Scandinavian Journal of Primary Health Care

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ipri20

A machine learning tool for identifying metastatic colorectal cancer in primary care

Eliya Abedi, Marcela Ewing, Elinor Nemlander, Jan Hasselström, Annika Sjövall, Axel C. Carlsson & Andreas Rosenblad

To cite this article: Eliya Abedi, Marcela Ewing, Elinor Nemlander, Jan Hasselström, Annika Sjövall, Axel C. Carlsson & Andreas Rosenblad (13 Mar 2025): A machine learning tool for identifying metastatic colorectal cancer in primary care, Scandinavian Journal of Primary Health Care, DOI: 10.1080/02813432.2025.2477155

To link to this article: https://doi.org/10.1080/02813432.2025.2477155

Published online: 13 Mar 2025.

SCANDINAVIAN JOURNAL OF PRIMARY HEALTH CARE
https://doi.org/10.1080/02813432.2025.2477155

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

A machine learning tool for identifying metastatic colorectal cancer in primary care

Eliya Abedi[a,b,c], Marcela Ewing[d], Elinor Nemlander[a,b,c], Jan Hasselström[a,b], Annika Sjövall[c,e], Axel C. Carlsson[a,b] and Andreas Rosenblad[a,c,f,g]

[a]Division of Family Medicine and Primary Care, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Solna, Sweden; [b]Academic Primary Health Care Centre, Region Stockholm, Stockholm, Sweden; [c]Regional Cancer Centre Stockholm-Gotland, Region Stockholm, Stockholm, Sweden; [d]Department of Community Medicine and Public Health, Sahlgrenska Academy, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden; [e]Division of Coloproctology, Department of Pelvic Cancer, Karolinska University Hospital, Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; [f]Department of Statistics, Uppsala University, Uppsala, Sweden; [g]Division of Clinical Diabetology and Metabolism, Department of Medical Sciences, Uppsala University, Uppsala, Sweden

ABSTRACT

Background: Detection of colorectal cancer (CRC) is mainly achieved by clinical assessment. As new treatments become available for metastatic CRC (MCRC), it is important to accurately identify these patients.

Aim: To develop a predictive model for identifying MCRC in primary health care patients using diagnostic data analysed with machine learning.

Design and setting: A case-control study utilising data on primary health care visits for 146 patients >18 years old diagnosed with MCRC in the Västra Götaland Region, Sweden during 2011, and 577 sex-, age, and primary health care centre-matched controls.

Method: Stochastic gradient boosting was used to construct a model for predicting the presence of MCRC based on diagnostic codes from primary health care consultations during the year before index (diagnosis) date and number of consultations. Variable importance was estimated using the normalised relative influence (NRI) score. Risks of having MCRC were calculated using odds ratios of marginal effects (OR_{ME}).

Results: The optimal model included 76 variables with non-zero influence, had an area under the curve of 76.5%, a sensitivity of 77.8%, and a specificity of 69.2%. The 10 most important variables had a combined NRI of 61.0%. Number of consultations during the year before index date had the highest NRI at 19.2%, with an OR_{ME} of 3.3.

Conclusion: A machine learning method based on primary health care consultation frequency and diagnoses may be used to identify important variables for predicting presence of MCRC. Both primary health care consultations and associated diagnostic codes need to be taken into consideration.

ARTICLE HISTORY

Received 4 September 2024
Accepted 3 March 2025

KEYWORDS

Artificial intelligence; cancer detection; family practice; gradient boosting; colorectal neoplasms

MICRO ABSTRACT

A study in Sweden developed a machine learning model to predict metastatic colorectal cancer (MCRC) in primary health care patients. The model used diagnostic codes and consultation data, achieving a sensitivity of 77.8% and specificity of 69.2%. The number of consultations with a general practitioner emerged as the most important variable. Consideration of both consultation frequency and diagnostic codes is crucial for MCRC detection.

CLINICAL PRACTICE POINTS

Colorectal cancer (CRC) is a significant global health concern, with higher mortality rates for patients with metastatic CRC (MCRC) compared to non-metastatic CRC (NMCRC). While screening programs help detect CRC at earlier stages, clinical assessment remains the primary method for diagnosis. This study aimed to develop a predictive model using machine learning to identify MCRC in primary healthcare patients. The model analysed diagnostic data and consultation frequency from electronic medical records. The machine learning model achieved a sensitivity of 77.8% and specificity of 69.2% in predicting MCRC. Key variables influencing MCRC prediction included the number of consultations, abdominal and pelvic pain, other anaemias, and senile

CONTACT Axel C. Carlsson axel.carlsson@ki.se Division of Family Medicine and Primary Care, NVS Department, Karolinska Institutet, Alfred Nobels Allé 23, SE-141 83, Huddinge, Sweden

cataract. Interestingly, hypertension, dorsalgia, and other joint disorders indicated a decreased risk of MCRC. The study highlights the potential of machine learning-based predictive models to aid in early detection and risk assessment of MCRC in primary health care. Successful implementation of such models could lead to improved clinical decision-making and contribute to better patient outcomes in the foreseeable future, but first further development, external validation, and testing to confirm model accuracy is needed.

POINT OF INTEREST

- We used artificial intelligence to predict the presence of metastatic colorectal cancer, in a case-control study based on all diagnostic codes in primary health care.
- Both number of consultations and associated diagnostic codes need to be considered.
- The prediction tool had a sensitivity of 77.8% and a specificity of 69.2%.

Introduction

Globally, colorectal cancer (CRC) is the third most common cancer, and the second most common cause of cancer-related death [1]. Although the 5-year-relative survival rate for patients with non-metastatic CRC (NMCRC; stage I–III) is high, for patients with metastatic CRC (MCRC; stage IV), the relative survival rate is 10–20% globally [2–4] and 15% in Sweden [5]. With more advanced treatments, the Swedish 5-year survival rates for CRC increased from 75% in 2007 to 79% in 2021. With new and more specific treatments for MCRC, the survival rates will likely increase further in the future [6,7].

Implementation of a synchronised CRC screening programme had in 2022 started in all 21 Swedish regions. Several studies have shown that screening leads to detection of cancer in earlier stages, which in the long term may lead to fewer patients being diagnosed with MCRC [8,9]. However, detection of CRC is still mainly achieved by clinical assessment of symptoms and signs, even in areas where screening is fully implemented [10,11]. Moreover, patients with CRC detected by clinical assessment often have a more advanced stage at diagnosis compared to patients detected by screening [10,12].

Most patients subsequently diagnosed with CRC initially present with symptoms and signs in primary health care [13]. Several risk prediction tools have been developed both in primary and secondary care to aid in identifying patients with increased risk of CRC [14–18]. Whether these risk prediction tools are better at detecting CRC than clinical assessment, is still unclear [19–22]. There are no risk prediction tools designed to find MCRC specifically, and there is little evidence on how symptoms and signs differ between MCRC and NMCRC.

In a previous study, we utilised artificial intelligence to analyse symptom presentations in patients with NMCRC and develop a predictive model. The analysis identified 16 variables with a normalised relative influence (NRI) > 1%. The model demonstrated a sensitivity of 73.3% and a specificity of 83.5%. The five variables with the highest NRI were *iron deficiency anaemia, other diseases of the anus and rectum, abdominal and pelvic pain, other anaemias*, and *haemorrhoids or other perianal venous thrombosis* [23]. The present study intends to develop a predictive model for patients with MCRC using the same method. An underlying question is whether symptoms and signs for patients with MCRC differ from those observed among patients with NMCRC.

Aim

To develop a predictive model for identifying MCRC among patients in primary health care using diagnostic data analysed with machine learning.

Methods and material

Design and setting

The present study utilised data from a population-based matched case-control study of patients diagnosed with breast, CRC, gynaecological, lung, prostate, or skin cancers in the Västra Götaland Region (VGR), Sweden during 2011 [23]. Inclusion criteria for this study were being > 18 years old and alive at the time of diagnosis, not being diagnosed with cancer during the 20 years before the diagnosis (index) date and having consulted a general practitioner in VGR during the year before the index date. Each patient with cancer (case) was matched on age, sex, and primary health care centre with up to four cancer-free controls selected using the same criteria. Cases and controls were identified through the Swedish Cancer Registry and the regional VEGA administrative healthcare database, which contains data on all healthcare use from primary health care providers in VGR. Cases without matched controls were excluded from the study. For the present analysis, only the 146 patients with MCRC and their 577 matched controls were included.

Variables

The collected data included information about each participant's age, sex, number of general practitioner consultations during the year before index date, and which medical diagnoses had been registered at each consultation. Medical diagnoses were registered in VEGA as ICD-10 or KSH97-P codes, the latter being an abbreviated version of ICD-10 adapted to Swedish primary health care.

In total, > 6000 different ICD-10/KSH97-P codes were reported for cases and controls. After excluding codes occurring in < 1% of the participants, four-character codes were merged to the closest common three-character code, or in some cases the closest block of three-character codes, according to clinical relevance. If a four-character code was deemed to be too different from other codes regarding clinical relevance, it was retained as a four-character code. Finally, codes in the ICD-10 code block D37-D48 (*Neoplasms of uncertain or unknown behaviour*) and diagnoses observed in < 10 participants were excluded. The remaining 178 codes were included as dichotomised variables in the present study, with the value 1 or 0 depending on if the code in question had been registered for the participant during the year before the index date or not. Additionally, a variable giving the number of different codes registered during the year before the index date was included.

Artificial intelligence

Machine learning, a subset of artificial intelligence, enables computer systems to autonomously learn and improve performance based on experience without explicit programming by employing statistical models to analyse data, identify patterns, and make predictions or decisions without human intervention [24]. Stochastic gradient boosting is a leading machine learning technique that improves predictive accuracy for classification and regression tasks by combining multiple base models. Building on the gradient boosting methodology, the technique introduces randomness by randomly sampling the training data at each iteration, reducing the risk of overfitting and enhancing the model's generalisation ability [25].

Statistical analyses

Categorical data are given as frequencies and percentages, n (%), while discrete and continuous data are given as means and standard deviations (SDs). Tests of equality between cases and controls for discrete data were performed using the Wilcoxon rank-sum test with continuity correction. To identify patients with MCRC (target), stochastic gradient boosting applied to classification decision trees was used, with 179 predictors (features): the 178 diagnostic codes and number of consultations during the year before index date.

To train the stochastic gradient boosting model, a training data set was constructed by randomly selecting 75% of the cases ($n = 110$), together with their matched controls ($n = 434$). The remaining 25% of cases ($n = 36$) and their matched controls ($n = 143$) were retained as a test data set to evaluate the performance of the model. The stochastic gradient boosting model was estimated using the R package 'gbm' version 2.1.8 by applying a Bernoulli loss function fitted to 20,000 trees, each having a maximum depth of 5 interactions, at least 10 observations in the terminal nodes of the trees, a shrinkage (learning rate) of 0.001, and a subsampling rate (bag fraction) of 0.5. The optimal number of trees to use for prediction was estimated using class-stratified 10-fold cross-validation.

Sensitivity, specificity, and area under the receiver operator characteristics (ROC) curve (AUC) were used to evaluate the performance of the stochastic gradient boosting model. To obtain the individual probabilities of having MCRC for each patient, the stochastic gradient boosting model with the optimal number of trees was applied to the test data set. These probabilities were used to calculate the value that maximised the sum of sensitivity and specificity, which was then used as cut-off value for classifying participants as having MCRC if the probability was larger than this cut-off value and otherwise as not having MCRC.

Variable importance was estimated using the NRI score. Odds ratios of marginal effects (OR_{ME}) of having MCRC were calculated using the probabilities of having MCRC obtained by integrating out all other variables in the model using the weighted tree traversal method. For diagnostic codes, OR_{ME} is given as a comparison between those with and without the diagnosis in question, while for discrete data it is given as a comparison between those at quartile 3 and those at quartile 1. All statistical analyses were performed using R 4.2.0 (R Foundation for Statistical Computing, Vienna, Austria), with p-values <0.05 considered statistically significant.

Results

Characteristics for the 146 cases and 577 controls included in the present study are given in Table 1. A slight majority of cases ($n = 78$; 53.4%) and controls ($n = 307$; 53.2%) were males, with mean (median; SD)

Table 1. Participant characteristics for the 146 cases and 577 controls included in the study.

Variable	Cases	Controls
Male sex, n (%)	78 (53.4)	307 (53.2)
Age at index date, mean (SD)[a]	65.7 (13.4)	65.6 (13.3)
<50 years, n (%)	21 (14.4)	81 (14.0)
50 to <65 years, n (%)	37 (25.3)	154 (26.7)
65 to <80 years, n (%)	66 (45.2)	264 (45.8)
≥80 years, n (%)	22 (15.1)	78 (13.5)
Number of consultations during the year before index date, mean (SD)[b]	5.8 (6.5)	5.6 (6.4)
Number of diagnoses during the year before index date, mean (SD)[c]	4.9 (3.5)	5.2 (3.7)

Note: [a]Median: 66.7 years for cases, 67.5 years for controls. *p*-value for testing equality of cases and controls using the Wilcoxon rank-sum test with continuity correction: [b]0.177, [c]0.371.
SD, standard deviation.

Table 2. Confusion matrix for predicting metastatic colorectal cancer status among the 179 patients in the test data set, using the optimal stochastic gradient boosting model from the training data set.

	Observed		
Predicted	Not MCRC	MCRC	Total
Not MCRC	99	8	107
MCRC	44	28	72
Total	143	36	179

Notes: Predictions based on 3888 trees. Sensitivity: 0.778, specificity: 0.692.

ages of 65.7 (66.7; 13.4) years for cases and 65.6 (67.5; 13.3) years for controls. The age group 65 to <80 years was the most common for both cases ($n = 66$; 45.2%) and controls ($n = 264$; 45.8%). There were no statistically significant differences between cases and controls regarding the number of consultations or number of diagnostic codes during the year before index date (*p*-values 0.177 and 0.371, respectively).

Predictive ability

Table 2 shows how well the model predicted metastatic colorectal cancer status for the 179 patients in the test dataset. These predictions were made using the best stochastic gradient boosting model, trained with 3,888 trees. Of the 36 patients with MCRC, the model correctly predicted presence of MCRC for 28 patients, resulting in a sensitivity of 77.8%. Likewise, for the 143 patients without MCRC, the model correctly predicted absence of MCRC for 99 patients, resulting in a specificity of 69.2%. The model had an AUC of 76.5% (95% confidence interval 67.5–85.6%) (Figure 1). Using the obtained model, 69.2% of

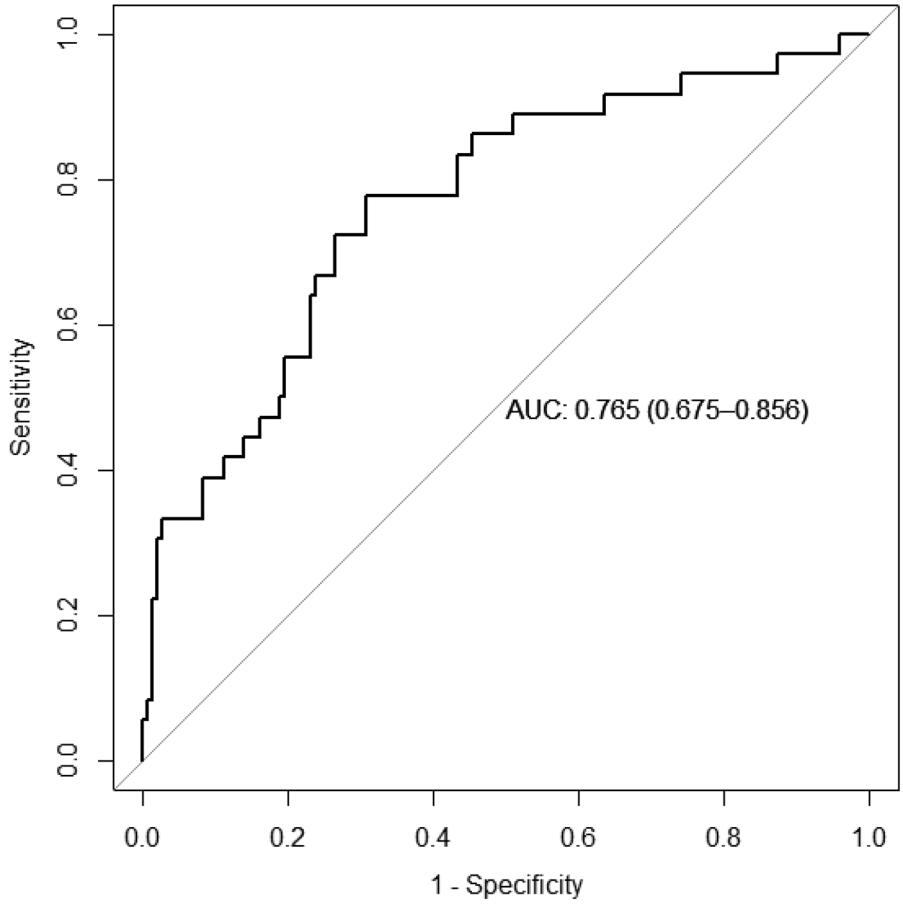

Figure 1. Receiver operator characteristics (ROC) curve and area under the ROC curve (AUC) for the optimal stochastic gradient boosting model from the training data set applied to the 179 patients in the test data set.

cancer-free individuals could thus be correctly excluded as not having MCRC, while 77.8% of individuals with MCRC were correctly diagnosed by the model as having the condition.

Variable importance

Of the 179 predictors included in the stochastic gradient boosting model, 76 (42.5%) had non-zero influence. The 10 variables with the highest NRI for predicting presence of MCRC, together with each variable's OR_{ME} of having MCRC, are given in Table 3. The combined NRI of these 10 variables was 61.0%. Notably, number of consultations during the year before index date had the by far highest NRI at 19.2%. Among the diagnostic codes, *abdominal and pelvic pain* (ICD-10 code R10) had the highest NRI at 12.1%, followed by *other anaemias* (D64) at 11.8%. Of the remaining variables, *senile cataract* (H25) with an NRI of 5.3%, was the only variable with an NRI > 5%.

Marginal effects

Other anaemias had the highest OR_{ME} at 45.8, followed by *senile cataract* at 11.8, *abdominal and pelvic pain* at 10.0, and *other diseases of anus and rectum* (ICD-10 code K62, excluding K625 *Haemorrhage of anus and rectum*) at 3.5. For number of consultations during the year before index date, the OR_{ME} of having MCRC was 3.3, when comparing quartile 3 (7 consultations) with quartile 1 (2 consultations). Notably, the three diagnostic codes *essential (primary) hypertension* (I10), *dorsalgia* (M54), and *other joint disorders, not elsewhere classified* (M25) all had OR_{ME}:*s* < 1. The presence of these diagnoses thus signalled a decreased risk of having MCRC.

Table 3. The 10 variables with highest normalised relative influence (NRI) for predicting presence of metastatic colorectal cancer (MCRC) using the optimal stochastic gradient boosting model with 3888 trees, together with odds ratios for marginal effects (OR_{ME}) of having MCRC.

ICD-10 code	Description	NRI (%)	OR_{ME}
	Number of consultations during the year before index date	19.2	3.3[a]
R10	Abdominal and pelvic pain	12.1	10.0
D64	Other anaemias	11.8	45.8
H25	Senile cataract	5.3	11.8
I10	Essential (primary) hypertension	2.5	0.9
K62[b]	Other diseases of anus and rectum	2.2	3.5
M54	Dorsalgia	2.1	0.9
M25	Other joint disorders, not elsewhere classified	2.1	0.4
R53	Malaise and fatigue	1.9	2.0
M75	Shoulder lesions	1.8	1.9

[a]OR_{ME} for comparing quartile 3 (7 consultations) with quartile 1 (2 consultations).
[b]Excluding K625 *Haemorrhage of anus and rectum*.

Discussion

We developed and tested a predictive model for identifying patients diagnosed with MCRC in primary health care using machine learning on all available diagnostic codes from electronic medical records and number of general practitioner consultations. The model had a sensitivity of 77.8% and a specificity of 69.2%. The sensitivity indicates that the model would miss 22.2% of patients with MCRC. However, given that this diagnosis is often overlooked in primary health care [26], we welcome the potential to identify nearly 80% of these cases. The specificity suggests that the model would incorrectly prompt investigations in approximately 30% of patients without MCRC, which is still significantly lower than the roughly 80% of cancer-free patients currently undergoing fast-track referrals for CRC investigation in Sweden [27]. Given the low incidence of CRC in the general population [28,29], the model is expected to produce more false positives than true positives. This is reflected in the test data set, where there were 44 false positives and 28 true positives at a case-to-control ratio of 1:4. However, false positives are not a main concern, as the primary focus is ensuring that individuals classified as cases have a high risk of MCRC. Furthermore, since these risk prediction models are intended for use in symptomatic patients rather than for screening, prioritising high sensitivity over specificity is warranted. This is because the likelihood of cancer that requires treatment is higher in a symptomatic population in primary health care than in a screening population [30].

Results in perspective

In the model, 178 diagnostic codes were used as predictors together with number of consultations during the year before index date, thereby taking not only number of consultations but also the diagnostic codes registered during the consultations into account while calculating NRI and OR_{ME}. Number of consultations during the year before the index date had the highest NRI (19.2%) for predicting MCRC. The OR_{ME} was 3.3 when comparing quartile 3 (7 visits) with quartile 1 (2 visits). Several studies have shown an increase of patient consultations in primary health care in the months leading to a cancer diagnosis [31–33]. However, in the present study, number of consultations did not differ significantly between cases and controls in the univariate analysis. The high NRI and OR_{ME} observed for number of consultations may therefore be explained by the intricate interactions built into the stochastic gradient boosting model, with up to five-way

interactions being possible between all included predictors. The absence of significant differences between cases and controls in number of consultations combined with high NRI and OR_{ME} for the same variable thus implies that a large number of consultations may only be interpreted as indicating an increased risk of MCRC when observed in connection with specific diagnoses.

As expected, *abdominal and pelvic pain* and *other anaemias*, both common symptoms and signs of CRC [14], had high NRIs, with the latter having the highest OR_{ME} at 45.8. More surprisingly, *senile cataract* had an NRI of 5.3% and an OR_{ME} of 11.8. A possible explanation might be that cataracts and CRC share some underlying risk factors, such as smoking, obesity and diabetes [34–36]. Dehydration (due to diarrhoea) is also a known risk factor for cataracts [34]. The observed association between senile cataract and MCRC has not been reported in previous studies and may thus be attributable to statistical errors due to the small sample size of the present study.

In line with our previous study on NMCRC [23], *essential (primary) hypertension*, *dorsalgia*, and *other joint disorders, not elsewhere classified* had OR_{ME}:*s* < 1, indicating that patients with these diagnoses had decreased risks of MCRC. Our hypothesis is that these diagnoses are associated with conditions that are monitored regularly at primary health care centres, enabling symptoms of CRC to be more easily detected [37–39]. It has also been reported that hypertension and benign musculoskeletal lesions are associated with lower risks of metastatic cancer [39]. Another explanation may be that individuals with hypertension more often have prescriptions for daily low-dose aspirin, inhibiting polyp formation and development of CRC [40].

Cancer risk prediction tools in general have yet to be proven useful in clinical practice [21,22]. A systematic review of machine learning methods compared with standard care for diagnostic predictions in emergency departments showed that machine learning methods outperformed standard care on a variety of clinical presentations and outcomes [41]. In a comparison with an oesophageal-gastric cancer risk prediction tool, artificial intelligence-based methods had higher overall accuracy and ability to identify more patients with cancer [42]. For CRC, the clinical utility of prediction models is restricted by inadequate and flawed validation in relevant clinical settings [22], and external validation and prospective testing is first needed before a tool can prove to be useful. Yet, to the best of our knowledge, no risk prediction tools have previously been developed specifically for MCRC.

When comparing the stochastic gradient boosting model for MCRC with the corresponding model for NMCRC [23] we noted that the MCRC model had four variables with an NRI > 5%: *number of consultations during the year before index date* (19.2%), *abdominal and pelvic pain* (12.1%), *other anaemias* (11.8%), and *senile cataract* (5.3%), whereas the NMCRC model had six variables with an NRI > 5%: *iron deficiency anaemia*, *other diseases of anus and rectum*, *abdominal and pelvic pain*, *other anaemias, haemorrhoids and perianal venous thrombosis*, and *change in bowel habit*. We believe the difference may be due to patients with MCRC having more diffuse symptoms such as anaemia, compared to patients with NMCRC, where a localised colorectal tumour more often gives distinct symptoms related to the gastrointestinal tract [43]. Notably, *abdominal and pelvic pain* and *other anaemias* had high influence for both MCRC and NMCRC, while *number of consultations during the year before index date* had four times higher NRI for MCRC than for NMCRC.

Strengths and limitations

This population-based study used data on all diagnostic codes for symptoms and diseases from all primary health care visits for all inhabitants in the second-largest region in Sweden during an entire year, thus allowing for a high degree of generalisability. The use of prospectively recorded ICD-codes reduced the risks of selection and recall bias. However, previous studies have shown that general practitioners do not code for all symptoms presented in consultations [44,45]. Therefore, the codes might not accurately reflect all presented symptoms. Other limitations related to coding are that chronic diagnoses are not coded annually for all patients in Swedish primary health care [46]. Moreover, the reimbursement system for the VGR primary health care is partly based on the disease burden mirrored by registered diagnoses of the listed patients, which may influence the registration of diagnostic codes. Since only codes registered during the year before index date are included, the effect of chronic diagnoses may not be accurately accounted for. This is unfortunate, since chronic diagnoses may impact cancer risk, detection, and outcome [39,47,48].

The ability to understand why certain symptoms influence the detection of MCRC is hindered by the lack of data on height, weight, smoking status, alcohol use, physical activity, family history, and blood test results. Access to electronical medical records including free text during the consultations would have resulted in more data to analyse, but at the risk of cluttering the analyses with inaccurately reported texts

of limited relevance for the current situation. Finally, the small sample size was a limitation, which might explain why number of consultations did not differ significantly between cases and controls.

Conclusions

The present study showed how a machine learning method based on consultation frequency and diagnostic codes may find variables of importance for prediction of MCRC in primary health care. The resulting MCRC risk prediction tool needs to be tested further, both nationally and internationally. If proven successful, it could be integrated in electronic medical records to automatically inform general practitioners of individuals with elevated risks of MCRC.

Ethical approval

The Regional Ethical Review Board in Gothenburg approved the study protocol (reference number: 252-12).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Regional Cancer Centre Stockholm-Gotland, Region Stockholm [grant number NSV-977323], and the Einar Belvén Foundation.

ORCID

Eliya Abedi http://orcid.org/0000-0002-6177-4136
Marcela Ewing http://orcid.org/0000-0002-6849-0654
Elinor Nemlander http://orcid.org/0000-0001-9589-4681
Jan Hasselström http://orcid.org/0000-0001-9521-2345
Annika Sjövall http://orcid.org/0000-0003-2221-5881
Axel C. Carlsson http://orcid.org/0000-0001-6113-0472
Andreas Rosenblad http://orcid.org/0000-0003-3691-8326

References

[1] Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Transl Oncol. 2021;14(10):101174. doi: 10.1016/j.tranon.2021.101174.
[2] Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet. 2014;383(9927):1490–1502. doi: 10.1016/S0140-6736(13)61649-9.
[3] Li N, Lu B, Luo C, et al. Incidence, mortality, survival, risk factor and screening of colorectal cancer: a comparison among China, Europe, and northern America. Cancer Lett. 2021;522:255–268. doi: 10.1016/j.canlet.2021.09.034.
[4] Gullickson C, Goodman M, Joko-Fru YW, et al. Colorectal cancer survival in sub-Saharan Africa by age, stage at diagnosis and Human Development Index: a population-based registry study. Int J Cancer. 2021;149(8):1553–1563. doi: 10.1002/ijc.33715.
[5] Regionalt Cancercentrum Norr. Koloncancer 2020. Nationell kvalitetsrapport för år 2020 från. Umeå: Svenska Kolorektalcancerregistret; 2021.
[6] André T, Shiu K-K, Kim TW, et al. Pembrolizumab in microsatellite-instability-high advanced colorectal cancer. N Engl J Med. 2020;383(23):2207–2218. doi: 10.1056/NEJMoa2017699.
[7] Cercek A, Lumish M, Sinopoli J, et al. PD-1 blockade in mismatch repair-deficient, locally advanced rectal cancer. N Engl J Med. 2022;386(25):2363–2376. doi: 10.1056/NEJMoa2201445.
[8] Larsen MB, Njor S, Ingeholm P, et al. Effectiveness of colorectal cancer screening in detecting earlier-stage disease-a nationwide cohort study in Denmark. Gastroenterology. 2018;155(1):99–106. doi: 10.1053/j.gastro.2018.03.062.
[9] Cardoso R, Guo F, Heisser T, et al. Colorectal cancer incidence, mortality, and stage distribution in European countries in the colorectal cancer screening era: an international population-based study. Lancet Oncol. 2021;22(7):1002–1013. doi: 10.1016/S1470-2045(21)00199-6.
[10] Moreno CC, Mittal PK, Sullivan PS, et al. Colorectal cancer initial diagnosis: screening colonoscopy, diagnostic colonoscopy, or emergent surgery, and tumor stage and size at initial presentation. Clin Colorectal Cancer. 2016;15(1):67–73. doi: 10.1016/j.clcc.2015.07.004.
[11] Hatch QM, Kniery KR, Johnson EK, et al. Screening or symptoms? How do we detect colorectal cancer in an equal access health care system? J Gastrointest Surg. 2016;20(2):431–438. doi: 10.1007/s11605-015-3042-6.
[12] Juul JS, Andersen B, Laurberg S, et al. Differences in diagnostic activity in general practice and findings for individuals invited to the Danish screening programme for colorectal cancer: a population-based cohort study. Scand J Prim Health Care. 2018;36(3):281–290. doi: 10.1080/02813432.2018.1487378.
[13] Rubin G, Berendsen A, Crawford SM, et al. The expanding role of primary care in cancer control. Lancet Oncol. 2015;16(12):1231–1272. doi: 10.1016/S1470-2045(15)00205-3.
[14] Hamilton W, Round A, Sharp D, et al. Clinical features of colorectal cancer before diagnosis: a population-based case-control study. Br J Cancer. 2005;93(4):399–405. doi: 10.1038/sj.bjc.6602714.
[15] Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract. 2012;62(594):e29–e37. doi: 10.3399/bjgp12X616346.
[16] Fijten GH, Starmans R, Muris JW, et al. Predictive value of signs and symptoms for colorectal cancer in patients with rectal bleeding in general practice. Fam Pract. 1995;12(3):279–286. doi: 10.1093/fampra/12.3.279.
[17] 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. doi: 10.1136/gut.2010.225987.
[18] Ewing M, Naredi P, Zhang C, et al. Identification of patients with non-metastatic colorectal cancer in primary care: a case-control study. Br J Gen Pract. 2016;66(653):e880–e886. doi: 10.3399/bjgp16X687985.
[19] Price S, Spencer A, Medina-Lara A, et al. Availability and use of cancer decision-support tools: a cross-sectional survey of UK primary care. Br J Gen Pract. 2019;69(684):e437–e443. doi: 10.3399/bjgp19X703745.
[20] Williams TGS, Cubiella J, Griffin SJ, et al. Risk prediction models for colorectal cancer in people with symptoms: a systematic review. BMC Gastroenterol. 2016;16(1):63. doi: 10.1186/s12876-016-0475-7.
[21] Medina-Lara A, Grigore B, Lewis R, et al. Cancer diagnostic tools to aid decision-making in primary care: mixed-methods systematic reviews and cost-effectiveness analysis. Health Technol Assess. 2020;24(66):1–332. doi: 10.3310/hta24660.
[22] Grigore B, Lewis R, Peters J, et al. Development, validation and effectiveness of diagnostic prediction tools for colorectal cancer in primary care: a systematic review. BMC Cancer. 2020;20(1):1084. doi: 10.1186/s12885-020-07572-z.
[23] Nemlander E, Ewing M, Abedi E, et al. A machine learning tool for identifying non-metastatic colorectal cancer in primary care. Eur J Cancer. 2023;182:100–106. doi: 10.1016/j.ejca.2023.01.011.
[24] Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593.
[25] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–378. doi: 10.1016/S0167-9473(01)00065-2.
[26] Fernholm R, Pukk Härenstam K, Wachtler C, et al. Diagnostic errors reported in primary healthcare and emergency departments: a retrospective and descriptive cohort study of 4830 reported cases of preventable harm in Sweden. Eur J Gen Pract. 2019;25(3):128–135. doi: 10.1080/13814788.2019.1625886.
[27] Regionala cancercentrum i samverkan. Antal patienter i standardiserade vårdförlopp (SVF) ; [2025; cited 2025 January 23]. Available from: https://cancercentrum.se/samverkan/vara-uppdrag/statistik/svf-statistik/antal-patienter-i-svf/
[28] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020. GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clinicians 2021;71(3):209–249.
[29] Åhlin E. Cancer i siffror 2023. Sweden: Socialstyrelsen; 2023.
[30] Neal RD, Tharmanathan P, France B, et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. Br J Cancer. 2015;112 Suppl 1(Suppl 1):S92–S107. doi: 10.1038/bjc.2015.48.
[31] Ewing M, Naredi P, Nemes S, et al. Increased consultation frequency in primary care, a risk marker for cancer: a case-control study. Scand J Prim Health Care. 2016;34(2):205–212. doi: 10.1080/02813432.2016.1183692.
[32] Lyratzopoulos G, Neal RD, Barbiere JM, et al. Variation in number of general practitioner consultations before hospital referral for cancer: findings from the 2010 National Cancer Patient Experience Survey in England. Lancet Oncol. 2012;13(4):353–365. doi: 10.1016/S1470-2045(12)70041-4.
[33] Hauswaldt J, Hummers-Pradier E, Himmel W. Does an increase in visits to general practice indicate a malignancy? BMC Fam Pract. 2016;17(1):94. doi: 10.1186/s12875-016-0477-0.
[34] Gupta VB, Rajagopala M, Ravishankar B. Etiopathogenesis of cataract: an appraisal. Indian J Ophthalmol. 2014;62(2):103–110. doi: 10.4103/0301-4738.121141.
[35] Andreasson A, Hagström H, Sköldberg F, et al. The prediction of colorectal cancer using anthropometric measures: a Swedish population-based cohort study with 22 years of follow-up. United European Gastroenterol J. 2019;7(9):1250–1260. doi: 10.1177/2050640619854278.
[36] Franks PW, Atabaki-Pasdar N. Causal inference in obesity research. J Intern Med. 2017;281(3):222–232. doi: 10.1111/joim.12577.
[37] Yu G-H, Li S-F, Wei R, et al. Diabetes and colorectal cancer risk: clinical and therapeutic implications. J Diabetes Res. 2022;2022:1747326–1747316. doi: 10.1155/2022/1747326.
[38] Sninsky JA, Shore BM, Lupu GV, et al. Risk factors for colorectal polyps and cancer. Gastrointestinal Endoscopy Clin N Am. 2022;32(2):195–213. doi: 10.1016/j.giec.2021.12.008.
[39] Renzi C, Kaushal A, Emery J, et al. Comorbid chronic diseases and cancer diagnosis: disease-specific effects and underlying mechanisms. Nat Rev Clin Oncol. 2019;16(12):746–761. doi: 10.1038/s41571-019-0249-6.
[40] Grancher A, Michel P, Di Fiore F, et al. Colorectal cancer chemoprevention: is aspirin still in the game?. Cancer Biol Ther. 2022;23(1):446–461. doi: 10.1080/15384047.2022.2104561.
[41] Kareemi H, Vaillancourt C, Rosenberg H, et al. Machine learning versus usual care for diagnostic and prognostic prediction in the emergency department: a systematic review. Acad Emerg Med. 2021;28(2):184–196. doi: 10.1111/acem.14190.
[42] Briggs E, de Kamps M, Hamilton W, et al. Machine learning for risk prediction of oesophago-gastric cancer in primary care: comparison with existing risk-assessment tools. Cancers (Basel). 2022;14(20):5023. doi: 10.3390/cancers14205023.
[43] Stapley S, Peters TJ, Sharp D, et al. The mortality of colorectal cancer in relation to the initial symptom at presentation to primary care and to the duration of symptoms: a cohort study using medical records. Br J Cancer. 2006;95(10):1321–1325. doi: 10.1038/sj.bjc.6603439.
[44] Ewing M, Naredi P, Zhang C, et al. Clinical features of patients with non-metastatic lung cancer in primary care: a case-control study. BJGP Open. 2018;2(1):bjgpopen18X101397. doi: 10.3399/bjgpopen18X101397.
[45] Ford E, Nicholson A, Koeling R, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol. 2013;13(1):105. doi: 10.1186/1471-2288-13-105.

[46] Carlsson AC, Wändell P, Ösby U, et al. High prevalence of diagnosis of diabetes, depression, anxiety, hypertension, asthma and COPD in the total population of Stockholm, Sweden - a challenge for public health. BMC Public Health. 2013;13(1):670. doi: 10.1186/1471-2458-13-670.

[47] Tu H, Wen CP, Tsai SP, et al. Cancer risk associated with chronic diseases and disease markers: prospective cohort study. BMJ. 2018;360:k134. doi: 10.1136/bmj.k134.

[48] Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatment. CA Cancer J Clin. 2016;66(4):337–350. doi: 10.3322/caac.21342.