Genetics, primary care records and lifestyle factors for short-term dynamic risk prediction of colorectal cancer: prospective study of asymptomatic and symptomatic UK Biobank participants

Samantha Ip*[123], Hannah Harrison*[1], Juliet A. Usher-Smith[4], Matthew Barclay[5], Jonathan Tyrer[1], Joe Dennis[1], Xin Yang[1], Michael Lush[1], Cristina Renzi[5,6], Nora Pashayan[1], Spiros Denaxas[7,8,9], Georgios Lyratzopoulos[5], Antonis C. Antoniou[1], Angela Wood[2,3,10,11]

[1] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[2] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge UK
[3] Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge UK
[4] The Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[5] Epidemiology of Cancer Healthcare & Outcomes (ECHO) Group, Department of Behavioural Science & Health, Institute of Epidemiology and Health Care (IEHC), University College London, UK
[6] University San Raffaele, Vita Salute, Milan, Italy
[7] Institute of Health Informatics, University College London, London, UK
[8] National Institute for Health Research University College London Biomedical Research Center, London, UK
[9] British Heart Foundation Data Science Center, Health Data Research UK, London, UK
[10] National Institute for Health and Care Research Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge, UK
[11] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
Cambridge Centre of Artificial Intelligence in Medicine

* Contributed equally.

ABSTRACT

Objectives: To quantify the contributions of polygenic scores, primary care records (presenting symptoms, medical history and common blood tests) and lifestyle factors, for short-term risk prediction of colorectal cancer (CRC) in both all and symptomatic individuals.

Design: Prospective cohort study.

Setting: UK Biobank with follow-up until 2018.

Participants: All participants with linked primary care records (n=160,526), and a subcohort of participants with a presentation of a symptom associated with CRC (n=50,728).

Main outcome measures: Outcome was the first recorded CRC diagnosis within two years. Dynamic risk models with time-varying predictors were derived in a super-landmark framework. Contributions to model discrimination were quantified using novel inclusion-order-agnostic Shapley values of Harrel's C-index using cross-validation.

Results: C-indices [95% CIs] were 0.74 [0.72-0.75] and 0.71 [0.67-0.77] for the models derived in all and symptomatic participants respectively. The Shapley contributions to model discrimination differed between the two groups of participants for different predictors: 31% (32% in the symptomatic participants) for core predictors (e.g., age, sex, smoking), 16% (12%) for polygenic scores, 27% (30%) for primary care blood tests, 14% (14%) for primary care medical history, 8% (0.5%) for additional lifestyle factors and 4% (12%) for symptoms.

Conclusions: Polygenic scores contribute substantially to short-term risk prediction for CRC in both general and symptomatic populations; however, the contribution of information in primary care records (including presenting symptoms, medical history and common blood tests) is greater. There is, however, only a small contribution by the additional lifestyle risk factors which are not routinely collected in primary care.

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer worldwide, with incidence expected to rise (1). Although early stage diagnosis improves survival - five-year survival rates for patients in England diagnosed with stage I and IV CRC are 91% and 11% respectively (2) - more than half of incident CRC cases are diagnosed at late stage (57%). Around 90% of CRC cases present with symptoms in primary care prior to diagnosis, however, around one in five of these patients go on to be diagnosed via emergency presentation (2,3). The widespread use of CRC screening has increased both early-stage diagnosis and survival rates, however, the programme is constrained by cost-effectiveness considerations and endoscopic capacity (4,5). Models that can predict the likelihood of a CRC diagnosis could support referral decisions for symptomatic patients in primary care (6) or identify high-risk people in the general population to prioritise for screening (e.g. earlier or more frequently) (7).

A range of individual level characteristics (including demographics and lifestyle factors) are associated with CRC risk. Additionally, the use of genetic predictors, such as polygenic scores (PGS) which characterise cancer predisposition, hold substantial potential (8). Healthcare systems have already adopted cancer risk assessments that incorporate genetic predictors (9) for example, the BOADICEA breast cancer model is currently under evaluation for potential implementation in a range of healthcare settings (10,11). Further, electronic health records, which encompass symptoms, blood test results and medical history, are well suited for developing risk assessment tools that can be implemented into healthcare systems. The longitudinal structure of these records supports the dynamic evaluation of risk, enabling periodic updates to predictors and risk as new information becomes available for each individual.

Previous models for predicting the risk of CRC diagnosis have incorporated genetic alongside phenotypic risk factors (12) or utilised electronic health records (13,14). However, no existing CRC model has fully harnessed the longitudinal information contained in electronic health records (7,15) in conjunction with genetics for short-term CRC diagnosis (16). Moreover, when evaluating the role of genetic risk in model performance for early cancer detection, previous studies have typically focused on the incremental improvement achieved by adding a PGS to a pre-existing model (12,17); we expect that this “inclusion-order dependent” approach underestimates the contribution of genetic risk factors.

In this study, we derive and internally validate a dynamic prediction model for the diagnosis of CRC using individual-level data from participants in UK Biobank (UKB), incorporating demographics, PGS, primary care data (presenting symptoms, medical history, and common blood tests) and additional lifestyle factors. We aimed to quantify the inclusion-order-agnostic contribution of these six predictor types to model performance in both a general and symptomatic population.

METHODS

Study population

We used data from UKB, a prospective population-based cohort (n=502,371) of UK residents aged 40-69 at enrolment (2006-2010) (18). All participants attended a baseline assessment that collected detailed information about demographics, medical history and lifestyle. Blood samples were taken and genotype data are available for 488,377 participants. Primary care records, including coded information from GP consultations and prescriptions, are available for approximately half of the cohort (n=228,913) with records available up to 2018.

Analyses were restricted to UKB participants with available genetic data and linked primary care data. First degree relatives were excluded by random selection of one of each pair of relatives. Participants with a diagnosis of any cancer, except non-melanoma skin cancer, before baseline assessment were also excluded, ensuring that identified cases of CRC in follow-up were incident primary cancers (Figure S1).

Super-landmark framework

We structured the cohort into landmark age datasets at landmark ages 40, 41, 42, … up to 74 years. Participants were included in a landmark age dataset if they were alive, had at least six months of continuous primary care records (no gaps >90 days; where multiple continuous periods are available we used the most recent) in the previous two years and had not previously received a cancer diagnosis (except non-melanoma skin cancer) before the landmark age. Participants could be included in multiple landmark age datasets as they aged through the cohort. The date when a participant entered a landmark age dataset (1st of the month in which they reach a landmark age) is referred to as the “index date”. Time-varying predictors and outcomes (defined below) were extracted for each participant at all their index dates. The landmark age datasets were stacked to make a “super-landmark dataset” subsequently referred to as the “study cohort” (19,20). For further details see Supplementary Methods 1.

Outcomes

Data on CRC incidence were available through linkage to national cancer registries (provided by: Medical Research Information Service, National Cancer Intelligence Network, NHS England, NHS Central Register, Scottish Cancer Registry and Public Health Scotland) - available up to June 2022 - using ICD codes (see GitHub repository). CRC diagnosis events (cases) were defined as individuals with a first cancer diagnosis of CRC in the two years following an index date. Follow-up was censored at the first diagnosis of any incident cancer (excluding non-melanoma skin cancer), death (via linkage to death registry), end of primary care data availability or two years after the index date.

Risk predictors

We selected candidate predictors as follows: (i) well-established CRC risk factors (21–23); (ii) systematic literature reviews of CRC risk models (6,7); (iii) analyses of CRC symptoms (24,25), pre-existing conditions (both modifiers of CRC risk and potential misdiagnoses) (26,27) and blood test results (28); and (iv) consultation with clinical and academic experts.

Some candidate predictors, such as family history and smoking behaviour, were collected during the UKB baseline assessment. There is missing data for a small number of variables drawn from baseline assessment (Supplementary Methods 2); missing data is included as a separate category for both smoking status and ethnicity, however, for all other variables with missing data (e.g. alcohol consumption) we used a complete case approach (Figure S1). We used the multi-ancestry PRS-CSx, a PGS developed in a cohort of individuals with European or East Asian ancestry to characterise genetic risk of CRC (29). Candidate predictors drawn from primary care records (medical history, symptoms and blood tests) were extracted respective to each index date. We included candidate predictors for both the occurrence and results of blood tests that could indicate iron deficiency anaemia or inflammation.

We grouped the predictors into six sets for analysis (Table S1): core (8 predictors), polygenic score, symptoms (16 predictors), medical history (12 predictors), common blood tests (4 predictors) and other lifestyle (5 predictors). See Supplementary Methods 2 for further details of predictor definitions and derivation.

Defining a Symptomatic Subcohort

We defined a symptomatic super-landmark cohort (subsequently referred to as the "symptomatic subcohort"), by identifying individuals in the study cohort with presenting symptoms relevant to CRC recorded in their primary care records in the two years prior to the landmark age. The relevant symptoms were selected from the 16 predictors in the symptom predictor set using the Akaike information criterion (AIC) based bidirectional stepwise selection with a Cox proportional hazard (Cox PH) model, resulting in the selection of four symptoms: new onset haemorrhoids, new onset constipation, recent record of rectal bleeding and recent record of diverticular disease (see Table S1, Fig S3, Supplementary Methods 2).

Model Development and Validation

We developed dynamic early detection models for CRC diagnoses within two years of an index date. A Cox PH model was developed using the super-landmark dataset with robust standard errors to account for the inclusion of participants multiple times. We selected statistically important predictors using AIC-based bidirectional stepwise selection

(Supplementary Methods 3). After predictor selection, we used 10-fold cross validation to assess model discrimination with Harrell's C-index and calibration with calibration decile plots at 2 years. We computed Shapley values for each of the six predictor sets using the mean C-index from cross-validation, in order to assess the contributions to model discrimination by the six predictor sets. We present the Shapley values as the percentage contribution of each predictor set to the overall C-index; they are independent of the order in which the sets are added to the model (order-agnostic) (28–30). For further details see Supplementary Methods 4.

Model development and validation is reported in line with the TRIPOD guidelines for reporting of clinical prediction models (Supplementary Methods 5)(30).

Sensitivity Analyses

We conducted the following sensitivity analyses: (i) repeating the predictor selection using the LASSO algorithm and comparing both methods to a model with all candidate predictors (no selection); (ii) repeating the analyses using a PGS (LDPred PGS (31)) developed in a European-ancestry only cohort (we note this may overestimate model performance and the genetic contribution as UKB cohort members were included in LDPred development); (iii) redefining the symptomatic subcohort as participants with any CRC symptom (listed in Table S1), except fatigue (not considered to be sufficiently specific), in the two years prior to each index date, and (iv) repeating the analysis excluding the 14,747 participants with English Vision primary care data (linked primary care records for participants registered English Vision who died pre-2017 are mostly unavailable) [(32)].

Patient and public involvement

The Centre for Cancer Genetic Epidemiology has a panel of 6 patients and members of the public (3 with personal cancer experience and 3 without) to advise on all aspects of the cancer risk modelling research incorporating/relating to genetic predisposition to cancer undertaken by members of the centre, under the CanRisk programme of work. They are involved in reviewing the research protocols, the co-design of research and provide feedback on research findings.

RESULTS

Cohort Characteristics

We identified 160,526 UKB participants with genetic data and linked primary care records (Figure S1); their characteristics were similar to the whole UKB cohort (Table S7). Within this study cohort we identified 43,701 symptomatic participants (the symptomatic subcohort). The characteristics of participants in the study cohort and symptomatic subcohort had some differences (Table 1); individuals in the symptomatic subcohort were

older (59.5 (IQR: 52.0-64.2) compared to 57.9 (IQR: 50.3-63.4) in the study cohort), more likely to be male (47.3% and 43.9%) and more likely to have ever smoked (45.4% and 44.4%). The most commonly recorded CRC symptoms in the study cohort were new-onset constipation (16.8%) and new-onset diarrhoea (15.3%). Most participants (76%) had at least one relevant medical history indicator, including: colonoscopy in the last ten years (23.6%) and regular use of non-steroidal anti-inflammatory drugs (NSAIDs) (14.8%). Most participants (87%) also had at least one recent blood test that could be used to determine the presence of inflammation (such as C-Reactive Protein) or iron-deficiency anaemia (such as ferritin), with 33% having at least one abnormal result. More symptoms, indicators of medical history and common blood tests were observed for older participants (Figure S2). We identified 2419 CRC diagnosis events (cases) in the study cohort and 341 in the symptomatic subcohort (Table S5).

Model Development

In the study cohort, 25 (of 39) candidate predictors were selected as important risk predictors of CRC (Figure 1a, Table S2a). This included predictors from each of the six predictor sets. The strongest predictors of CRC (hazard ratio (HR) [95% CI]) other than age were: iron-deficiency anaemia (3.8 [3.3-4.4]) and rectal bleeding (2.7 [2.0-3.7]). The PGS also had a strong association with CRC (1.4 [1.3-1.5] per SD increase).

In the symptomatic subcohort, 19 (of 39) candidate predictors were selected (Figure 2a, Table S2b), 16 of which had also been selected in the study cohort. The 3 distinct predictors were abdominal bloating, diverticular disease and education level. Predictors exhibiting the strongest associations with CRC were: iron-deficiency anaemia (3.9, [2.8-5.4]), male sex (2.0, [1.5-2.7]) and rectal bleeding (2.0, [1.4-2.9]). Again, the PGS had a strong association with CRC (1.3 [1.2-1.5] per SD increase).

Model Performance

The overall discriminative ability (C-index [95% CIs]) of the CRC early detection models developed in the study cohort and symptomatic subcohort was 0.74 [0.72-0.75] and 0.72 [0.67-0.77] respectively. Calibration decile plots at 2 years indicated good calibration for both the study cohort and symptomatic subcohort (Figure S9a-b).

Predictors in the core set (demographics and smoking) made the greatest contribution (31%) to the discriminative ability of the models in the study cohort (Figure 1, Table S3a), followed by common blood tests (27%), PGS (16%) and medical history (13%). These predictor sets by themselves had discriminatory ability of 0.67 [0.65-0.69], 0.63 [0.61-0.65], 0.60 [0.58-0.61] and 0.59 [0.58-0.61] respectively. Selected predictors in the symptom group made the smallest contribution to overall predictive ability (3.6%). The version of the model with only the four highest contributing predictor sets (i.e., without additional lifestyle factors and symptoms) had similar discrimination to the full model (0.73 [0.71-0.75]).

In the symptomatic subcohort the largest contributions to discriminative ability were made by the core predictors (32%), primary care blood tests (30%) and medical history (14%) (Figure 2, Table S3b). The contribution of symptoms was more than three times higher than in the study cohort model (12%) and was similar to that made by the PGS (12%). Additional lifestyle predictors make the smallest contribution (0.5%); removing this predictor set does not lower overall model discrimination (0.73 [0.67-0.78]).

Sensitivity Analyses

Lasso selected more (and slightly different) predictors than bidirectional stepwise (Figures 5a-b), although the discriminative ability of the models using both selection methods were similar, and also equal to that of the model with all predictors (no selection) in the study cohort (Figures S4).

As expected, replacing the PRS-CSx with LPred PGS resulted in models (in both cohorts) with higher hazard ratios for the PGS (HR per SD increase was 2.3 [2.1-2.4] in the study cohort and 2.1 [1.8-2.4] in the symptomatic subcohort), higher discrimination (C-index was 0.79 [0.78-0.80] in the study cohort and 0.76 [0.72-0.81] in the symptomatic subcohort), and with larger contributions from the PGS to the discriminative ability (40% in study cohort and 38% in the symptomatic subcohort) (Table S5, Figures S6a-b).

Using any CRC presenting symptoms except fatigue to define the symptomatic individuals resulted in a larger symptomatic subcohort (n=70,241) than in the main analysis. The predictors selected and their performance were almost identical, although with a slightly higher discriminative ability (0.73 [0.69-0.77]) (Table S5, Figure S7).

Excluding 14,747 participants with English Vision primary care records did not change the model discrimination or the ranking of the predictor sets in either the study cohort or symptomatic subcohort (Table S5, Figure S8c), although the symptom set of predictors makes a larger contribution to model performance in the symptomatic subcohort for this sensitivity analysis (18.7% compared to 12.4% in the main analysis).

DISCUSSION

Summary

In this study we have quantified the contributions of polygenic scores, primary care records (including presenting symptoms, medical history and common blood tests) and lifestyle factors for the short-term risk prediction of colorectal cancer (CRC) amongst participants in UK Biobank. Polygenic scores make substantial contributions to short-term risk prediction for CRC in both a representative sample of UK Biobank participants and a symptomatic sub-cohort (16% and 12% respectively); however, the contribution of information in primary care records (including presenting symptoms, medical history and common blood tests) is greater; the large contribution of common blood tests to the discriminatory ability

of the models (27% and 30% respectively) is particularly of note. In contrast, the additional lifestyle risk factors make only a small contribution to the models, especially in the symptomatic subcohort (0.5%).

This study also presents novel applications of several cutting-edge methodological approaches to clinical prognostic modelling, including the use of the super-landmark framework (to incorporate longitudinal data from primary care records) and Shapley values (to quantify the contributions of six predictor types to model discrimination).

The role of genetics in short-term CRC risk prediction

Several studies have investigated the possibility of adding genetics to other predictors for longer-term CRC risk prediction. Kachuri et al. measure the change in discrimination when a PGS is added to other predictors (demographic, lifestyle and family history) for CRC over a five year follow-up; the C-index [standard error] increases from 0.686 [0.006] to 0.716 [0.006] (33). Briggs et al. showed that adding a PGS to the QCancer-10 model (includes demographic, lifestyle and medical history predictors) increased discrimination for predicted CRC risk, whereas the C-index (in women) increased from 0.65 [0.63-0.66] to 0.69 [0.67-0.70], with authors concluding that there is no clear justification for including a genetic element in risk-stratification for CRC screening (12). However, in previous studies only the effect of adding a PGS after a phenotypic model had been developed was measured. Evaluation of how genetic information can be used to predict short-term cancer risk has had less attention, although Green et al. found that adding a PGS to age in a symptomatic cohort improved discrimination for prostate cancer from 0.68 [0.65-0.71] to 0.77 [0.74-0.80] for a two year follow-up (17).

In this study, we treat genetic and phenotypic predictors equally, which permits a more objective comparison of their relative contribution. Similar to Briggs et al. (12), a modest improvement in discrimination is seen when a PGS is added to all other included predictors (from 0.72 [0.70-0.74] to 0.74 [0.72-0.75] in the study cohort). However, we demonstrate that genetics makes a substantial contribution to the overall performance of the model (16% and 12% in the study cohort and symptomatic cohort, respectively). In the sensitivity analysis using LDPred, the PGS makes a larger contribution to model performance (40% in the study cohort) making it the most influential predictor set. LDPred may be better optimised for this cohort (development only included individuals with European ancestry), however, we note that there are risks of overfitting (and hence inflation of the genetic component) due to the use of UK Biobank individuals in LDPred development.

The role of primary care records in short-term CRC risk prediction

Risk assessment tools using primary care records are well suited for implementation into clinical practice. The C-index of the model that includes the three primary care data

predictor sets (presenting symptoms, medical history and common blood tests) was 0.69 [0.65-0.73] in the symptomatic subcohort (or 0.71 [0.66-0.76] when combined with the core predictors).

In particular, we note that the common blood test predictor set makes the second largest contribution after the core predictor set to the discriminatory ability of the models in both the study cohort (27%) and symptomatic subcohort (30%). Previous studies have found anaemia (24), raised haemoglobin(34) and inflammatory markers (28,35) to be strongly predictive of diagnosis with colorectal cancer in isolation, with increased levels seen up to 9 months prior to diagnosis (35). Previous models for the early detection of colorectal cancer have included predictors for low haemoglobin levels (13,14), however, they did not consider the wider range of common tests that may indicate the underlying clinical state, for example low ferritin may also indicate iron-deficiency anaemia (details of the derivation of these predictors in Supplementary Methods 2). The large contribution blood tests make to the discriminatory performance of the model demonstrates the value that these common blood tests have for triaging patients for referral in a primary care setting, including as part of a multifactorial risk assessment.

As in other studies (24,34), we find that individual symptoms are strongly predictive of colorectal cancer diagnosis. For example, rectal bleeding has a hazard ratio [95% CI] of 2.7 [2.0-3.7] in the model developed in the study cohort (and 2.0 [1.4-1.9] in the symptomatic subcohort). However, the overall contribution of the symptom predictor set to discriminatory ability is relatively low, especially in the study cohort (3%). This may be due to the relatively low incidence of many of these predictors, especially in the study cohort. For example, only 4% of the cohort have a record of rectal bleeding in the two years before any included landmark age. Additionally, the modelling used relatively long-time scales (e.g. two year lookback period and one year intervals between landmark ages) compared to the expected timeline from symptom reporting to CRC diagnosis (e.g. urgent referral for rectal bleeding via the two week wait pathway); the predictive value of symptoms over short periods (<1 year), therefore, may not be well captured in this analysis and should be explored further.

The role of additional lifestyle information in short-term CRC risk prediction

The predictor set “additional lifestyle” included information about patients (collected during UK Biobank baseline assessment) that is not routinely available in primary care records; these include dietary variables (e.g. red meat consumption) and level of education; note that smoking status, which should be routinely collected in primary care, is included in the core predictor set (36). Three (of 5) were selected for the model in the study cohort where this predictor set made a modest contribution to model discrimination (8.4%), however, in the symptomatic subcohort only one was selected and the contribution to model discrimination was very small (0.5%). Measuring these types of predictors requires data collection via self-reporting (e.g. questionnaire or interview) which is resource-intensive

and prone to recall bias (37,38); given their relatively small contribution to the discriminative ability of the models in this study, this may not be an efficient use of resources when considering implementation of a multifactorial risk assessment for CRC early detection, especially in a symptomatic primary care population.

Strengths and Limitations

Within this study, we have implemented a range of cutting-edge methods to make the best use of available data and address the challenges of prognostic modelling for the early detection of cancer. The super-landmark framework allowed us to maximise the number of cases included in the analysis, include information from across the whole study period (from baseline assessment to the end of primary care linkage) in model development and enabled the dynamic derivation of predictors from electronic health records. The use of Shapley values to measure the inclusion-order-agnostic contribution of the predictor sets is a novel approach in the context of prognostic modelling for clinical outcomes.

Additionally, the characteristics of the UKB cohort - including its size, extensive data collection (including genotyping) and linkage (to national registries and primary care records) - has made the analysis described in this study possible. The use of linked primary care records has permitted the derivation of predictors describing a range of clinical events (symptoms, medical history and common blood tests). All codelists used to identify predictors of interest from primary care records were harmonised between the four coding frameworks used by the data providers, and newly developed codelists were checked by clinical experts.

However, there are several limitations to these analyses. First, UKB participants are not representative (39); with only a small number reporting non-white ethnicity (5.0% in the study cohort) and lower cancer incidence and fewer symptoms than the UK general population (40). Despite these discrepancies, there is evidence that associations between predictors and outcomes (including cancer incidence) may be consistent between UK Biobank and the general population (41). In fact, measured discrimination of the models may be higher in the general population due to greater heterogeneity of some risk predictors (42). External validation in a more representative population would be required to demonstrate generalisability, however, datasets that include both detailed primary care records and genetic data are scarce.

Second, primary care records do not give a comprehensive or unbiased view of the health status of cohort members, given the requirement for patients to seek healthcare and for a clinician to code any event of interest (we note that much information is primary care is recorded as free text which is never made available for research). Coding practice in UK primary care is known to vary by data provider (43,44) and over time (44), which may explain the variation seen in the results (symptomatic subcohort only) in the sensitivity analysis excluding participants with English Vision records. Some promising predictors

(such as faecal immunochemical tests (45)) were not included in the analysis due to low counts in the available primary care records (linkage ends 2016-2018 before widespread use of this test in clinical practice).

Conclusions

We have quantified the contributions of six predictor types to the discrimination of dynamic models predicting the short-term risk of CRC in both a study cohort of UK Biobank participants registered with a primary care provider and a symptomatic subcohort. We included a range of predictors (and data) types - including demographics, genetics, symptoms, medical history, primary care blood tests and lifestyle - and assessed their contributions using an order-agnostic method. We evidence the meaningful independent contribution of genetics and data from primary care records, especially common blood tests, to identify people at high risk of CRC diagnosis in the near future. In comparison, only a small contribution is measured for additional lifestyle risk factors which are not routinely collected in primary care.

We anticipate that the resources developed for this study (available online) - including the novel application of the super-landmark framework (incorporating longitudinal data from primary care records) and Shapley values to clinical prognostic modelling - will be of interest to researchers working with electronic health records.

What is already known on this topic:

- Associations between colorectal cancer diagnosis and a range of individual level characteristics - including demographics, lifestyle, polygenic scores, symptoms, medical history, and primary care blood tests - are well-established.
- Models using a range of predictor types offer a way to identify people at high short-term risk of a colorectal cancer diagnosis, but the relative contribution of the different predictor types to overall model performance is unclear.
- Polygenic scores for colorectal cancer are promising for long-term follow-up (5 or more years), but have not yet been explored for short-term follow-up.

What this study adds:

- Quantifies the contribution of six predictor sets to the short-term prediction of colorectal cancer using a novel application of Shapley values to health care data in both a general and symptomatic cohort.
- Demonstrates the substantial contribution of polygenic scores and data from primary care records (particularly common blood tests) to models for the early detection of colorectal cancer.
- Uses a super-landmark framework to incorporate longitudinal data from primary care records into a robust analysis pipeline for model development and validation.

Ethics approval and consent to participate

The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee (reference number 06/MRE09/ 65), and at recruitment all participants gave informed written consent to participate in UK Biobank and be followed up, using a signature capture device.

Data storage and access for this project was strictly controlled in accordance with the data agreements between the researchers and UK Biobank. Access to this data was approved by UK Biobank under projects 64351 (Risk-stratified early detection and diagnosis of cancer using linked electronic health records (RREDD-EHR)) and 28126 (Validating risk prediction models for common hormonal cancers). Throughout the project the data were stored and accessed only via a secure server at the Centre for Cancer Genetic Epidemiology (DPHPC, University of Cambridge).

Availability of data and materials

This paper uses data from UK Biobank that the authors do not have permission to distribute. Bona-fide researchers can apply for access to this data, including linked primary care records, from UK Biobank https://www.ukbiobank.ac.uk/

Resources developed for this project including codelists and analysis code can be found at [TBC].

Conflict of Interests

The authors have no conflicts of interest to declare.

Funding

The work was supported by the International Alliance for Cancer Early Detection, a partnership between Cancer Research UK (C18081/A31373), Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London, and the University of Manchester. SI is additionally supported by Cancer Research UK (EDDPMA-May22\100062) and HH and MB by CRUK International Alliance for Cancer Early Detection (ACED) Pathway Awards (EDDAPA-2022/100001 and EDDAPA-2022/100002, respectively). GL was supported by a Cancer Research UK (C18081/A18180) Advanced Clinician Scientist Fellowship. CR acknowledges funding from Cancer Research UK Early Detection and Diagnosis Committee (grant number EDDCPJT\100018). JUS is supported by a National Institute of Health Research Advanced Fellowship (NIHR300861). SI and AW are supported by the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014; NIHR203312) [*]. AW and SD are part of the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No 116074. SD is supported by the BHF Data Science Centre, the NIHR-UKRI CONVALESCENCE study, the Longitudinal Health and Wellbeing COVID-19 National Core Study, the BHF Accelerator Award (AA/18/6/24223) and Health Data Research UK.

*The views expressed are those of the author(s) and not necessarily those of CRUK, NIHR, NHSBT or the Department of Health and Social Care.

FIGURES & TABLES

Table 1: Participant characteristics: Study and symptomatic cohorts[1]

Predictor	Level	N (%) for categorical; Median (IQR) for continuous	
		Study cohort (N=160526)	Symptomatic subcohort[2] (N=42788)
Core			
Age (baseline assessment)	Years	57.9 (50.3, 63.4)	59.5 (52.0, 64.3)
Body Mass Index	kg/m^2	26.8 (24.2, 29.9)	27.1 (24.5, 30.5)
Birth year	Years	1951 (1945,1958)	1949(1944,1957)
Ethnicity	White	152531 (95.0%)	40347 (94.3%)
	SE Asian	3258 (2.0%)	1077 (2.5%)
	Black	1732 (1.1%)	543 (1.3%)
	Mixed	881 (0.5%)	231 (0.5%)
	Other	1659 (1.0%)	465 (1.1%)
	Missing	465 (0.3%)	125 (0.3%)
Sex (genetic)	Female	84638 (52.7%)	24050 (56.2%)
	Male	75888 (47.3%)	18738 (43.8%)
Smoking status	Never	88710 (55.3%)	22424 (52.4%)
	Current	16755 (10.4%)	4705 (11.0%)
	Former	54586 (34.0%)	15502 (36.2%)
	Missing	475 (0.3%)	157 (0.4%)
Townsend deprivation score	/	-2.2 (-3.7, 0.4)	-2.1 (-3.6, 0.7)
Lifestyle			
Alcohol consumption	Units/day	1.7 (0.1, 3.6)	1.4 (0.1, 3.3)
Education (highest qualification)[3]	Higher Education	53625 (33.4%)	11822 (27.6%)
	A-level	17842 (11.1%)	4320 (10.1%)
	GCSE	33960 (21.2%)	9347 (21.8%)
	CSE	8314 (5.2%)	2351 (5.5%)
	Vocational	10797 (6.7%)	3153 (7.4%)
	Professional	8508 (5.3%)	2294 (5.4%)
	None	27480 (17.1%)	9501 (22.2%)
Fibre consumption score	Units/day	13.6 (9.9, 17.6)	13.6 (9.8, 17.7)
Processed meat consumption	Servings/week	1.0 (0.5, 3.0)	1.0 (0.5, 3.0)
Red meat consumption	Servings/week	2.0 (1.5, 2.5)	2.0 (1.5, 2.5)
Symptoms*			
Abdominal bloating		3174 (2.0%)	1210 (2.8%)
Abdominal pain		22822 (14.2%)	7331 (17.1%)
Change in Bowel habits		3715 (2.3%)	1516 (3.5%)
Constipation		26984 (16.8%)	26984 (63.1%)
Diarrhoea		24593 (15.3%)	6384 (14.9%)
Diverticular		6589 (4.1%)	6589 (15.4%)
Fatigue		13806 (8.6%)	2607 (6.1%)
Haemorrhoids		16427 (10.2%)	16427 (38.4%)
Irritable Bowel Syndrome		4133 (2.6%)	1986 (4.6%)
Rectal bleed		6652 (4.1%)	6652 (15.5%)
Stomach disorders		6580 (4.1%)	1551 (3.6%)

[1] Unless otherwise specified, sections with * indicates that the predictor level is set by default to “has at least one positive entry in the super-landmark dataframe”.
[2] The "symptomatic" subcohort consists of individuals with any of four selected symptoms during their follow-up: new-onset haemorrhoids, new-onset constipation, recent rectal bleeding and recent diverticular disease (see Table S1, Fig S3, Supplementary Methods 2). While these four symptoms formed the eligibility criteria for inclusion in this subcohort, some patients also presented with additional symptoms.
[3] Or equivalent qualifications.

Medical history*			
Aspirin		7675 (4.8%)	3099 (7.2%)
Colonoscopy in last 10 years		37879 (23.6%)	17772 (41.5%)
Diabetes T2 ever		12481 (7.8%)	3720 (8.7%)
Eligible for bowel cancer screening		103781 (64.7%)	25153 (58.8%)
Family history of bowel cancer	True	17405 (10.8%)	5227 (12.2%)
Family history of breast cancer	True	16200 (10.1%)	4359 (10.2%)
Family history of lung cancer	True	19705 (12.3%)	5658 (13.2%)
Gallbladder calc ever		5386 (3.4%)	1918 (4.5%)
Inflammatory bowel disease ever		3313 (2.1%)	1218 (2.8%)
Multimorbidity Score (residual)	Subject mean	-0.137 (-0.237, 0.139)	0.050 (-0.190, 0.438)
NSAIDs (non-aspirin)		23738 (14.8%)	9164 (21.4%)
Primary care blood tests*			
Inflammation	Abnormal	46968 (29.3%)	16257 (38.0%)
	Measured	138152 (86.1%)	39002 (91.2%)
Iron deficiency	Abnormal	13207 (8.2%)	4959 (11.6%)
	Measured	133768 (83.3%)	39051 (91.3%)
Polygenic Score			
Polygenic Score		0.3 (0.2, 0.4)	0.3 (0.2, 0.4)

Figure 1a Hazard ratios from bidirectional stepwise Cox regression for the **study cohort**; **Figure 1b** Discriminative contribution of predictors using Shapley values (C-index > 0.5); **Figure 1c** C-indices from 10-fold cross-validation for each coalition of predictor sets. Colour-coding indicates the predictor set in all figures.

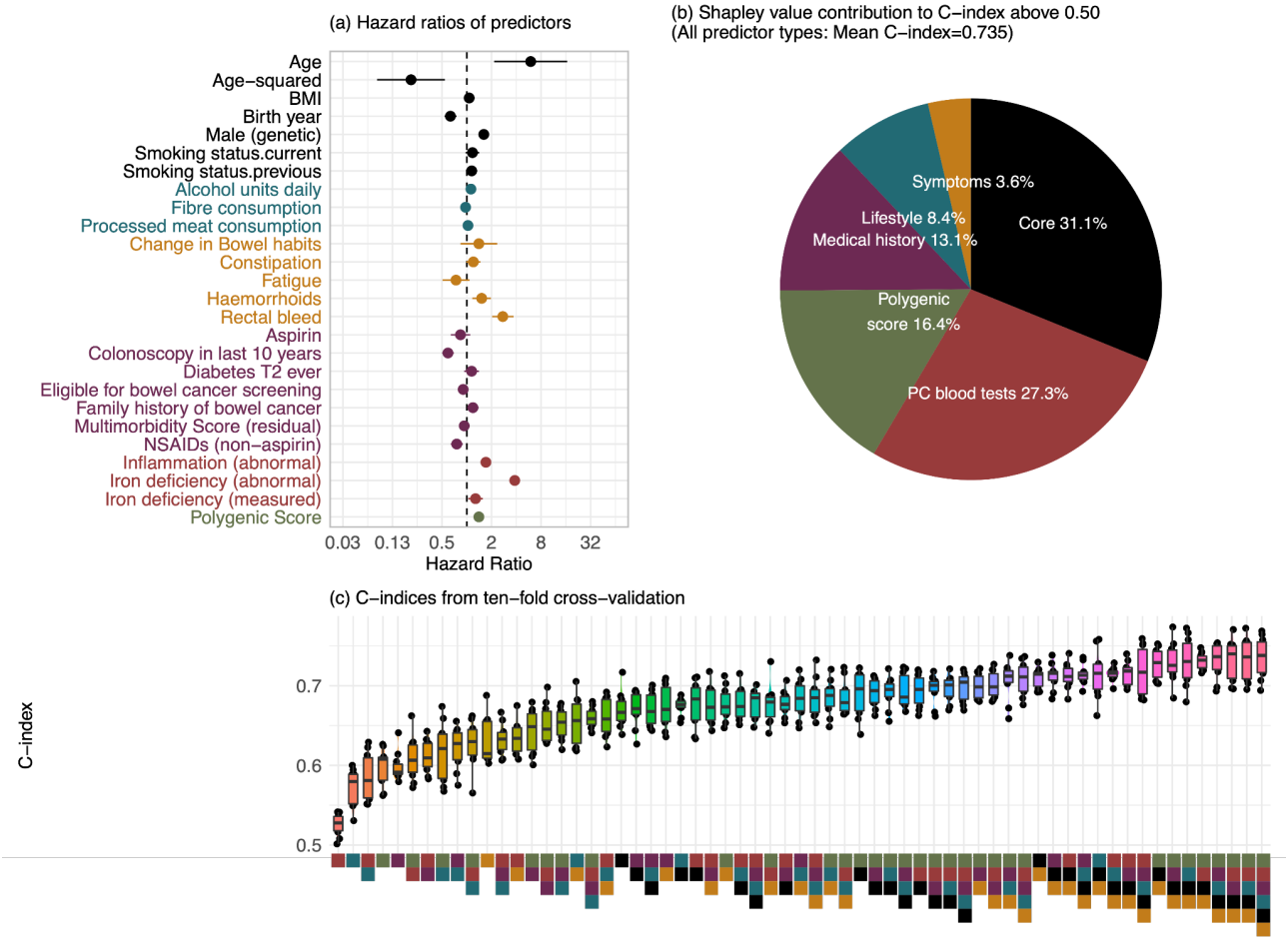
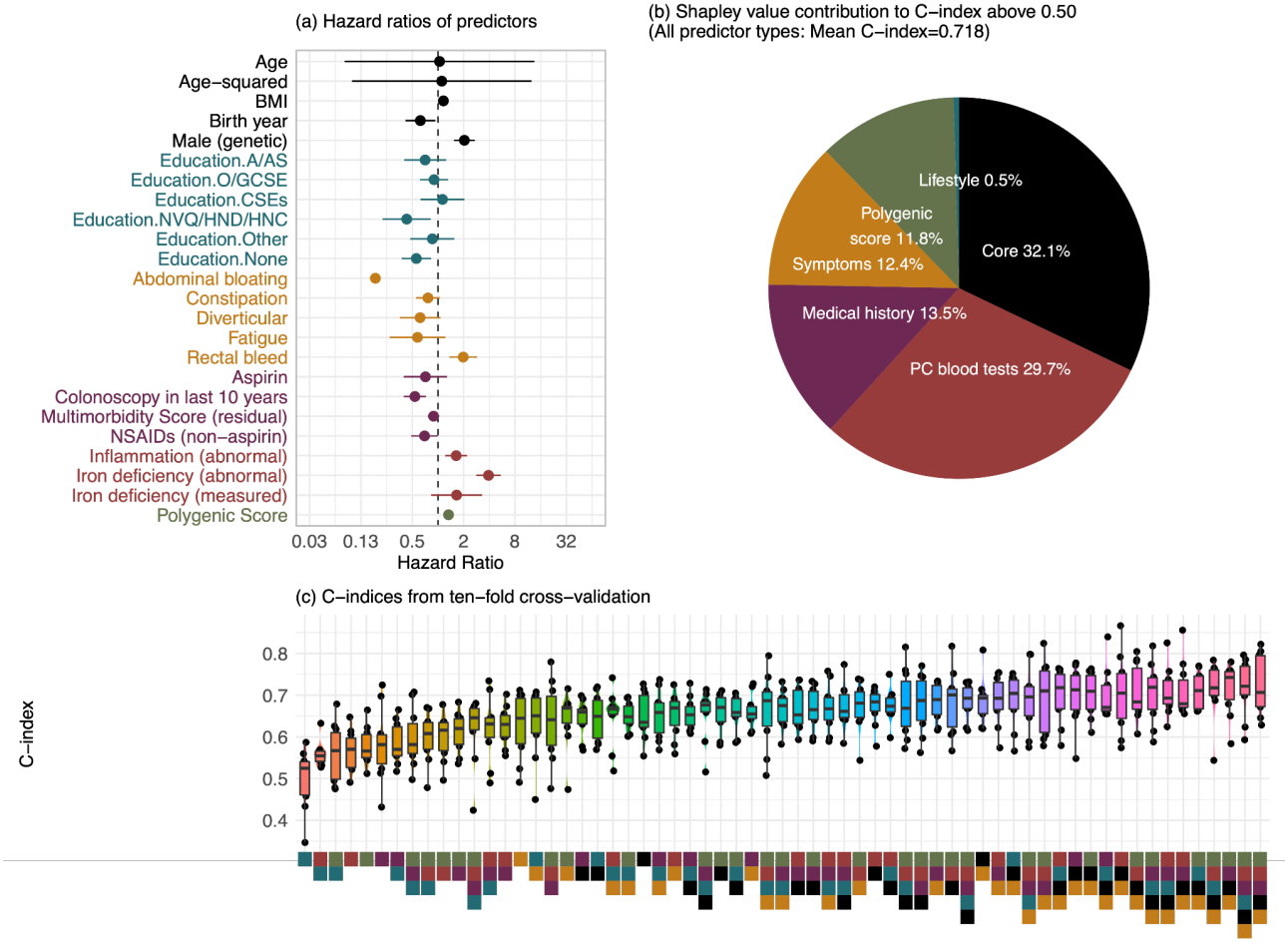


Figure 2a Hazard ratios from bidirectional stepwise Cox regression for the **"symptomatic" cohort**, characterised by haemorrhoids, constipation, rectal bleeding and diverticular disease, as shown in Supplementary Figure 3; **Figure 2b** Discriminative contribution of predictors using Shapley values (C-index > 0.5); **Figure 2c** C-indices from 10-fold cross-validation for each coalition of predictor sets.



REFERENCES

1. Gunter MJ, Alhomoud S, Arnold M, Brenner H, Burn J, Casey G, et al. Meeting report from the joint IARC-NCI international cancer seminar series: a focus on colorectal cancer. Ann Oncol. 2019;30(4):510–9.
2. NHS digital. Cancer Survival in England, cancers diagnosed 2016 to 2020, followed up to 2021 [Internet]. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/cancer-survival-in-england/cancers-diagnosed-2016-to-2020-followed-up-to-2021
3. Renzi C, Lyratzopoulos G, Card T, Chu TPC, Macleod U, Rachet B. Do colorectal cancer patients diagnosed as an emergency differ from non-emergency patients in their consultation patterns and symptoms? A longitudinal data-linkage study in England. Br J Cancer. 2016;115(7):866–75.
4. Logan RFA, Patnick J, Nickerson C, Coleman L, Rutter MD, Wagner C von. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. Gut. 2012;61(10):1439.
5. Gill MD, Bramble MG, Rees CJ, Lee TJW, Bradburn DM, Mills SJ. Comparison of screen-detected and interval colorectal cancers in the Bowel Cancer Screening Programme. Br J Cancer. 2012;107(3):417–21.
6. Williams TGS, Cubiella J, Griffin SJ, Walter FM, Usher-Smith JA. Risk prediction models for colorectal cancer in people with symptoms: a systematic review. BMC Gastroenterol. 2016;16(1):63–63.
7. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. Cancer Prev Res Phila. 2016;9(1):13–26.
8. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. Nat Med. 2021;27(11):1876–84.
9. US Preventive Services Task Force. Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2019;322(7):652–65.
10. Archer S, Fennell N, Colvin E, Laquindanum R, Mills M, Dennis R, et al. Personalised Risk Prediction in Hereditary Breast and Ovarian Cancer: A Protocol for a Multi-Centre Randomised Controlled Trial. Cancers. 2022;14(11).
11. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med Off J Am Coll Med Genet. 01/15 ed. 2019;21(8):1708–18.
12. Briggs SEW, Law P, East JE, Wordsworth S, Dunlop M, Houlston R, et al. Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study. BMJ. 2022;379:e071707.
13. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. Br J Gen Pr. 2013;63(606):e11-21.
14. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract. 2013;63(606):e1.
15. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic

review. J Am Med Inform Assoc JAMIA. 2017;24(1):198–208.
16. Saunders CL, Kilian B, Thompson DJ, McGeoch LJ, Griffin SJ, Antoniou AC, et al. External validation of risk prediction models incorporating common genetic variants for incident colorectal cancer using UK Biobank. Cancer Prev Res (Phila Pa). 2020;canprevres.0521.2019.
17. Green HD, Merriel SWD, Oram RA, Ruth KS, Tyrrell J, Jones SE, et al. Applying a genetic risk score for prostate cancer to men with lower urinary tract symptoms in primary care to predict prostate cancer diagnosis: a cohort study in the UK Biobank. Br J Cancer. 2022;127(8):1534–9.
18. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–9.
19. Paige E, Barrett J, Stevens D, Keogh RH, Sweeting MJ, Nazareth I, et al. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. Am J Epidemiol. 2018;187(7):1530–8.
20. Houwelingen H van, Putter H. Dynamic Prediction in Clinical Survival Analysis. CRC Press; 2011. 250 p.
21. NHS. Symptoms of Bowel Cancer [Internet]. Available from: https://www.nhs.uk/conditions/bowel-cancer/symptoms/
22. NHS. Causes of Bowel Cancer [Internet]. Available from: https://www.nhs.uk/conditions/bowel-cancer/causes/
23. CRUK. Symptoms of Bowel Cancer. Available from: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/symptoms
24. Herbert A, Rafiq M, Pham TM, Renzi C, Abel GA, Price S, et al. Predictive values for different cancers and inflammatory bowel disease of 6 common abdominal symptoms among more than 1.9 million primary care patients in the UK: A cohort study. PLOS Med. 2021;18(8):e1003708.
25. White B, Renzi C, Barclay M, Lyratzopoulos G. Underlying cancer risk among patients with fatigue and other vague symptoms: a population-based cohort study in primary care. Br J Gen Pract. 2023;73(727):e75.
26. Renzi C, Kaushal A, Emery J, Hamilton W, Neal RD, Rachet B, et al. Comorbid chronic diseases and cancer diagnosis: disease-specific effects and underlying mechanisms. Nat Rev Clin Oncol. 2019;16(12):746–61.
27. Koo M, Swann R, McPhail S, Abel G, Renzi C, Rubin G, et al. Morbidity and measures of the diagnostic process in primary care for patients subsequently diagnosed with cancer. Fam Pract. 2022;
28. Watson J, Salisbury C, Banks J, Whiting P, Hamilton W. Predictive value of inflammatory markers for cancer diagnosis in primary care: a prospective cohort study using electronic health records. Br J Cancer. 04/24 ed. 2019;120(11):1045–51.
29. Xin J, Du M, Gu D, Jiang K, Wang M, Jin M, et al. Risk assessment for colorectal cancer via polygenic risk score and lifestyle exposure: a large-scale association study of East Asian and European populations. Genome Med. 2023;15(1):4.
30. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162(1):55–63.
31. Thomas M, Sakoda LC, Hoffmeister M, Rosenthal EA, Lee JK, van Duijnhoven FJB, et al. Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. Am J Hum Genet. 08/05 ed. 2020;107(3):432–44.
32. UK Biobank. UK Biobank: Primary Care Linked Data [Internet]. 2019 [cited 2022].

Available from: https://biobank.ndph.ox.ac.uk/ukb/ukb/docs/primary_care_data.pdf
33. Kachuri L, Graff RE, Smith-Byrne K, Meyers TJ, Rashkin SR, Ziv E, et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. Nat Commun. 2020;11(1):6084.
34. Hamilton W, Lancashire R, Sharp D, Peters TJ, Cheng K, Marshall T. The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. BMC Med. 2009;7(1):17.
35. Moullet M, Funston G, Mounce LTA, Abel GA, de Wit N, Walter FM, et al. Pre-diagnostic clinical features and blood tests in patients with colorectal cancer: a retrospective linked-data study. Br J Gen Pract. 2022;72(721):e556.
36. Szatkowski L, Lewis S, McNeill A, Coleman T. Is smoking status routinely recorded when patients register with a new GP? Fam Pract. 2010;27(6):673–5.
37. Whitton C, Ramos-García C, Kirkpatrick SI, Healy JD, Dhaliwal SS, Boushey CJ, et al. A Systematic Review Examining Contributors to Misestimation of Food and Beverage Intake Based on Short-Term Self-Report Dietary Assessment Instruments Administered to Adults. Adv Nutr. 2022;
38. Tevik K, Bergh S, Selbæk G, Johannessen A, Helvik AS. A systematic review of self-report measures used in epidemiological studies to assess alcohol consumption among older adults. PLoS One. 2021;16(12):e0261292.
39. Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank: Current status and what it means for epidemiology. Health Policy Technol. 2012;1(3):123–6.
40. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017/06/24 ed. 2017;186(9):1026–34.
41. Batty GD, Gale CR, Kivimäki M, Deary IJ, Bell S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. BMJ. 2020;368:m131.
42. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172(8):971–80.
43. Tulloch JS, Beadsworth MB, Vivancos R, Radford AD, Warner JC, Christley RM. GP coding behaviour for non-specific clinical presentations: a pilot study. BJGP Open. 2020;4(3):bjgpopen20X101050.
44. Walker AJ, MacKenna B, Inglesby P, Tomlinson L, Rentsch CT, Curtis HJ, et al. Clinical coding of long COVID in English primary care: a federated analysis of 58 million patient records in situ using OpenSAFELY. Br J Gen Pract. 2021;71(712):e806.
45. Crooks CJ, Banerjea A, Jones J, Chapman C, Oliver S, West J, et al. Understanding colorectal cancer risk for symptomatic patients in primary care: A cohort study utilising faecal immunochemical tests and blood results in England. Aliment Pharmacol Ther. 2023;58(4):443–52.