year-old CRC population. This is

Table 6 ColonFlag® odds ratios for colorectal cancer by age group and time window

Time window (days after complete blood count test)	Age group (years)	Odds ratio by specificity level			
		99%	97%	95%	90%
0–180 days	40–89	34.7	19.7	14.6	10.0
	50–75	29.6	15.7	10.8	7.4
181–360 days	40–89	20.4	11.5	8.7	7.8
	50–75	14.7	7.1	5.3	4.8

Kaiser Permanente Northwest, Portland, Oregon, USA

Table 7 ColonFlag® odds ratios for selected non-cancerous bowel conditions that can cause internal bleeding by specificity levels

Condition	N	Odds ratio by specificity level		
		97%	95%	90%
Diverticulosis/diverticular disease	928	4.7	3.6	2.9
Angiodysplasia/angioectasia	13	9.7	8.4	7.7
Inflammatory bowel disease (colitis, ulcerative colitis, Crohn's disease)	40	4.6	2.1	1.6
Gastrointestinal ulcers (rectal, gastric, duodenal)	127	6.0	4.2	4.1
Upper gastrointestinal inflammations (esophagitis, duodenitis, gastritis)	1791	1.9	1.7	1.5
Hemorrhoids	1234	2.0	1.6	1.5

Kaiser Permanente Northwest, Portland, Oregon, USA

cooperate in raising the rate of CRC screening in the USA to 80% of the at-risk population by 2018 [27]. Amidst increased screening rates is evidence of screening and surveillance colonoscopy overuse, programs that target patients at increased risk for CRC may help to better target colonoscopy resources [28–32].

The lower performance of our CRC detection model for detecting tumors in the sigmoid colon and rectum (Table 5) may relate to the ability for persons to visualize fresh blood from these left-sided tumors through hematochezia; this symptom often leads to a clinical presentation and subsequent diagnosis of an underlying CRC. Older CBC tests still have meaningful predictive value for CRC (Table 6), but an analysis of CBC test results in <180 days has higher predictive accuracy and enables earlier detection of potentially treatable disease. Ideally, the ColonFlag® CRC detection model can be computed after every CBC test and incorporated into the reports to ordering physicians.

Prior "Big Data" algorithms utilizing patient data have had limitations. Hippisley-Cox et al. [33] recently developed a range of innovative algorithms for identifying individuals suspected to CRC by analyzing primary care data. These algorithms identify suspected individuals by taking into account "alarm" symptoms which may indicate the existence of as yet undiagnosed cancer. As part of their studies, Hippisley-Cox et al. [33] developed and validated algorithms for detecting individuals at high risk of current CRC. These algorithms make use of symptoms—recorded within primary care consultations—which are known to

indicate the existence of CRC (such as rectal bleeding, weight loss, anemia, and other symptoms). Although these symptoms may also be associated with other types of cancer, the algorithms are able to use these general parameters to specifically identify individuals with high chances to be diagnosed with CRC within a period of 2 years. The reported receiver operating characteristics (ROC) curve statistics for these algorithms were 0.89 (females) and 0.91 (males). The top 10% risk score of the validated population had 90.1% specificity and 70.6% sensitivity for diagnosing CRC in the following 2 years. Yet, the algorithms presented by the Hippisley-Cox team have several limitations. They use parameters based on self-reported symptoms, which may not always be collected or reliably reported by the patients. Moreover, models based on patient complaints or visible clinical signs of cancer are unable to identify the cancer at an early stage before there are any visible alarming signs.

Our CRC detection model algorithm reliably identifies individuals in curable stages of CRC (0/I/II), and flags CRC tumors 180–360 days prior to a CRC diagnosis. ColonFlag® performs better than single Hgb threshold screening. Our detection model also demonstrates useful detection performance for other clinical conditions that generate increases or decreases in Hgb values. Other currently available risk scores for CRC utilizing age, sex, body mass index (BMI) as well as medical history, diet, exercise, and other predictive factors have been shown to either have poor discriminatory power [33, 34], require collection of
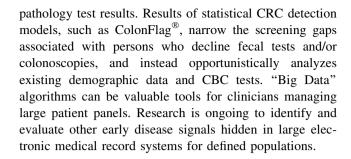
patient-reported information, or focus on the estimation of individual lifetime risk of CRC, which is quite different from current risk [34]. Work is beginning on re-estimating our CRC detection model using US HMO data (KPNW). We expect this tailoring will improve the model's detection performance for the KPNW membership.

An efficient CRC screening program has a high compliance rate, but also targets patients at increased risk for CRC. The reported compliance rate varies tremendously between CRC screening programs worldwide (10–71%), depending on socioeconomic status, ethnicity, age, gender, psychological factors, and other factors [21]. Whereas newer CRC screening programs based on mailed fecal immunochemical tests and screening colonoscopy can reach a majority of patients in some settings [35, 36], there is concern that fecal immunochemical tests may be less sensitive than colonoscopy for right-sided colorectal cancers [37]. Colonoscopy resources are also limited [38], and there is evidence of overuse of screening and surveillance colonoscopy in the USA [31, 32], which may reduce access for others with higher risk of CRC. Our CRC detection model can be applied to broad populations to identify persons at increased risk of CRC (in particular, right-sided CRC); this can enable organized health systems to more effectively target colonoscopy resources.

Strengths of this study include innovative use of electronic medical record data, a large number of CRC cases, a large control sample, and a sophisticated machine learning detection algorithm. A policy-relevant limitation of the ColonFlag® CRC detection model algorithm is that it cannot characterize the risk of individuals who avoid contact with the health care system. We suggest additional research on identifying characteristics predictive of undiagnosed cancer risks among non-users, such as age, gender, last BMI, and length of time since last physician visit.

Conclusions

The ColonFlag® model had higher sensitivity for detecting CRC cases among true CRC cases compared to Hgb alone in the first and second 6 months after the CBC tests. It also had higher sensitivity for identifying CRC cases diagnosed the first 180 days as compared to 181–360 days before CRC diagnosis, and for detecting CRC cases among the 40–89-year-old CRC population age range compared to the 50–75-year-old CRC population. ColonFlag® has been integrated into a population-based CRC screening program by MHS in Israel [20]. This study similarly demonstrates its feasibility for its use in a US-based HMO adult population with a comprehensive electronic medical record systems that includes a NAACCR-certified tumor registry, clinical diagnosis and procedure codes, and laboratory and pathology test results. Results of statistical CRC detection models, such as ColonFlag®, narrow the screening gaps associated with persons who decline fecal tests and/or colonoscopies, and instead opportunistically analyzes existing demographic data and CBC tests. "Big Data" algorithms can be valuable tools for clinicians managing large patient panels. Research is ongoing to identify and evaluate other early disease signals hidden in large electronic medical record systems for defined populations.

Compliance with ethical standards

Conflict of interest ColonFlag® (previously MeScore®) is a registered product with a granted patent. This manuscript does not cover any product under development. RG, EC, and YK are employees of Medial EarlySign, Inc.; this employment does not alter the authors' adherence to the journal's policies on sharing data and materials. All other authors declare that they have no conflicts of interest.

References

1. American Cancer Society. *Cancer Facts and Figures 2016*. Washington DC. http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079. Last accessed October 14, 2016.
2. Qaseem A, Denberg TD, Hopkins RH Jr, et al. Screening for colorectal cancer: a guidance statement from the American College of Physicians. *Ann Intern Med*. 2012;156:378–386.
3. Segnan NPJ, von Karsa L (eds). *European Guidelines for Quality Assurance in Colorectal Cancer Screening and Diagnosis*. 1st ed. Brussels: European Union; 2010.
4. Committee ACNCCGR. *Clinical Practice Guidelines for the Prevention, Early Detection and Management of Colorectal Cancer*. 2005; Available from: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/cp106_0.pdf.
5. 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 U.S. Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology*. 2008;134:1570–1595.

6. Lin JS, Piper MA, Perdue LA, et al. Screening for colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA*. 2016;315:2576–2594. doi:10.1001/jama.2016.3332.

7. U.S. Preventive Services Task Force, Bibbins-Domingo K, Grossman DC, et al. Screening for colorectal cancer: U.S. Preventive Services Task Force Recommendation Statement. *JAMA*. 2016;315:2564–2575. doi:10.1001/jama.2016.5989.

8. Doubeni CA, Corley DA, Quinn VP, et al. Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: a large community-based study. *Gut*. 2016. doi:10.1136/gutjnl-2016-312712 (Epub ahead of print).

9. Driver JA, Gaziano JM, Gelber RP, Lee IM, Buring JE, Kurth T. Development of a risk score for colorectal cancer in men. *The American Journal of Medicine*. 2007;120:257–263.

10. Spell DW, Jones DV Jr, Harper WF, David Bessman J. The value of a complete blood count in predicting cancer of the colon. *Cancer Detect Prev*. 2004;28:37–42.

11. Bafandeh Y, Khoshbaten M, Eftekhar Sadat AT, Farhang S. Clinical predictors of colorectal polyps and carcinoma in a low prevalence region: results of a colonoscopy based study. *World J Gastroenterol*. 2008;14:1534–1538.

12. Dominguez-Ayala M, Diez-Vallejo J, Comas-Fuentes A. Missed opportunities in early diagnosis of symptomatic colorectal cancer. *Revista Espanola de Enfermedades Digestivas: Organo Oficial de la Sociedad Espanola de Patologia Digestiva*. 2012;104:343–349.

13. Goodman D, Irvin TT. Delay in the diagnosis and prognosis of carcinoma of the right colon. *The British Journal of Surgery*. 1993;80:1327–1329.

14. Hafstrom L, Johansson H, Ahlberg J. Does diagnostic delay of colorectal cancer result in malpractice claims? A retrospective analysis of the Swedish board of malpractice from 1995–2008. *Patient Safety in Surgery*. 2012;6:13.

15. Barillari P, de Angelis R, Valabrega S, et al. Relationship of symptom duration and survival in patients with colorectal carcinoma. *Eur J Surg Oncol*. 1989;15:441–445.

16. Acher PL, Al-Mishlab T, Rahman M, Bates T. Iron-deficiency anaemia and delay in the diagnosis of colorectal cancer. *Colorectal Disease: The Official Journal of the Association of Coloproctology of Great Britain and Ireland*. 2003;5:145–148.

17. Rai S, Hemingway D. Iron deficiency anaemia–useful diagnostic tool for right sided colon cancers? *Colorectal Disease: The Official Journal of the Association of Coloproctology of Great Britain and Ireland*. 2005;7:588–590.

18. Goldshtein I, Neeman U, Chodick G, Shalev V. Variations in hemoglobin before colorectal cancer diagnosis. *Eur J Cancer Prev*. 2010;19:342–344.

19. Cancer Screening—United States, 2010. *MMWR*. 2012; 61:41–45.

20. Kinar Y, Kalkstein N, Akiva P, et al. 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*. 2016;23:879–890. doi:10.1093/jamia/ocv195.

21. Kinar Y, Akiva P, Choman E, et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. *PLoS ONE*. 2017;12:e0171759. doi:10.1371/journal.pone.0171759.

22. Israel National Cancer Registry. Available from: http://www.health.gov.il/icr.

23. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of the health improvement network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Informatics in Primary Care*. 2011;19:251–255.

24. Davis M, Oaten M, Occhipinti S, Chambers SK, Stevenson RJ. An investigation of the emotion of disgust as an affective barrier to intention to screen for colorectal cancer. *Eur J Cancer Care (Engl)*. 2016. doi:10.1111/ecc.12582 (Epub ahead of print).

25. Power E, Miles A, von Wagner C, Robb K, Wardle J. Uptake of colorectal cancer screening: system, provider and individual factors and strategies to improve participation. *Future Oncology*. 2009;5:1371–1388.

26. Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2010.

27. http://nccrt.org/roundtable-members/member-organizations/. Last accessed April 26, 2017.

28. Shahidi NC, Homayoon B, Cheung WY. Factors associated with suboptimal colorectal cancer screening in U.S. Immigrants. *Am J Clin Oncol*. 2013;36:381–387.

29. Shapiro JA, Klabunde CN, Thompson TD, Nadel MR, Seeff LC, White A. Patterns of colorectal cancer test use, including CT colonography, in the 2010 National Health Interview Survey. *Cancer Epidemiol Biomark Prev*. 2012;21:895–904.

30. Inadomi JM, Vijan S, Janz NK, et al. Adherence to colorectal cancer screening: a randomized clinical trial of competing strategies. *Arch Intern Med*. 2012;172:575–582.

31. Goodwin JS, Singh A, Reddy N, Riall TS, Kuo YF. Overuse of screening colonoscopy in the Medicare population. *Arch Intern Med*. 2011;171:1335–1343. doi:10.1001/archinternmed.2011.212 (Epub 05/09/2011). PubMed PMID: 21555653; PubMed Central PMCID: PMC3856662.

32. Kruse GR, Khan SM, Zaslavsky AM, Ayanian JZ, Sequist TD. Overuse of colonoscopy for colorectal cancer screening and surveillance. *J Gen Intern Med*. 2015;30:277–283. doi:10.1007/s11606-014-3015-6 (Epub 08/30/2014). PubMed PMID: 25266407; PubMed Central PMCID: PMC4351286.

33. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *The British Journal of General Practice*. 2012;62:e29–e37.

34. Baxter NN, Goldwasser MA, Paszat LF, Saskin R, Urbach DR, Rabeneck L. Association of colonoscopy and death from colorectal cancer. *Ann Intern Med*. 2009;150:1–8.

35. van der Vlugt M, Grobbee EJ, Bossuyt PM, et al. Adherence to colorectal cancer screening: four rounds of faecal immunochemical test-based screening. *Br J Cancer*. 2017;116:44–49. doi:10.1038/bjc.2016.399. PubMed PMID: 27923037; PubMed Central PMCID: PMC5220157.

36. Liles EG, Schneider JL, Feldstein AC, et al. Implementation challenges and successes of a population-based colorectal cancer screening program: a qualitative study of stakeholder perspectives. *Implement Sci*. 2015;10:41–57.

37. Castro I, Estevez P, Cubiella J, et al. Diagnostic performance of fecal immunochemical test and sigmoidoscopy for advanced right-sided colorectal neoplasms. *Dig Dis Sci*. 2015;60:1424–1432. doi:10.1007/s10620-014-3434-6. PubMed PMID: 25407805.

38. Joseph DA, Meester RG, Zauber AG, et al. Colorectal cancer screening: estimated future colonoscopy need and current volume and capacity. *Cancer*. 2016;122:2479–2486.