Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A dynamic time-lagged machine learning approach to early colorectal cancer detection in primary care using electronic health records

Emma Bailey
Queen Mary University of London
Andrew Snelling
Queen Mary University of London
Carol Coupland
Queen Mary University of London
Julia Hippisley-Cox
Queen Mary University of London
Jun Wang
j.a.wang@qmul.ac.uk

Queen Mary University of London

Article

Keywords:

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11037540/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. J.H-C. is an unpaid director of QResearch, a not-for-profit organisation which is a partnership between QMUL and EMIS Health who supply the QResearch database used for this work. Until 9 Aug 2023, J.H.C. had a 50% shareholding in ClinRisk Ltd, co-owning it with her husband, who was an executive director. On 9 August 2023, 100% of the share capital was donated to Endeavour Health Care Charitable Trust and the company was renamed to Endeavour Predict Ltd. J.H-C. is a consultant to Endeavour Predict Ltd. and her husband is a non-executive director to cover the transition. The company licences software both to the private sector and to NHS bodies or bodies that provide services to the NHS (through GP electronic health record providers, pharmacies, hospital providers and other NHS providers). This software implements algorithms (including QCancer) developed from access to the QResearch database.

Abstract

Early detection of colorectal cancer (CRC) remains a major unmet need, as current diagnostic pathways rely on symptoms which often appear only at advanced stages. Longitudinal electronic health records (EHRs) provide the means to identify earlier signals, and are well suited to machine learning approaches capable of identifying complex temporal patterns associated with CRC. We present DetectCRC, an XGBoost-based model for early detection of CRC using real-world EHR data. Using 13.6 million QResearch primary care records, the model is developed and evaluated in a matched nested case-control study (22,406 CRC cases; 559,849 controls), incorporating demographics, clinical risk factors, and two years of longitudinal symptom and blood test data using time-lagged features in 6-month advancing windows prior to diagnosis. Longitudinal analyses show that symptom signals are largely confined to the 6 months preceding diagnosis, whereas changes in blood test results, particularly full blood count markers, are detectable up to two years prior. Although multiple biomarkers are evaluated, optimal performance is achieved using routinely available full blood count tests alongside symptoms and risk factors. DetectCRC achieves a C-index of 0.874 in a held-out regional validation set and maintains strong performance on sparse data, demonstrating the potential of routinely collected blood tests for earlier CRC detection.

Introduction

Colorectal cancer (CRC) is the 4th most common cancer in the UK, and the 2nd most common cause of cancer death [1]. Unfortunately, the UK has lower five-year survival rates than comparable high-income countries. A key contributor to poor survival rates is delayed diagnosis: CRC is often detected at a later stage, when outcomes are substantially worse. Five-year survival exceeds 90% for patients diagnosed at stage I, but falls to below 10% at stage IV [2]. This stark contrast highlights the urgent need for improved early detection pathways. With progression from a polyp to CRC estimated to take place over a 10–15 year period [3], there exists a prolonged window during which early pathological changes may be detectable, offering strong potential for earlier diagnosis.

Diagnosis of CRC commonly depends on patients presenting to primary care with symptoms such as rectal bleeding or changes in bowel habit. An awareness initiative in England has been established to increase the public's awareness of cancer-associated symptoms, with the aim of increasing early referrals [4]. However, symptomatic presentation is frequently associated with more advanced disease compared with cancers detected through screening [5]. To promote earlier detection, from 2025 individuals in England aged 50–74 are invited to biennial CRC screening using faecal immunochemical testing (FIT) [6]. However, screening effectiveness is limited by individual uptake and excludes those under 50 years of age, an age group that has experienced a marked rise in CRC incidence over the past 25 years [7].

There are currently limited cancer risk-prediction algorithms using routinely collected primary care electronic health record (EHR) data. QCancer[8] is a publicly available tool to identify individuals at a high risk of certain cancers, however, it is based mainly on symptoms; established risk factors and apart from anaemia, does not incorporate blood test results, which can indicate the presence of CRC [9]. Additionally, this tool is based on single measures of variables at cohort entry rather than repeated measures over time. More advanced models incorporating multiple blood tests have recently been published and shown to outperform the QCancer model [10], however, these models also rely on single timepoint measurements and do not account for longitudinal trends. Furthermore, these approaches utilise logistic regression, which does not natively accommodate the missingness inherent to EHR data, and therefore requires missing data imputation, which may introduce additional uncertainty. In contrast, another model[11] did utilise a machine learning approach using decision-trees, with a range of longitudinal complete blood count results, however the cohort was limited to those aged ≥ 40 and no symptoms or risk factors were included in this model.

Here we develop DetectCRC, an XGBoost [12] based early detection model for CRC using longitudinal EHR data. Unlike previous algorithms, DetectCRC can natively handle sparse and longitudinal data, reducing reliance on missing data imputation. These properties make it well suited for potential integration into routine clinical risk prediction workflows. Within the DetectCRC framework, we investigated the utility of routinely collected blood tests as early biomarkers of CRC, including full blood counts, liver function tests and inflammatory markers, by analysing their longitudinal trends over the five years prior to diagnosis. We then developed and evaluated predictive models which integrate blood test data with symptoms, clinical risk factors, and demographic information using time-lagged features in 6-month advancing time windows preceding diagnosis, to assess the predictive value of combining multi-modal, longitudinal data for CRC early-detection.

Results

A matched nested case-control sample of 22,406 cases and 559,849 controls (1:25 case-control matching) was extracted from a total cohort of 13.6M patients from the QResearch database according to the selection criteria shown in Fig. 1. For each patient, demographic information and five years of relevant clinical history prior to diagnosis or the equivalent index date were obtained from primary care EHRs for trend analysis and model development (Table S1). Longitudinal data were partitioned into ten consecutive six-month time windows to represent the five-year pre-diagnostic period.

The resulting dataset was inherently sparse, particularly for blood tests and lifestyle variables. Across all sparse predictors and time windows, the proportion of observed values ranged from 0.7% (iron level) to 34.7% (bilirubin). Full details of data completeness across the ten time windows are provided in Table S2.

Trend Analysis

To assess which time windows were important for subsequent analysis, general trends were assessed for blood tests and BMI (Fig. 2), and symptoms (Figure S1) reported across the 5 years prior to diagnosis/index date for both cases and controls. These results were based on averages across the entire cohort at each 6-month time window.

Differences in full blood count tests between cases and controls became apparent from the 18–23 month window for mean corpuscular volume (MCV), haemoglobin and platelets. These differences increased up until diagnosis at 0 months, with a particular increase in trend steepness in the 0–5 month window. Haemoglobin and MCV showed a decrease in cases compared to controls, whereas white blood count and platelets showed an increase. White blood count (WBC) appeared to show very minor elevation in cases from 3–4 years prior to diagnosis, but again, the most notable increase was seen from the 18–23 month window onwards.

Similarly, differences in trends for liver function tests were also apparent from 18–23 months prior to diagnosis/index date for albumin, bilirubin and alanine aminotransferase (ALT), with a gradual decrease seen for all until diagnosis at 0 months (and a slight consistent decrease also seen for controls over the full 5 years). In contrast, the difference in trend for gamma-glutamyl transferase (GGT) was only seen in the 0–5 month window, with a sudden increase seen for cases compared to controls (which showed a slight continual increase over the full 5 years).

A decrease in iron level was evident in cases compared to controls from 18–23 months prior to diagnosis onwards, with a steep drop seen in the 0–6 months prior to diagnosis. Ferritin did appear to be slightly lower in cases compared to controls as early as 3–4 years prior to diagnosis, but a more noticeable deviation in trend was seen from 18–23 months prior to diagnosis onwards, when a gradual decrease in ferritin is observed until a slight uptick is seen in the 0–5 month window. In contrast to cases, controls exhibit a continual increase in ferritin over the 5-year period, whilst iron levels showed a slight decrease.

Across the population, BMI appeared to be higher on average in CRC cases compared to controls, until the 0–5 months prior to diagnosis when this dropped rapidly to below the average for controls. The total cholesterol to HDL ratio (TC/HDL) showed a gradual decline over the 5-year period for both cases and controls, with minimal differences seen between the groups.

When looking at inflammation markers, erythrocyte sedimentation rate (ESR) was elevated in cases compared to controls, starting with a gradual increase from 18–23 months and transforming to a steeper increase by 0–5 months prior to diagnosis. C-reactive protein (CRP) was similar in controls and cases, only showing an increase in cases during the 0–5 months prior to diagnosis.

For all symptoms assessed, differences between trends in cases and controls were only visible in the 0–5 months prior to diagnosis, when a sharp increase was seen in symptoms reported for cases (excluding back pain which showed only a very minor increase). These symptoms included abdominal pain, abdominal mass, appetite loss, bowel change, constipation, diarrhoea, rectal bleed, tiredness and weight loss.

Odds ratios: blood tests, health metrics and symptoms

Logistic regression analyses estimated odds ratios (OR) to assess the magnitude of association between each predictor and the likelihood of CRC, with blood tests and BMI modelled using log-transformed values (Fig. 3). Based on the results of the trend analysis, this analysis was focused on the four 6-month time windows from 0–23 months prior to diagnosis. The strongest associations were observed in the 0–5 months prior to diagnosis, particularly for full blood count tests: CRC cases showed higher inflammatory markers, reflected in elevated platelet (OR 12.20, 95% CI 11.50-12.94) and WBC levels (OR 3.79, 95% CI 3.58–4.01), alongside markedly lower anaemia-related markers, MCV and haemoglobin (ORs approaching zero), compared with controls. Other markers of iron deficiency were also evident in cases compared with controls to a slightly lesser degree, with reduced iron (OR 0.12, 95% CI 0.10–0.13) and ferritin levels (OR 0.53, 95% CI 0.52–0.54). Additional inflammatory markers, ESR (OR 1.89, 95% CI 1.83–1.96) and CRP (OR 1.69, 95% CI 1.65–1.73) were also elevated in cases. However, only modest differences were generally observed for liver function tests, including slightly higher GGT (OR 1.28, 95% CI 1.23–1.32) and lower bilirubin (OR 0.45, 95% CI 0.43–0.47) and ALT (OR 0.45, 95% CI 0.43–0.47), although a stronger decrease was seen for albumin (OR 0.01, 95% CI 0.01–0.01). A reduction of BMI was seen in cases during this time (OR 0.50, 95% CI 0.44–0.58), however no significant association was seen for TC/HDL.

Associations weakened in the 6–11 month period prior to diagnosis. Differences in full blood count parameters remained apparent, despite reduced elevation of WBC and platelet counts in cases. Additionally, associations with CRP and ESR inflammatory markers noticeably weakened. Liver function tests signals were weaker, however, evidence of decreased albumin levels remained strong.

Associations continued to lessen through the 12–17 month period, and the 18–23 months prior to diagnosis, however differences in blood test profiles between cases and controls remained at 18–23 months, for full blood count tests in particular. There was a noticeable decrease in MCV (OR 0.21, 95% CI 0.15–0.30), haemoglobin (OR 0.24, 95% CI 0.20–0.30) and albumin (OR 0.29, 95% CI 0.23–0.38) in cases compared with controls, and a slight but significant elevation of platelets (OR 1.38, 95% CI 1.27–1.51) and white blood cell count (OR 1.29, 95% CI 1.17–1.42).

Reports of symptoms showed a similar pattern to blood test results, with the signals predominantly appearing in the 0–5 months prior to diagnosis, but then fading more rapidly in prior time windows. This can be seen clearly when viewing the reported frequencies of all symptoms included in the analysis (Table S3), with 2.18% of cases reporting abdominal pain and 0.75% reporting rectal bleeding in the 18–23 month window, increasing to 15.30% reporting abdominal pain

and 13.76% reporting rectal bleeding in the 0–5 month window, whilst reports of symptoms amongst controls remained steady across all time windows.

The odds ratio analyses confirm these findings, with the signals being strongest in the 0–5 months prior to diagnosis, most notably for reports of rectal bleeding (OR 11.10, 95% CI 10.56–11.68), bowel change (OR 9.29, 95% CI 8.73–9.89) and abdominal mass in CRC patients (OR 6.93, 95% CI 6.08–7.91). There was also evidence of increased reports of appetite loss, abdominal pain, diarrhoea, constipation, weight loss and tiredness amongst cases. Reports of symptoms in cases dropped considerably by 6–11 months prior to diagnosis, and by 12–17/18–23 months prior to diagnosis there was almost no difference in symptoms reported between cases and controls, excluding rectal bleeding which was still reported slightly more for cases than controls (OR 1.34, 95% CI 1.21–1.49 at 18–23 months prior to diagnosis).

Odds ratios: comorbidities and family history

The percentages of cases and controls reporting family history of cancer as well as the comorbidities Crohn's and Ulcerative Colitis (UC) are shown in Table S4. The proportions with reported family history of both non-CRC cancer and CRC specifically, were slightly higher in cases than controls. This is reflected in the logistic regression analysis, which found that a family history of colorectal cancer was associated with a 38% increase in the odds of CRC (OR 1.38, 95% CI 1.22–1.57). A family history of other cancers was associated with a 5% increase in the odds of CRC (OR 1.05, 95% CI 1.02–1.09).

In terms of comorbidities, although absolute prevalence was low, both Crohn's disease and ulcerative colitis were more frequently reported among cases than controls. Individuals with Crohn's disease had 31% higher odds of CRC (OR 1.31, 95% CI 1.10–1.57), while those with ulcerative colitis had 34% higher odds (OR 1.34, 95% CI 1.18–1.51).

At least one record of alcohol consumption and smoking status was available for 61.55% and 82.53% of patients respectively, within the 5 years prior to the diagnosis/index date. Both alcohol consumption and smoking were associated with increased risk of CRC: alcohol intake was associated with a 4% increase in odds per 10-unit/week increase (OR 1.04, 95% CI 1.03–1.05), whilst compared to non-smokers, ex-smokers had an 18% increase in odds of CRC (OR 1.18, 95% CI 1.15–1.22), light smokers had a 14% increase in odds (OR 1.14, 95% CI 1.07–1.22), moderate smokers had a 13% increase in odds (OR 1.13, 95% CI 1.06–1.21), and heavy smokers had a 15% increase in odds (OR 1.15, 95% CI 1.08–1.22).

DetectCRC development

Guided by the trend and odds ratio analyses, model development incorporated full blood count tests and symptoms up to 23 months prior to diagnosis, together with key demographic and clinical risk factors in time-lagged 6-month advancing time windows preceding diagnosis. Additional models were assessed following inclusion of liver function tests and other anaemia and inflammation markers.

Initial models were trained and benchmarked using XGBoost and logistic regression (Table 1), on both non-imputed (i.e. complete cases only) and imputed data (using MICE) using the training and validation split (Fig. 1), as both methods have shown strong performance in model predictions using EHR data [10,13,14]. Additional machine learning models were also evaluated, with LightGBM achieving near-identical performance to XGBoost, while Random Forest performed less well overall.

To allow a fair and direct comparison between XGBoost and logistic regression, which cannot natively handle sparse data, the same complete-case filtering strategies were applied to both models, restricting analyses to patients with complete full blood count data across the defined history windows. Cases had lower levels of missingness for full blood count measurements than controls, resulting in slight shifts in the case:control ratio after complete-case filtering. The numbers of cases and controls retained under each filtering strategy are shown in Table S5.

Table 1
Comparison of performance metrics in the validation dataset for a range of machine learning and logistic regression models, using differing time frames of historical electronic health record data, for the early detection of CRC on non-imputed data.

Model	History	C-index	Sensitivity	Specificity	PPV	NPV	Pred value
lightGBM	**0–5 months**	0.873 (0.866–0.880)	0.813	0.784	0.244	0.980	> 0.35
Random Forest	**0–5 months**	0.776 (0.767–0.786	0.832	0.453	0.115	0.969	> 0.35
XGBoost	**0–5 months**	0.873 (0.866–0.880)	0.812	0.785	0.244	0.980	> 0.35
	0–11 months	0.870 (0.858–0.883)	0.850	0.717	0.186	0.984	> 0.3
	0–23 months	0.859 (0.836–0.881)	0.828	0.704	0.149	0.985	> 0.25
Logistic Regression	**0–5 months**	0.860 (0.853–0.868)	0.810	0.764	0.227	0.979	> 0.1
	0–11 months	0.855 (0.842–0.868)	0.910	0.447	0.111	0.985	> 0.05
	0–23 months	0.853 (0.828–0.878)	0.871	0.584	0.116	0.986	> 0.05

The models were trained on the same features – full blood count tests (haemoglobin, mean corpuscular volume, white blood count and platelets), symptoms (rectal bleed, bowel change, abdominal mass, appetite loss, abdominal pain, diarrhoea, constipation, weight loss and tiredness) as well as age, sex, ethnicity, BMI, diagnosis of UC / Crohn's, family history of cancer / CRC. Models shown here were trained on non-imputed data only.

To allow a direct comparison of XGBoost and logistic regression, the same filtering strategy was applied to each – for 0–5 months history, rows were kept if a full set of full blood count tests was recorded within the 0–5 month window; for 0–11 month history, if recorded in both the 0–5 and 6–11 month window; for 0–23 month history if recorded within each of the 0–5, 6–11, 12–17 and 18–23 month windows.

The 'Pred value' refers to the model's score threshold, which was chosen to optimise sensitivity, to define patients at a high risk of having CRC.

Using non-imputed data, XGBoost showed a small but consistent improvement in performance compared with logistic regression on the validation set. Incorporating patient history beyond the 0–5 month window resulted in a slight reduction in performance. Models trained on imputed data performed marginally worse than those trained on non-imputed data (Table S6).

After demonstrating strong performance on complete imputed data, XGBoost models were subsequently trained on sparse data without imputation to evaluate performance in real-world clinical settings. These models were trained on 2 years of historical data, with varying combinations of predictors (Table 2).

The best-performing model incorporated both symptoms and full blood count tests, achieving a C-index of 0.874 (95% CI, 0.867–0.881). Model performance decreased slightly (0.806–0.817) when additional blood tests - including liver function tests, ferritin, iron levels, and inflammatory markers (ESR and CRP) - were included. Performance was lower (0.799) when models used full blood count tests alone without symptoms, and substantially lower (0.579) when symptoms were used in isolation without full blood count data.

Table 2
Comparison of performance metrics in the validation dataset for a range of XGBoost models, using differing predictors and 2 years of historical electronic health record data, for the early detection of CRC

Predictors used by XGBoost	C-index	Sensitivity	Specificity	PPV	NPV	Pred value
† **Symptoms + FBC**	0.874 (0.867–0.881)	0.910	0.491	0.084	0.991	> 0.2
Symptoms + FBC + LFT	0.806 (0.798–0.814)	0.808	0.597	0.093	0.984	> 0.3
Symptoms + FBC + LFT + ferritin + iron level	0.810 (0.803–0.818)	0.846	0.501	0.080	0.985	> 0.25
Symptoms + FBC + LFT + ferritin + ESR + CRP	0.817 (0.809–0.825)	0.870	0.461	0.076	0.986	> 0.25
FBC only	0.799 (0.791–0.807)	0.821	0.558	0.087	0.984	> 0.3
Symptoms only	0.579 (0.571–0.588)	0.840	0.227	0.053	0.965	> 0.45

Patients were retained for model prediction if at least one set of full blood counts (haemoglobin, platelet count, white blood cell count, and mean corpuscular volume) was recorded within the 2 year history window. Symptoms, when included, were as follows: rectal bleed, bowel change, abdominal mass, appetite loss, abdominal pain, diarrhoea, constipation, weight loss and tiredness. All models also included age, sex, ethnicity, bmi, diagnosis of UC / Crohn's, family history of cancer / CRC, smoking status and alcohol intake. Models were trained and tested on 0–23 months of historical data. The 'Pred value'

refers to the model's score threshold, which was chosen to optimise sensitivity, to define patients at a high risk of having CRC.

† denotes the best performing model which was used for the feature importance and risk distribution plots in Fig. 4

Models were further trained and evaluated across each of the four time windows to assess changes in predictive performance at increasing time intervals before diagnosis (Table 3). Model performance declined substantially beyond the 0–5 month window and continued to decrease gradually at earlier timepoints.

Model performance was also assessed with differing requirements for patient inclusion based on their availability of recent full blood count test results, i.e. requiring a full blood count test in the prior 0–5 months, 0–11 months and 0–23 months (Tables S5 and S7). Whilst the best performance was seen when a full blood count test had been performed in the prior 0–5 months (C-index: 0.886), good performance was still seen when the requirements were relaxed to allow patients with a full blood count in the prior 0–11 months or even the prior 0–23 months, with only a very minor decrease in C-index (0.880 and 0.874 respectively).

Table 3
Comparison of performance metrics in the validation dataset for a range of XGBoost models for the early detection of CRC, looking at sliding 6-month intervals prior to diagnosis

	C-index	Sensitivity	Specificity	PPV	NPV	Threshold
	All symptoms + FBC					
XGBoost – lag0 **(0–5 months history)**	0.873 (0.866–0.880)	0.812	0.785	0.244	0.980	> 0.35
XGBoost – lag1 **(6–11 months history)**	0.649 (0.635–0.663)	0.734	0.436	0.053	0.975	> 0.42
XGBoost – lag2 **(12–17 months history)**	0.593 (0.578–0.608)	0.904	0.138	0.042	0.972	> 0.48
XGBoost – lag3 **(18–23 months history)**	0.554 (0.539–0.569)	0.859	0.189	0.041	0.971	> 0.48

Patients were retained for model prediction if a full blood count test (haemoglobin, platelet count, white blood cell count, and mean corpuscular volume) was recorded in the 0–5 month window for the lag0 model, the 6–11 month window for the lag1 model, the 12–17 month window for the lag2 model and the 18–23 month window for the lag3 model. All models were trained on the full blood count tests; the following symptoms: rectal bleed, bowel change, abdominal mass, appetite loss, abdominal pain, diarrhoea, constipation, weight loss and tiredness; as well as age, sex, ethnicity, bmi, diagnosis of UC / Crohn's, family history of cancer / CRC, smoking status and alcohol intake. The 'Pred value' refers to the model's score threshold, which was chosen to optimise sensitivity, to define patients at a high risk of having CRC.

To assess model stability of the final model (full blood count and symptoms in Table 2), five-fold cross-validation was performed within the training set, with the model trained on four folds and evaluated on the held-out fold in each iteration. The resulting ROC curves shown in Fig. 4A were highly consistent across folds, demonstrating stable predictive performance and limited variability in AUC estimates.

Model evaluation

We further evaluated the final DetectCRC model in terms of its prediction scores and feature importance. Overall, in the validation set, the CRC case samples had significantly markedly higher risk prediction scores than control samples (median score of 0.202 for controls, and 0.836 for cases, $p < 2.2e\text{-}16$, Fig. 4B). A threshold of 0.2 was selected as it maintained high sensitivity (91.01%) while improving specificity compared to a lower threshold of 0.15 (49.09% vs 24.95%), as shown in Table 4. Higher thresholds reduced sensitivity (≤ 85.42%), with the proportion of missed cases increasing from approximately 1 in 11 at 0.2 to 1 in 7 at 0.25, a level likely too high for clinical screening.

Table 4
Distribution of predicted classes when using a range of cut off values to assign high risk of CRC

Pred value	True Positive	False Positive	True Negative	False Negative	Accuracy	Sensitivity	Specificity	PPV	NPV
0.15	4,134	62,987	20,941	159	28.42%	96.30%	24.95%	6.16%	99.25%
0.20	3,907	42,727	41,201	386	51.13%	91.01%	49.09%	8.38%	99.07%
0.25	3,667	26,494	57,434	626	69.26%	85.42%	68.43%	12.16%	98.92%
0.30	3,520	19,145	64,783	773	77.42%	81.99%	77.19%	15.53%	98.82%

The distribution of predicted classes within the validation set using an XGBoost model trained on 2-years of historical blood test and symptom data – corresponds to the model denoted by 'ŧ' in Table 2.

To assess feature contribution, we produced a SHAP summary plot (Fig. 4C). The features which contributed most strongly to the model as indicated by the highest mean absolute SHAP values, were three full blood count measures (haemoglobin, platelets and MCV), followed by several symptoms, including rectal bleeding, abdominal pain and bowel change. For each of these features, measurements from the 0–5 month window prior to diagnosis were of greatest importance. The direction of these effects was broadly consistent with the trend and odds ratio analyses, with lower haemoglobin and mean corpuscular volume contributing to increased predicted risk, while higher platelet count and white blood cell count were associated with increased risk. Symptom reports within this time window, such as rectal bleeding, abdominal pain, diarrhoea and constipation also increased predicted risk, with repeated reports of such symptoms within the window contributing more strongly.

Historical full blood count results continued to contribute to model predictions, with measurements from the 6–11, 12–17 and 18–23 month windows appearing among the top 25 features. However, the directionality of these associations became less distinct further from diagnosis, suggesting a diminishing signal over time. Symptom reports outside the initial 0–5 month window showed comparatively low importance; with symptoms reported further from diagnosis generally associated with reduced predicted risk, suggesting earlier symptoms are less specific and more often benign. An exception to this was seen for rectal bleeding, which remained positively associated with risk across earlier time windows, indicating that its predictive signal persists further from diagnosis than other symptoms.

Age was the fourth most important predictor, with younger individuals tending to show a higher predicted risk. However, as cases and controls were age matched, this should not be interpreted as younger age conferring a higher overall risk of CRC. Instead, SHAP values reflect how features influence predictions conditional on the other variables in the model; consequently, CRC-associated clinical features such as symptom reports or anaemia may exert a stronger influence on risk prediction in younger individuals, where these findings are less expected.

Sex fell within the top ten predictors, with male sex contributing modestly towards higher model-predicted risk, although cases and controls were sex matched. Ethnicity was the eleventh most important predictor, however, the

cohort was heavily skewed towards white patients, with relatively few cases in other ethnic groups, limiting the reliability of any observed differences in risk by ethnicity.

Factors such as smoking status, alcohol intake and a diagnosis of ulcerative colitis also contributed to the final model but showed much lower overall importance compared to sex and age, while a diagnosis of Crohn's disease and family history of cancer did not appear among the top 50 features at all.

Discussion

We have developed and validated an early detection algorithm to predict the likelihood of a patient having a current but not yet diagnosed case of CRC, with the potential for detection up to 2-years earlier than current diagnostic timelines. After assessing the performance of multiple algorithms, using different combinations of blood tests and time frames, the final best performing model incorporated a range of predictors, including full blood counts, relevant symptoms and associated risk factors, utilising up to 2 years of historical data.

Our analysis demonstrates that XGBoost is well suited for developing predictive models using electronic health record data. Its performance is comparable to that of recently developed logistic regression–based approaches [10], while avoiding the need for explicit data imputation. Although multiple imputation by chained equations (MICE) is widely regarded as a robust approach, its performance has been shown to deteriorate under conditions of high missingness (i.e. >50%), particularly in longitudinal health data [15]. In our cohort, missingness in blood test variables substantially exceeded this threshold, making imputation likely to introduce inaccuracies. Consistent with this, a slight reduction in predictive performance was observed in models trained on imputed data compared with those restricted to cases and controls with complete data. XGBoost's ability to work directly on the sparse data effectively addresses these issues.

After assessing the 5-year trends of CRC-associated biomarkers in cases compared to controls, the model was based on 2 years of historical data, with four consecutive 6-month time windows prior to diagnosis. When assessing model performance across each individual time window, performance declined with increasing time from diagnosis. This aligns with the natural progression of CRC, as physiological changes intensify nearer to diagnosis. Trend analysis showed that symptoms tend to appear only in the 5 months immediately preceding diagnosis, while blood test abnormalities are detectable up to two years prior and become increasingly pronounced over time.

It has previously been reported that the median time from symptom onset to diagnosis of CRC is 14 weeks [16], with the most frequently reported symptoms being rectal bleeding, abdominal pain and change in bowel habits. This coincides with our analysis, with these three symptoms showing the strongest contributions to our model when assessing SHAP scores; and with our trend analysis, which showed that symptoms were generally only reported in the 0–5 months prior to diagnosis. Whilst symptoms are a key feature of predictive CRC models, their usefulness for early detection before this initial time frame is limited, highlighting the benefit of including other clinical biomarkers.

Blood test results provide a larger scope for early detection of CRC. Trend analysis showed that markers of anaemia (haemoglobin, MCV, ferritin and iron level) began to decline up to 2 years prior to diagnosis in cases, showing a clear separation from controls. Studies have found that many CRC patients present with anaemia at diagnosis [17–20], and several studies have also identified anaemia as a marker of CRC up to 2 years prior to diagnosis [21–24]. However, these studies often rely on fixed clinical thresholds to define anaemia, which may miss earlier changes occurring while levels remain within the normal range, and additionally does not account for the patient's own baseline. Using longitudinal trends of these anaemia markers is therefore likely to be more informative than defined thresholds for early detection.

Among markers of anaemia, haemoglobin and MCV emerged as the strongest early indicators of CRC, showing more consistent and pronounced associations across the four 6-month time windows than ferritin or iron levels. In addition, being part of a standard full blood count, these markers were substantially more frequently recorded (approximately 30% of records), compared with ferritin and iron levels, which were rarely available. This combination of stronger predictive signal and greater availability supports their suitability for an early detection model.

Interestingly, trend analysis showed a slight uptick in ferritin levels for CRC patients, in the final 0–5 month period prior to diagnosis, whilst all other markers of anaemia continued to decline. This may be the result of increased inflammation in CRC patients. Inflammation has been shown to trigger the synthesis of ferritin, with the cytokine network thought to modulate iron metabolism in a manner resulting in low serum iron and elevated ferritin [25]. This correlates with our findings, with trend analysis showing clear signs of elevated markers of inflammation. Platelets, WBC, ESR and CRP all show a notable increase in CRC patients, particularly in this 0–5 month period prior to diagnosis, with WBC and platelet levels showing divergence from controls at an earlier time point than CRP, making them well-suited markers for our model. Other studies have also shown that an increase in inflammation markers is associated with CRC. High levels of CRP, WBC, platelets and ESR have all been associated with increased risk of CRC prior to diagnosis, particularly in the 6–9 months prior [22,26,27].

Platelet levels and WBC are routinely recorded as part of a full blood count and were therefore included in the early detection model as standard markers of inflammation. We then assessed model performance with the addition of ESR and CRP, but observed little change in performance. This suggests that the relevant inflammatory signal was already captured through the full blood count markers, and given the high level of missingness for ESR and CRP, these markers were not retained in the final model.

There is limited research examining the relationship between liver function tests and early detection of CRC, and to our knowledge, longitudinal trends in liver function test results up to two years prior to CRC diagnosis have not been well explored. Our trend analysis showed that albumin, bilirubin and ALT declined more steeply in CRC cases than in controls up to two years prior to diagnosis. In contrast, GGT levels only diverged close to diagnosis, with a marked increase observed in cases in the final 0–5 months. Other studies have shown similar trends. One study examined blood test results in the 1–6 month period prior to diagnosis and reported reduced ALT, albumin and bilirubin levels alongside elevated GGT in CRC cases compared with controls [28]. Whilst another study concluded that lower levels of ALT, albumin, bilirubin and GGT were associated with a higher risk of CRC [29], citing a potential link between liver function and the gut microbiota in contributing to CRC risk. The disparity in findings regarding GGT could be because this study examined overall CRC risk over a 10-year follow-up period as opposed to focusing on markers in the few months directly before diagnosis.

Overall, our findings suggest that liver function markers may have potential as early indicators of CRC. However, as values largely remained within normal clinical ranges, it appears that longitudinal trends, rather than absolute levels, carry the relevant signal. When incorporated into the predictive model alongside full blood counts and symptoms, inclusion of liver function markers resulted in a slight reduction in performance. This suggests that the predictive information captured by these markers may already be reflected in the full blood counts, or that they introduce additional noise. This noise could be due to the presence of noticeable downward trends also being observed for these markers in controls (except for GGT which shows a slight upward trend over time). These trends are likely related to ageing, with previous studies reporting declining ALT and increasing GGT levels with age [30].

After identifying full blood counts and symptoms as the key markers for DetectCRC, the model performance was evaluated under varying eligibility criteria, requiring at least one full blood count to have been recorded within 6, 12, or 24 months prior to diagnosis. Performance was highest when a test had been performed within 6 months prior, but

declined only modestly when this window was extended to 12 or 24 months. As over one third of the cohort had at least one full blood count within 6 months, and more than two thirds within 24 months, this suggests the model would be applicable to a substantial proportion of patients in routine clinical practice. These findings could also warrant more regular full blood count testing for patients. Annual measurements could enhance capture of patient-specific trends and increase the efficacy of early detection models, although the feasibility and cost-effectiveness of such an approach would require further evaluation.

While this study demonstrates promising findings, limitations should also be noted when interpreting the results. In primary care, blood tests are often performed in response to symptoms or clinical concern, meaning patients developing CRC are more likely to undergo testing than controls. This was reflected in our dataset, where cases had lower levels of missingness for full blood count measurements than controls. As XGBoost can learn from missingness, part of the predictive signal may reflect testing behaviour in addition to biological change. Although this may influence predictions across groups with different testing rates, it reflects current real-world practice and may support clinical applicability. Moreover, although unseen regions of the UK (South East and West Midlands) were used as the testing split, further validation in independent national and international datasets is required to confirm generalisability.

Summary

Our final best-performing model, termed DetectCRC, included 2 years of historical full blood count results and clinically relevant symptoms, in addition to key demographic and clinical risk factors using time-lagged features in 6-month advancing time windows prior to diagnosis. Although other markers demonstrate detectable longitudinal changes, such as liver function markers, and additional markers for inflammation and anaemia, they do not meaningfully enhance predictive performance when full blood counts are already included.

The identification of full blood counts as the primary predictors is advantageous, as these tests are among the most routinely requested in primary care, increasing the potential applicability of the model in real-world settings. Additionally, these markers act as a composite signature: while anaemia markers (MCV and haemoglobin) or inflammatory markers (WBC and platelets) alone may be indicative of multiple conditions, their combined pattern forms a distinctive fingerprint which is more specific to CRC.

Our model demonstrated strong predictive performance, with a c-index of 0.874. Its ability to operate directly on sparse data from primary care electronic health records, while incorporating patient-specific longitudinal trends rather than relying on fixed cut-offs, supports its potential utility as an early detection tool for CRC and could be used by GPs to flag at risk patients for further testing, such as FIT.

Methods

Study design and data sources

Data was obtained from the electronic medical record database, QResearch. This database comprises anonymised medical record data collected across England from hospitals, GP practices as well as the cancer registry. QResearch is based on the EMIS (Egton Medical Information Systems) clinical system, which hosts electronic health records (EHRs).

CRC cases were defined using ICD-10 site codes C18 (colon), C19 (rectosigmoid junction) or C20 (rectum). Tumours with histology codes corresponding to non-epithelial or non-relevant morphologies (ICD-O histology codes 9050–9055, 9140, 9590, and 9992) were excluded. The criteria for selecting CRC patients for the cohort was a first-time diagnosis with CRC recorded in the cancer registry, between January 2010 and 31st December 2018, with at least 1 year of

practice registration prior to the diagnosis date and no other cancer diagnosed in the cancer registry in the 2 years prior to CRC diagnosis (excluding non-melanoma skin cancer).

The CRC cases were matched to controls by GP practice, sex and age ±10 years. This matching was performed at a ratio of 1:25 cases to controls, broadly reflecting the rarity of CRC in the general population. Where > 25 eligible controls were available, matches were prioritised by smallest absolute age difference. The only inclusion criterion for controls was an absence of CRC diagnosis at the index date and at least 1 year of practice registration prior to the index date (which was the diagnosis date of the matched CRC case).

Patients aged 20–100 at the time of diagnosis (or the index date for controls) were eligible for the cohort. Patients were excluded if ethnicity data was absent – this was to prevent bias in the model, since the majority of CRC patients had ethnicity data present from the cancer registry, whereas ethnicity data was frequently absent in potential controls. Patients from the West Midlands and South East regions were left out of the training cohort, and formed an external validation set. All other regions were included in the training cohort. A summary of the demographics of the full cohort are included in Table S8.

Predictor Variables

Potential predictors included variables which are used in clinical practice and recorded in GP records. This included factors which are associated with baseline probability of CRC, such as smoking, alcohol, family history of CRC and other cancers and comorbidities (UC and Crohn's), as well as predictors which could facilitate detection of an undiagnosed cancer (e.g. symptoms and blood tests). Symptoms were selected based on existing prediction models and the literature [31–35], including both red-flag symptoms (those which warrant further assessment) and more general symptoms. Blood tests were chosen based on those frequently requested by clinicians if cancer is suspected, as well as existing prediction models and the literature [36–39]. Blood test results outside biologically plausible ranges were excluded, such as negative values.

Odds Ratios

Odds ratios were estimated using unconditional logistic regression models (*glm*, binomial family) on non-imputed values, using log x + 1 for blood tests / BMI. No transformation was applied to symptoms.

Model development

Models were developed using XGBoost, as well as multivariable logistic regression for benchmarking. Unlike logistic regression, XGBoost natively handles sparse data, therefore, most analyses were performed directly on the sparse data. However, multiple imputation with chained equations (MICE) was used to generate an imputed version of the dataset to determine whether imputation provided any meaningful performance advantage.

The R package *mice* v3.19.0 [40] was used to replace missing values for smoking (which was categorised as never, former, light, moderate, or heavy smoker) and alcohol intake (recorded as a continuous variable using units/week), only for patient rows where at least one record was present (i.e. at least one 0–5 month time window); and for blood tests and health metrics, which were all stored as continuous variables (body mass index, haemoglobin, platelets, white blood cell count, mean corpuscular volume, erythrocyte sedimentation rate, total cholesterol: HDL cholesterol ratio, iron level, ferritin, albumin, alanine aminotransferase, bilirubin and gamma-glutamyl transferase), only for rows where at least two records were present.

Binary static variables (family history of CRC and comorbidities) were coded as present if these were listed in the GP record and otherwise coded as absent. Family history of cancer excluding CRC was encoded as 0–3, based on the number of types of cancer recorded, representing 0 to ≥ 3.

Time-varying variables were represented using a lagged feature approach. For each individual, clinical variables from the five years preceding the index date were encoded as 6-monthly lags: 0 to < 6 months, 6 to < 12 months and so on, up to 54 to < 60 months, resulting in a set of temporally ordered predictors (i.e. lag0–lag9) that capture historical information. If multiple values were present for continuous variables (e.g. blood tests) within any of these 6-month windows, then an average of these test results was taken. Symptoms were recorded as present only within the 6-month window that they were reported. Symptoms were encoded as values from 0–3, with 0 representing no reports for that symptom within the time window, 1 representing a single report, 2 representing the symptom having been reported twice and 3 representing the symptom having been reported 3 or more times.

The predictor variables were first assessed using a trend analysis over the 5-years prior to diagnosis / index date, to decide their suitability for the model. It was then decided to use 2 years of history for model development and predictions.

Logistic regression models were trained using *glm* with the 'binomial' argument. To account for class imbalance, a weight of 2 was assigned to cases and 1 to controls; more extreme weighting proportion to class prevalence resulted in poorer performance.

The R package *xgboost* v1.7.8.1[12] for R was used to train models using a binary logistic objective. Model training used a learning rate of 0.01, maximum tree depth of 6, and L2 regularisation ($\lambda = 1$), with additional regularisation through a minimum child weight of 5 and a gamma penalty of 1. Subsampling of observations and features was applied (subsample = 0.8; colsample_bytree = 0.8) to reduce overfitting. Models were trained using the histogram-based tree method. To account for class imbalance, positive cases were upweighted using the scale_pos_weight parameter, set according to the observed case-to-control ratio in the training set. This value varied slightly across model runs where different cohort filtering criteria were applied. *shap.prep* was used to obtain the SHAP scores from the model, which represent the contribution of parameters to the model.

Model evaluation

The XGBoost model generates a probability between 0 and 1 for each individual. This probability reflects the degree to which the individual's pattern of symptoms, blood test results and clinical characteristics resembles those observed among CRC cases compared with controls in the training dataset. Values closer to 1 indicate a stronger model-estimated likelihood of CRC, whereas values closer to 0 indicate a lower likelihood. These values do not represent absolute risk of CRC and instead should be interpreted as model-derived probabilities. These values were used for risk stratification and performance evaluation.

For all models, the validation set was used to assess performance by calculating the C-statistic (equivalent to the Area under the Receiver Operating Curve), sensitivity, specificity, PPV and NPV. The threshold of model-derived probability for the patient having CRC was set per model and is stated alongside each report of model performance. This threshold was decided by assessing the confusion matrix and prioritising case detection (sensitivity) while maintaining reasonable discrimination of controls.

Declarations

Data availability

The electronic health data generated in this study are derived from the QResearch database under accession code OX309: https://www.qresearch.org/research/research-programs-and-projects/maars-multimodal-active-adaptive-risk-stratification-for-cancer/. The raw data are not available under restricted access for to ensure compliance with patient

privacy, ethical and license conditions for access to patient electronic health records in England. The processed data access can be obtained by following the published information on the QResearch website (www.qresearch.org).

Code availability

Clinical codes are published under a Creative Commons licence here: https://www.qresearch.org/data/qcode-group-library/ with accompanying details in Supplementary Table 1.

The code for the trained models is publicly available on Github https://github.com/BioInforCore-BCI/DetectCRC.

Acknowledgments

This work was funded jointly by Cancer Research UK (CRUK), the Medical Research Council (MRC) and the Engineering & Physical Sciences Research Council (EPSRC), under the CRUK Early Detect and Diagnosis Innovation Sandpit Award (Ref: EDDISA-Jan22\100001). E.B. and J.W. acknowledge support from the Barts Charity Strategic Award to Barts Centre for Squamous Cancer, QMUL (Ref: G-002030), and Cancer Research UK Clinical Research Biomarker Project Award (Ref: CRCBPA-Nov24/100003). E.B. and J.W. also acknowledge support from CRUK City of London Centre Award [CTRQQR-2021\100004] to Barts Cancer Institute, and the Accelerator Award Program funded by Cancer Research UK (C355/A26819) and FC AECC and AIRC. J.W. also acknowledges support from the Academy of Medical Sciences Springboard Award (SBF003\1025). J.H-C. acknowledges support from NIHR Senior Investigator Award scheme.

We acknowledge the contribution of EMIS practices who contribute to QResearch and Queen Mary University of London for expertise in developing or supporting the QResearch database. This project involves data derived from patient-level information collected by the NHS, as part of the care and support of cancer patients. The hospital, cancer and mortality data are collated, maintained and quality assured by the National Disease Registration Service which is part of NHS England. Access to the data was facilitated by the NHS England Data Access Request service. NHS England bears no responsibility for the analysis or interpretation of the data.

Author contributions

J.H-C. and J.W. conceived the study and acquired the funding. E.B. and J.W. developed the methodology and performed the data analysis and model development. E.B., A.S., C.C., J.H-C., and J.W. conducted the data curation and extraction. C.C. and J.H-C. contributed to the methodology and data interpretation. E.B. and J.W. interpreted the data and wrote the manuscript. All authors reviewed the manuscript and approved the final version of the manuscript.

Competing interests

J.H-C. is an unpaid director of QResearch, a not-for-profit organisation which is a partnership between QMUL and EMIS Health who supply the QResearch database used for this work. Until 9 Aug 2023, J.H.C. had a 50% shareholding in ClinRisk Ltd, co-owning it with her husband, who was an executive director. On 9 August 2023, 100% of the share capital was donated to Endeavour Health Care Charitable Trust and the company was renamed to Endeavour Predict Ltd. J.H-C. is a consultant to Endeavour Predict Ltd. and her husband is a non-executive director to cover the transition. The company licences software both to the private sector and to NHS bodies or bodies that provide services to the NHS (through GP electronic health record providers, pharmacies, hospital providers and other NHS providers). This software implements algorithms (including QCancer) developed from access to the QResearch database.

References

1. Bowel cancer statistics | Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#BowelCS0.
2. Araghi, M. *et al.* Colon and rectal cancer survival in seven high-income countries 2010–2014: variation by age and stage at diagnosis (the ICBP SURVMARK-2 project). *Gut* 70, 114–126 (2021).
3. Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. Colorectal cancer. *The Lancet* 394, 1467–1480 (2019).
4. Richards, M. A. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br. J. Cancer* 101, S1 (2009).
5. Novotny, S. A. *et al.* Prognostic Study of Colorectal Cancer: Differences between Screen-Detected and Symptom-Diagnosed Patients. *Cancers (Basel).* 16, 3363 (2024).
6. NHS England. Bowel cancer screening: programme overview - GOV.UK. https://www.gov.uk/guidance/bowel-cancer-screening-programme-overview (2025).
7. Vuik, F. E. R. *et al.* Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. *Gut* 68, 1820 (2019).
8. 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* 62, e29 (2011).
9. Virdee, P. S. *et al.* The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers (Basel)*. 12, 2348 (2020).
10. Hippisley-Cox, J. & Coupland, C. A. Development and external validation of prediction algorithms to improve early diagnosis of cancer. *Nat. Commun.* 16, 3660 (2025).
11. Kinar, Y. *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. *Journal of the American Medical Informatics Association* 23, 879–890 (2016).
12. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
13. Guan, X. *et al.* Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. *BMC Med. Inform. Decis. Mak.* 23, 107 (2023).
14. Ke, X. *et al.* Predicting early gastric cancer risk using machine learning: A population-based retrospective study. *Digit. Health* 10, 20552076241240904 (2024).
15. Junaid, K. P., Kiran, T., Gupta, M., Kishore, K. & Siwatch, S. How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. *Population Health Metrics 2025 23:1* 23, 2- (2025).
16. Majumdar, S. R., Fletcher, R. H. & Evans, A. T. How does colorectal cancer present? Symptoms, duration, and clues to location. *Am. J. Gastroenterol.* 94, 3039–3045 (1999).
17. Beale, A. L., Penney, M. D. & Allison, M. C. The prevalence of iron deficiency among patients presenting with colorectal cancer. *Colorectal Disease* 7, 398–402 (2005).
18. Edna, T. H., Karlsen, V., Jullumstrø, E. & Lydersen, S. Prevalence of anaemia at diagnosis of colorectal cancer: assessment of associated risk factors. *Hepatogastroenterology* 59, 713–716 (2012).
19. Ho, C. H., Yu, Y. Bin & Wu, P. H. The prevalence of iron deficiency anemia and its clinical implications in patients with colorectal carcinoma. *Journal of the Chinese Medical Association* 71, 119–122 (2008).

20. Mashlab, S. *et al.* Anaemia as a risk stratification tool for symptomatic patients referred via the two-week wait pathway for colorectal cancer. https://doi.org/10.1308/rcsann.2018.0030 100, 350–356 (2018).
21. Hamilton, W. *et al.* The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. *BMC Med.* 7, 17 (2009).
22. Moullet, M. *et al.* Pre-diagnostic clinical features and blood tests in patients with colorectal cancer: a retrospective linked-data study. *British Journal of General Practice* 72, e556–e563 (2022).
23. Virdee, P. S., Patnick, J., Watkinson, P., Birks, J. & Holt, T. A. Trends in the full blood count blood test and colorectal cancer detection: a longitudinal, case-control study of UK primary care patient data. *NIHR Open Research* 2, 32 (2022).
24. Schneider, C., Bodmer, M., Jick, S. S. & Meier, C. R. Colorectal cancer and markers of anemia. *European Journal of Cancer Prevention* 27, 530–538 (2018).
25. Feelders, R. A. *et al.* Regulation of iron metabolism in the acute-phase response: interferon gamma and tumour necrosis factor alpha induce hypoferraemia, ferritin production and a decrease in circulating transferrin receptors in cancer patients. *Eur. J. Clin. Invest.* 28, 520–527 (1998).
26. Rafiq, M. *et al.* Trends in primary care blood tests prior to lung and colorectal cancer diagnosis—A retrospective cohort study using linked Australian data. *Cancer Med.* 13, e70006 (2024).
27. Allin, K. H., Bojesen, S. E. & Nordestgaard, B. G. Inflammatory biomarkers and risk of cancer in 84,000 individuals from the general population. *Int. J. Cancer* 139, 1493–1500 (2016).
28. Goshen, R. *et al.* Predicting the presence of colon cancer in members of a health maintenance organisation by evaluating analytes from standard laboratory records. *Br. J. Cancer* 116, 944 (2017).
29. He, M. ming *et al.* Circulating liver function markers and colorectal cancer risk: A prospective cohort study in the UK Biobank. *Int. J. Cancer* 148, 1867–1878 (2021).
30. Petroff, D. *et al.* Age Dependence of Liver Enzymes: An Analysis of Over 1,300,000 Consecutive Blood Samples. *Clinical Gastroenterology and Hepatology* 20, 641–650 (2022).
31. Biswas, M., Ades, A. E. & Hamilton, W. Symptom lead times in lung and colorectal cancers: what are the benefits of symptom-based approaches to early diagnosis? *Br. J. Cancer* 112, 271–277 (2015).
32. Renzi, C., Lyratzopoulos, G., Hamilton, W. & Rachet, B. Opportunities for reducing emergency diagnoses of colon cancer in women and men: A data-linkage study on pre-diagnostic symptomatic presentations and benign diagnoses. *Eur. J. Cancer Care (Engl)*. 28, e13000 (2019).
33. Jones, R., Latinovic, R., Charlton, J. & Gulliford, M. C. Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database. *BMJ* 334, 1040 (2007).
34. Hippisley-Cox, J. & Coupland, C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* 63, e1–e10 (2013).
35. Hippisley-Cox, J. & Coupland, C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* 63, e11–e21 (2013).
36. Virdee, P. S., Patnick, J., Watkinson, P., Holt, T. & Birks, J. Full Blood Count Trends for Colorectal Cancer Detection in Primary Care: Development and Validation of a Dynamic Prediction Model. *Cancers (Basel)*. 14, 4779 (2022).
37. He, M. *et al.* Circulating liver function markers and colorectal cancer risk: A prospective cohort study in the UK Biobank. *Int. J. Cancer* 148, 1867–1878 (2021).
38. Gunter, M. J. *et al.* A Prospective Study of Serum C-Reactive Protein and Colorectal Cancer Risk in Men. *Cancer Res.* 66, 2483–2487 (2006).

39. Kantor, E. D. *et al.* Adolescent body mass index and erythrocyte sedimentation rate in relation to colorectal cancer risk. *Gut* 65, 1289–1295 (2016).
40. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45, 1–67 (2011).

Figures

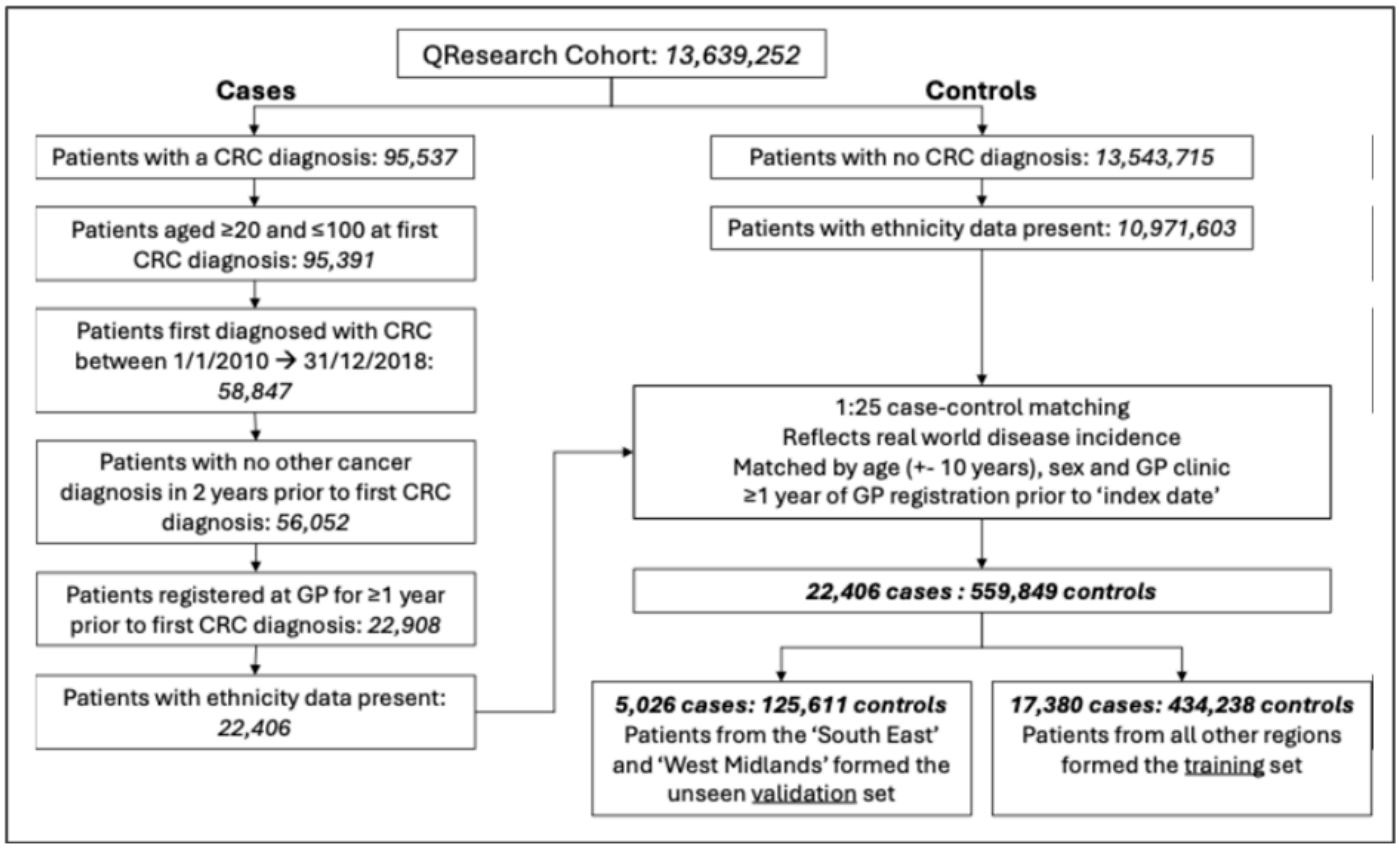


Figure 1

Flowchart showing selection criteria for the cohort used for model training and validation

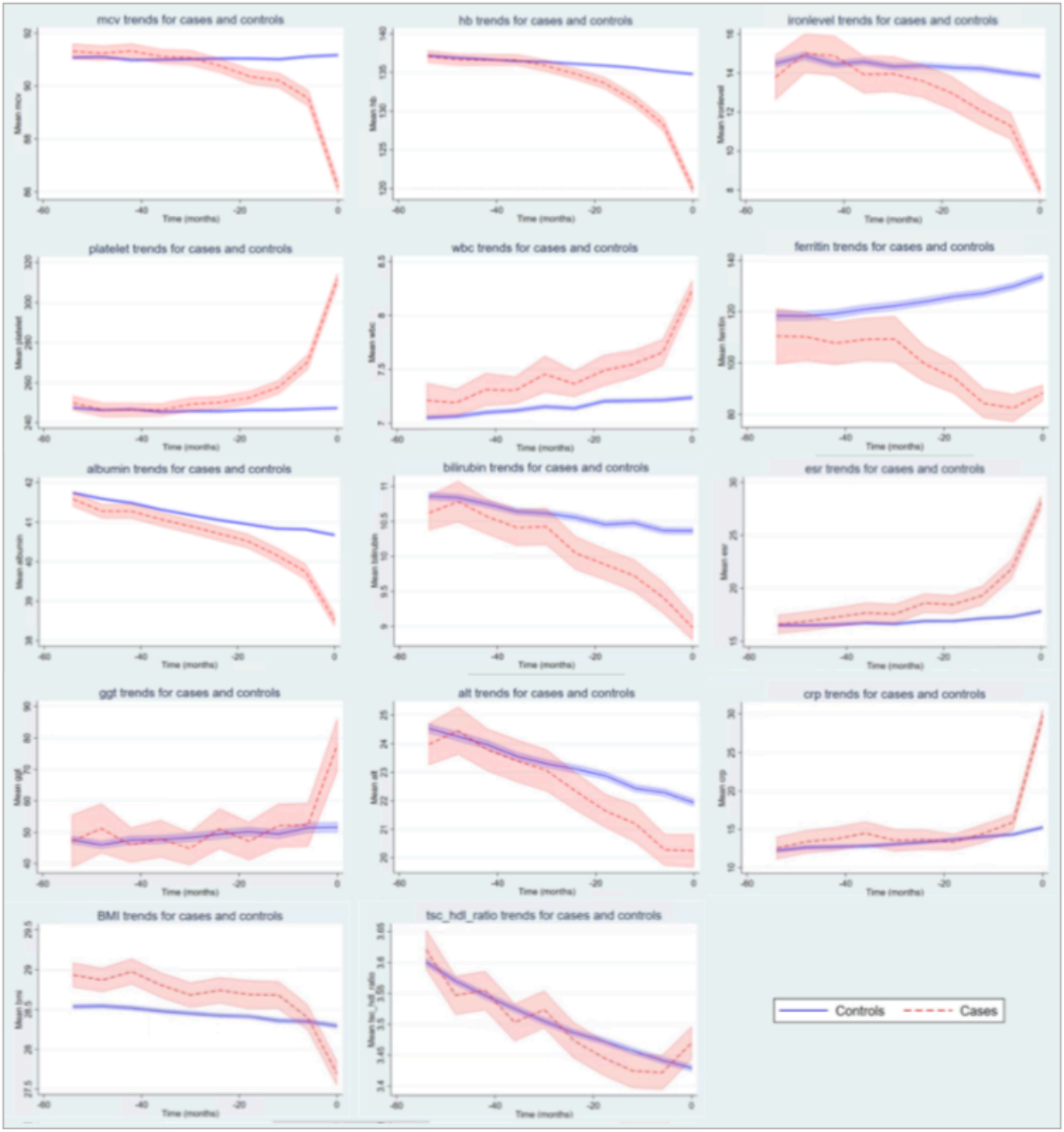


Figure 2

Trend analysis of full blood count tests (mean corpuscular volume, haemoglobin, platelets and white blood count), liver function tests (albumin, bilirubin, gamma-glutamyl transferase and alanine aminotransferase), iron level, ferritin, inflammation markers (erythrocyte sedimentation rate and C-reactive protein) and BMI and total cholesterol to HDL ratio across the 0-5 years prior to diagnosis for cases and controls. *Cases shown in red and controls in blue. The x-axis represents the time in months prior to diagnosis, from 0 to -54 months (representing the start of the final 54-59 month window). 95% confidence intervals were shown (shaded areas) along with the average values.*

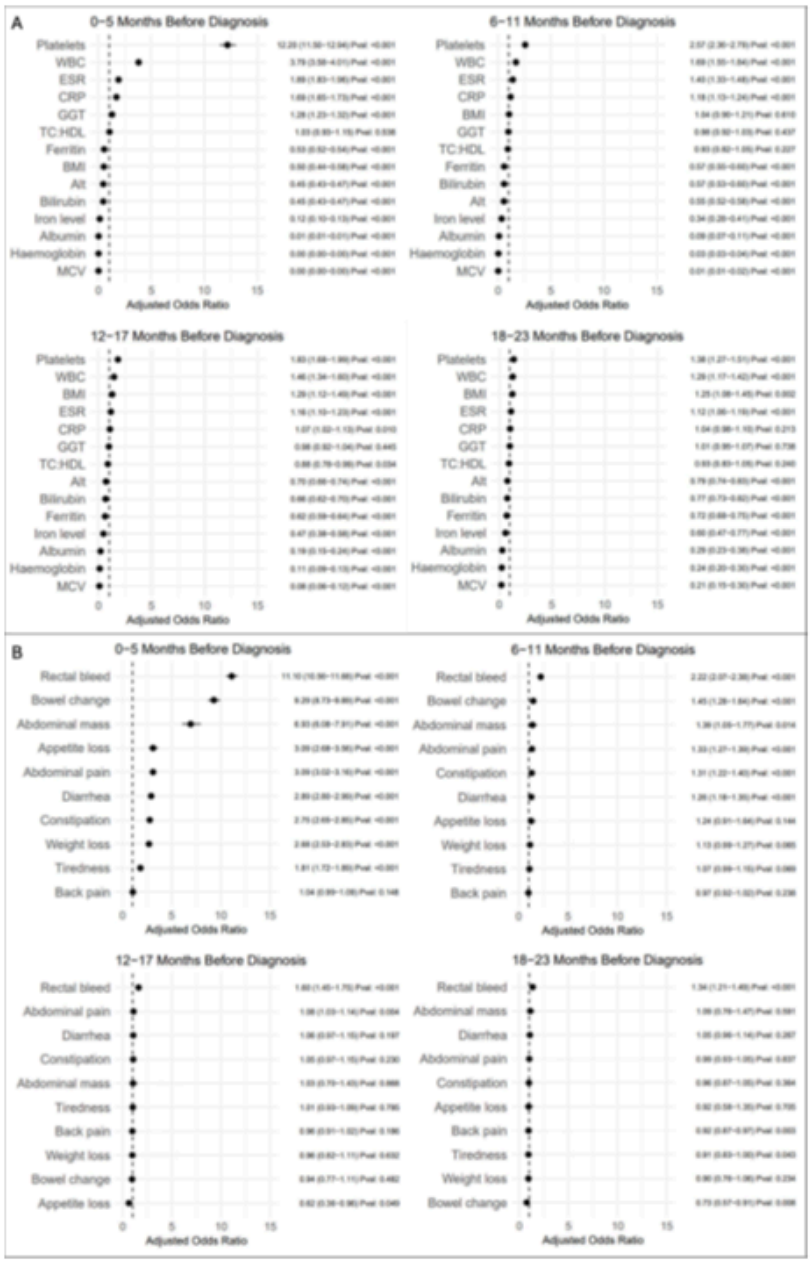


Figure 3

Odds ratio analyses of (A) common blood tests and (B) CRC-related symptoms. Adjusted for age and sex. Ordered by absolute odds ratios. Odds ratios rounded to two decimal places. 95% confidence intervals shown in brackets.

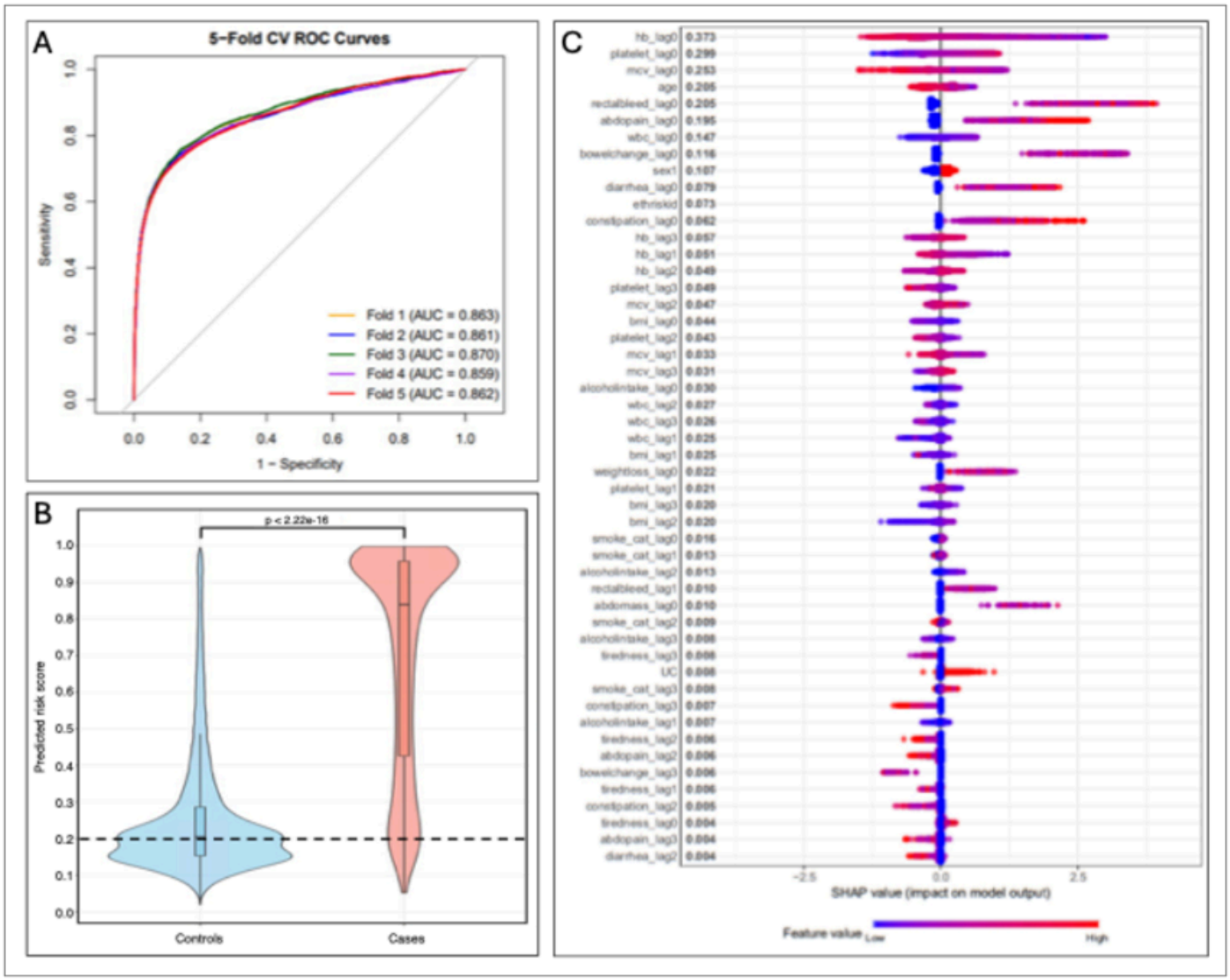


Figure 4

A. ROC curves from five-fold cross-validation of the full blood count and symptoms XGBoost model, with fold-specific AUC values indicated. B. The distribution of risk scores predicted on the validation set using an XGBoost model trained on 2-years of historical blood test and symptom data – corresponds to the model denoted by '†' in table 2. Controls are shown in blue, and CRC cases in red. The dashed line represents the cutoff used to assign high-risk vs low-risk patients for CRC. Risk scores represent model-estimated risk rather than absolute CRC risk. **C. SHAP summary plot illustrating the magnitude and direction of each feature's contribution to the model output-** corresponds to the model denoted by '†' in table 2. Only the top 50 features are shown here. Values on the left represent the mean absolute SHAP value for the corresponding variable. Ethnicity scores were excluded from the plot, as categorical values would lack interpretability. The terminology 'lag0' refers to symptoms/blood tests recorded in the 0-5 months prior to diagnosis, 'lag1' to 6-11 months, 'lag2' to 12-17 months and 'lag3' to 18-23 months. For 'sex1', 0 = female and 1 = male.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData.docx