Development and validation of a risk prediction model to diagnose Barrett's oesophagus (MARK-BE): a case-control machine learning approach

*Avi Rosenfeld, David G Graham, Sarah Jevons, Jose Ariza, Daryl Hagan, Ash Wilson, Samuel J Lovat, Sarmed S Sami, Omer F Ahmad, Marco Novelli, Manuel Rodriguez Justo, Alison Winstanley, Eliyahu M Heifetz, Mordehy Ben-Zecharia, Uria Noiman, Rebecca C Fitzgerald, Peter Sasieni, Laurence B Lovat, on behalf of the BEST2 study group**

Summary

Background Screening for Barrett's oesophagus relies on endoscopy, which is invasive and few who undergo the procedure are found to have the condition. We aimed to use machine learning techniques to develop and externally validate a simple risk prediction panel to screen individuals for Barrett's oesophagus.

Methods In this prospective study, machine learning risk prediction in Barrett's oesophagus (MARK-BE), we used data from two case-control studies, BEST2 and BOOST, to compile training and validation datasets. From the BEST2 study, we analysed questionnaires from 1299 patients, of whom 880 (67·7%) had Barrett's oesophagus, including 40 with invasive oesophageal adenocarcinoma, and 419 (32·3%) were controls. We randomly split (6:4) the cohort using a computer algorithm into a training dataset of 776 patients and a testing dataset of 523 patients. We compiled an external validation cohort from the BOOST study, which included 398 patients, comprising 198 patients with Barrett's oesophagus (23 with oesophageal adenocarcinoma) and 200 controls. We identified independently important diagnostic features of Barrett's oesophagus using the machine learning techniques information gain and correlation-based feature selection. We assessed multiple classification tools to create a multivariable risk prediction model. Internal validation of the model using the BEST2 testing dataset was followed by external validation using the BOOST external validation dataset. From these data we created a prediction panel to identify at-risk individuals.

Findings The BEST2 study included 40 diagnostic features. Of these, 19 added information gain but after correlation-based feature selection only eight showed independent diagnostic value including age, sex, cigarette smoking, waist circumference, frequency of stomach pain, duration of heartburn and acidic taste, and taking antireflux medication, of which all were associated with increased risk of Barrett's oesophagus, except frequency of stomach pain, with was inversely associated in a case-control population. Logistic regression offered the highest prediction quality with an area under the receiver-operator curve (AUC) of 0·87 (95% CI 0·84–0·90; sensitivity set at 90%; specificity of 68%). In the testing dataset, AUC was 0·86 (0·83–0·89; sensitivity set at 90%; specificity of 65%). In the external validation dataset, the AUC was 0·81 (0·74–0·84; sensitivity set at 90%; specificity of 58%).

Interpretation Our diagnostic model offers valid predictions of diagnosis of Barrett's oesophagus in patients with symptomatic gastro-oesophageal reflux disease, assisting in identifying who should go forward to invasive confirmatory testing. Our predictive panel suggests that overweight men who have been taking antireflux medication for a long time might merit particular consideration for further testing. Our risk prediction panel is quick and simple to administer but will need further calibration and validation in a prospective study in primary care.

Funding Charles Wolfson Charitable Trust and Guts UK.

Introduction

Oesophageal cancer has a long-term survival rate of only 12%, but 59% of cases are preventable.[1] Early diagnosis is crucial to change disease outcome but symptoms in early oesophageal adenocarcinoma are often either absent or indistinguishable from uncomplicated gastro-oesophageal reflux disease. Barrett's oesophagus is the only known precursor lesion to oesophageal adenocarcinoma, increasing the risk by 30–60 times.[2] Nevertheless, the annual incidence of oesophageal adenocarcinoma in patients with Barrett's oesophagus is low—eg, approximately 0·1–0·2% among people from the Netherlands[3]—and therefore the merits of endoscopic screening are controversial. The cytosponge test is less invasive and might add an important triaging step because it can be administered in general practice and is more acceptable to patients (cytosponge device was designed by RCF and her research team in 2009–10; patents and a trademark were filed in 2010 by the UK Medical Research Council [MRC]; cytosponge was specifically designed for the BEST2 study in 2010; in 2013, the MRC licensed the technology to Covidien GI

Lancet Digital Health 2020; 2: e37–48

Published Online December 5, 2019 https://doi.org/10.1016/S2589-7500(19)30216-X

See Comment page e6

*Listed in the appendix

Department of Industrial Engineering (A Rosenfeld PhD) **and Department of Health Informatics** (E M Heifetz PhD, M Ben-Zecharia PhD, U Noiman PhD)**, Jerusalem College of Technology, Jerusalem, Israel; GENIE GastroENterological IntervEntion Group, Department for Targeted Intervention, University College London, London, UK** (A Rosenfeld, D G Graham MBBS, S Jevons PhD, J Ariza RGN, D Hagan MSc, A Wilson BSc, S J Lovat, S S Sami MBBS, O F Ahmad MBBS, Prof L B Lovat MBBS)**; Gastrointestinal Services** (D G Graham, J Ariza, S S Sami, O F Ahmad, Prof L B Lovat) **and Department of Pathology** (Prof M Novelli MBChB, M Rodriguez Justo MBBS, A Winstanley MBBS)**, University College London Hospital, London, UK; MRC Cancer Unit, University of Cambridge, Cambridge, UK** (Prof R C Fitzgerald MBChB)**; Cancer Prevention Trials Unit, Queen Mary University of London, London, UK** (Prof P Sasieni PhD)**; and School of Cancer and Pharmaceutical Sciences, King's College London, London, UK** (Prof P Sasieni)

Correspondence to: Prof Laurence B Lovat, Division of Surgery and Interventional Science, University College London, London W1W 7TS, UK **l.lovat@ucl.ac.uk**

See Online for appendix

Research in context

Evidence before this study

We searched PubMed for publications in English from database inception until June 30, 2019, on models to identify the presence of Barrett's oesophagus using the terms "Barrett's esophagus", "prediction model", "risk factors", and "risk prediction models". Previous studies have identified multiple risk factors but—with two recent exceptions, one involving a small cohort of patients and the other looking at familial risk—they have either not synthesised the model to create a comprehensive risk factor panel or have not validated it in a completely independent dataset.

Added value of this study

Here we took two large datasets (BEST2 and BOOST) that together include more than 1600 patients and controls. We used robust machine learning methods to create a stable algorithm to predict the presence of Barrett's oesophagus from the BEST2 cohort. These algorithms were tested internally in a separate subset of the cohort and then validated externally in the BOOST cohort. A reliable and stable risk prediction panel was created, comprising eight risk factors, that can now be prospectively tested in a primary care cohort.

Implications of all the available evidence

A successful risk prediction panel would, for the first time, potentially allow routine non-invasive identification of patients who are at high risk of having Barrett's oesophagus. The machine learning approach we used to develop this risk prediction panel could be used for other medical conditions to aid diagnosis and avoid unwarranted and low yield invasive testing.

Solutions, now part of Medtronic [Dublin, Ireland]).[4] The cytosponge is in a capsule on a string that is swallowed by a patient. The capsule disintegrates upon entering the stomach and the sponge unravels within 5 min. The sponge is then pulled back out of the mouth using the string and as it travels out of the body it picks up cells from the lining of the oesophagus. These cells can then be tested for the presence of Barrett's oesophagus. Another alternative endoscopic test uses a video capsule that photographs the gut; however, because the capsule traveses the oesophagus very quickly, this test is not very useful for Barrett's oesophagus.[5] Therefore, an important question is which patients with suspected Barrett's oesophagus should be screened with these tests.

Obvious target groups would have symptoms and known risk factors. These include age, sex, race, reflux symptoms, obesity, cigarette smoking, family history, and use of anticholinergic drugs.[6,7] We previously tried to identify patients at risk by analysing these factors using statistical approaches, with relatively poor success.[8] Therefore, whether targeting these groups would work in clinical practice is unclear.

Machine learning applies mathematical models to generate computerised algorithms, which can create novel prediction models. Machine learning involves a computer that learns important features of a dataset to enable predictions about other unseen data. This approach can be particularly useful to create models to predict which individuals have a disease.[9]

We hypothesised that machine learning could yield better and more reproducible discrimination between patients with and without Barrett's oesophagus than other statistical models. Previous studies in this area have not validated their results[10,11] or found large reductions in model accuracy in validation cohorts.[12] Additionally, most previous studies focused on only a few symptoms, making between-study comparisons difficult. The risk factors identified include older age,[13] male sex,[12,14] Caucasian race,[15] gastro-oesophageal reflux disease,[13,16] smoking,[17,18] and central obesity (ie, high waist circumference).[18] Only two studies considered all of these factors together, of which one included only 235 patients with Barrett's oesophagus[19] and the other focused on familial disease.[20] Here, we used a large dataset to train and then test a model for detection of Barrett's oesophagus. We added an additional independent validation dataset to confirm the robustness of the tool to prescreen patients for this condition.

Methods

Study design and participants

In this prospective study, machine learning risk prediction in Barrett's oesophagus (MARK-BE), we collected data from two case-control studies done in the UK to construct training, testing, and external validation datasets. We collected data on patients with Barrett's oesophagus and controls, both as defined in the inclusion criteria of the studies. All patients with a diagnosis of dysplastic Barrett's oesophagus or oesophageal adenocarcinoma were included in the Barrett's oesophagus group and those with ultra-short segment Barrett's oesophagus (Prague classification of less than C1Mx or C0M3) were removed from the analysis completely to create a clear distinction between the groups.

BEST2 (ISRCTN 12730505) was a case-control study undertaken nationwide in 14 UK hospitals, with patients recruited in 2011–14, that compared the accuracy of the cytosponge-trefoil factor 3 test for the detection of Barrett's oesophagus with endoscopy and biopsy as the reference standard.[4,21] Barrett's oesophagus was defined as endoscopically visible columnar-lined oesophagus (Prague classification C1 or M3), with histopathological evidence of intestinal metaplasia on at least one biopsy sample. Controls were symptomatic patients without Barrett's oesophagus referred for routine endoscopy. Of 1299 patients, 880 (67·7%) had Barrett's oesophagus, 40 (3%) had invasive oesophageal adenocarcinoma, and 419 (32·3%) were controls. In parallel to assessing the

accuracy of the cytosponge test, patients were asked to complete a questionnaire giving details of 40 symptoms and risk factors of their condition to analyse whether these symptoms and risk factors could be used to stratify patients by risk, such as we have done previously.[8] Questionnaire data were collected from all 1299 participants. For the current study, we randomly split this large dataset (6:4) using a computer algorithm into a training dataset (n=776) and a testing dataset (n=523). We split the dataset using this ratio to allow sufficient training data to quantify the model's complexity while maintaining adequate data to validate the model.

BOOST (ISRCTN 58235785) was a case-control study undertaken in four European hospitals (two in the UK in London and Nottingham, one in Leuven, Belgium, and one in Madrid, Spain), with patients recruited in 2013–15, that used enhanced endoscopic techniques to target high-risk lesions that occur in patients with Barrett's oesophagus.[22] Clinical and demographic data were collected. Controls were patients referred by their primary care physician with suspected oesophageal cancer who had neither Barrett's oesophagus nor oesophageal adenocarcinoma and were analogous to those in BEST2. Although BOOST was a multicentre study, questionnaires were only collected from 398 patients at a single site, University College London Hospial, London, UK. 197 (50%) of 398 participants who completed questionnaires were controls and 24 (6%) of 398 had oesophageal adenocarcinoma. Patients were asked to complete a questionnaire similar to that in BEST2. This questionnaire was designed from the outset to include the same questions as in the BEST2 questionnaire so that the cohort could be used as a validation dataset for a symptom-based algorithm that was to be generated from the BEST2 dataset in line with TRIPOD guidelines.[23] However, some extra questions were included relating to food intake, anxiety, and depression. We used this dataset as the external validation dataset.

The primary outcome of both studies was a diagnosis of Barrett's oesophagus, which was ascertained by histopathologists who were masked to predictor variables.

For BEST2, symptoms of gastro-oesophageal reflux disease (GERD) were collected with a questionnaire adapted from the GERD Impact Scale[8] together with the GERD questionnaire.[10] BOOST also included the hospital anxiety and depression scale. The total number of variables reported in BEST2 was 40 and in BOOST was 204. In both studies, data were collected on paper case-report forms and transferred into electronic databases by investigators.

Data handling and machine learning approaches

We imputed missing data for nominal and numerical features with the modes and means of the training data. Here we describe how predictors were handled, and the workflow is shown in figure 1. We used feature analysis to process data and identify important predictors.

For the training dataset, we analysed data using two accepted feature selection filters: information gain and correlation-based feature selection. Information gain is a machine learning univariate filter that compares each feature separately and its correlation with the class. Features are chosen on the basis of how much each one discriminates between the groups being investigated; in our case, Barrett's oesophagus versus no Barrett's oesophagus. Correlation-based feature selection filtering is a multivariable filter that specifically considers features' correlation to each other and removes redundant features that are highly correlated. The final set of features is then used to generate the analysis model.

Both information gain and correlation-based feature selection are filter feature selection methods and thus have the advantage of being fast, scalable, and independent of the classifier.[24] Independence from the classifier is crucial to our study because it allows us to understand which features are being selected by the algorithm and their medical importance. As made clear by Nie and colleagues,[25] filters that are independent of the classifier enable improved interpretability. They should also lead to more stable algorithms than conventional statistical approaches, such as backward logistic regression, because they minimise data overfitting. Similar to our previous work,[8] we initially identified k features that had at least a minimal correlation to Barrett's oesophagus. We then plotted the change in mean area under the receiver operator curve (AUC) for prediction of Barrett's oesophagus using between 1 and k features.

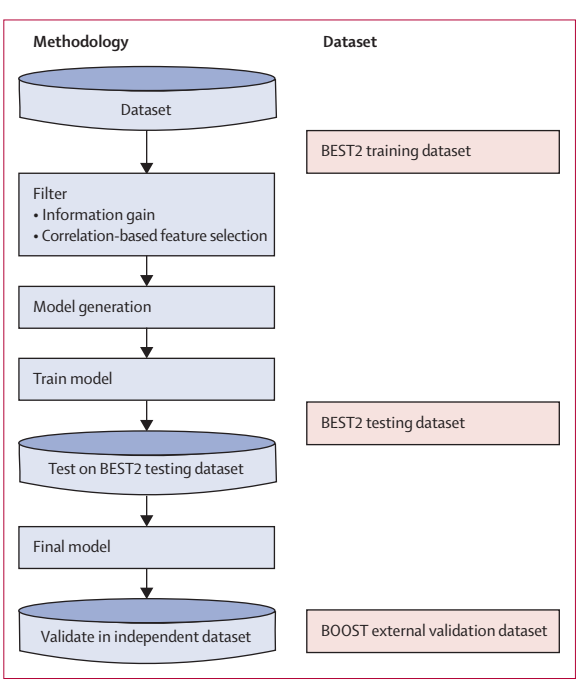

Figure 1: **Machine learning workflow for data processing and model development**

We identified the smallest number of independent features in the BEST2 training dataset to create our model. The smaller the set of predictors, the more stable and robust the model, which minimises the risks of overfitting the data.

Once our features were defined, we considered five different machine learning methods: logistic regression, a decision tree based on the Gini measure of quality, a naive Bayes classifier assuming a Gaussian distribution, a support vector machine using the radial bias function kernel, and a random forest classifier using ten trees. These five algorithms were chosen for comparison because they are well accepted machine learning methods that are typically of use in medical applications.[26–31] The relative strengths and weaknesses of learning algorithms remain a major research topic but in principle, when training data are restricted, simpler models usually perform better because they will generalise more reliably—eg, linear and logistic regression models. Random forest and decision trees usually perform better when training data are abundant and a complex interaction exists between features. Support vector machines can be extremely robust if the number of predictive features is very large compared with the number of training examples; a situation in which overfitting often occurs. Naive Bayes should be preferred over logistic regression if data are sparse but one is confident of the modelling assumptions.[32] We also considered using deep neural networks, but given the lack of dimensionality of our data, these models are substantially less accurate and interpretable.[33–35] Although we considered several options for building a supervised prediction model, unless otherwise specified, we present the results from a logistic regression prediction model.

We developed a prediction model using 90% of the BEST2 training dataset to train and 10% to internally test the model (figure 1). This process was repeated ten times. We used the mean AUC to determine which model performed best, which was then tested with the BEST2 testing dataset. Finally, we validated the model on the BOOST external validation dataset. For the AUC calculation, we set the sensitivity of the model to 90%, because we considered this sensitivity to be a clinically important.

Because the AUC measurements might have restricted accuracy for imbalanced datasets, we calculated precision recall and log loss to show the stability of the derived model. We calculated and present extended metrics for the machine learning application for the training model when applied to the BEST2 testing dataset, and for the BEST2 training model after external validation on the BOOST dataset. We present accuracy, which is the ratio of the correctly labelled participants to the whole dataset; recall, which is equivalent to sensitivity (of all the people with Barrett's oesophagus, how many could we correctly predict?); precision, which is equivalent to positive predictive value (how many of those labelled with Barrett's oesophagus actually have it?) measured at the highest point on the receiver operator curve; and the F-measure, which is the harmonic mean (average) of the precision and recall.

The input datasets included obvious biases, such as different sex prevalences in the Barrett's oesophagus and control groups and duration of symptoms. Patients with Barrett's oesophagus are known to have a higher prevalence of long-term gastro-oesophageal reflux disease.[13,16] Additionally, controls presented with new symptoms whereas those with Barrett's oesophagus were mostly in surveillance programmes. We reconstructed the datasets so that race, sex ratios, and age profiles were similar across all datasets. We also removed all features relating to symptom duration. We then repeated all machine learning with this reconstructed dataset to build a new risk prediction panel. The risk prediction panel was tested on both the BEST2 testing dataset and the BOOST independent validation dataset with the actual diagnoses withheld. Once the panel had predicted the diagnoses, the results were compared with the true diagnoses and the accuracy of the model was then calculated.

Statistical analysis

This Article is reported in alignment with TRIPOD guidelines.[23] No generally accepted approaches exist to estimate sample size requirements for derivation and validation studies of risk prediction models. We used all available data to maximise the power and generalisability of our results. Model reliability was enhanced by our use of an external validation cohort.

We present discrete variables as numbers and percentages and continuous variables as mean (SD). We calculated p values for the association of each factor with presence and absence of Barrett's oesophagus using Student's *t* test or the χ^2 method. We calculated AUCs by generating a univariate logistic regression model using only that feature.

We present the ranked features from the training dataset using regression coefficients of the association of each feasture in the final prediction model. We present the risk of Barrett's oesophagus associated with each feature using odds ratios and 95% CIs.

We did all analyses using the RWeka, cvAUC and pROC packages in R (version 3.6.1).

Role of the funding sources

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. PS had access to all the BEST2 data and LBL and AR had access to all data. The corresponding author had final responsibility for the decision to submit for publication.

Results

Demographic and symptom characteristics for all three datasets are shown in table 1. Patients with Barrett's oesophagus were generally older than those without and

	BEST2 training dataset (n=776)				BEST2 testing dataset (n=523)				BOOST external validation dataset (n=398)			
	Barrett's oesophagus present	Barrett's oesophagus absent	p value	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value	AUC
n	528 (68%)	248 (32%)	..	..	352 (67%)	171 (33%)	..	..	198 (50%)	200 (50%)	..	..
Sex												
Male	436/525 (83%)	105/248 (42%)	<0·0001	0·70 (0·67–0·74)	279/352 (79%)	74/171 (43%)	<0·0001	0·68 (0·64–0·72)	155/197 (79%)	94/199 (47%)	<0·0001	0·66 (0·61–0·70)
Female	91/525 (17%)	143/248 (58%)	..	..	73/352 (21%)	97/171 (57%)	..	..	42/197 (21%)	105/199 (53%)	..	..
Age, years	67·09 (11·99)	61·53 (14·37)	<0·0001	0·61 (0·57–0·66)	66·96 (11·93)	58·94 (15·06)	<0·0001	0·66 (0·61–0·71)	67·49 (11·66)	59·94 (15·38)	<0·0001	0·66 (0·60–0·71)
Waist circumference, cm	101·83 (12·49)	91·87 (13·40)	<0·0001	0·70 (0·66–0·74)	100·04 (12·33)	93·66 (13·51)	<0·0001	0·64 (0·58–0·69)	90·83 (9·91)	86·18 (10·86)	0·0001	0·62 (0·56–0·68)
Cigarettes per day	16·41 (13·33)	10·61 (8·30)	<0·0001	0·63 (0·57–0·68)	16·27 (13·77)	11·26 (9·52)	0·0026	0·63 (0·57–0·68)	32·17 (32·93)	19·74 (17·28)	0·0093	0·66 (0·56–0·75)
Taking antireflux medication												
No	31/525 (6%)	102/243 (42%)	<0·0001	0·68 (0·65–0·71)	23/349 (7%)	69/171 (40%)	<0·0001	0·67 (0·63–0·71)	17/190 (9%)	67/175 (38%)	<0·0001	0·65 (0·61–0·69)
Yes	494/525 (94%)	141/243 (58%)	..	..	326/349 (93%)	102/171 (60%)	..	..	173/190 (91%)	108/175 (62%)	..	..
Stomach pain frequency												
Never	371/525 (71%)	73/243 (30%)	<0·0001	0·73 (0·69–0·76)	238/348 (68%)	66/171 (38%)	<0·0001	0·67 (0·62–0·72)	130/188 (69%)	66/177 (37%)	<0·0001	0·69 (0·64–0·74)
Occasionally*	108/525 (21%)	82/243 (34%)	..	..	70/348 (20%)	46/171 (27%)	..	..	24/188 (13%)	15/177 (8%)	..	..
Weekly	28/525 (5%)	39/243 (16%)	..	..	17/348 (5%)	22/171 (13%)	..	..	19/188 (10%)	42/177 (24%)	..	..
Daily	18/525 (3%)	49/243 (20%)	..	..	23/348 (7%)	37/171 (22%)	..	..	15/188 (8%)	54/177 (31%)	..	..
Time since acidic taste started												
Never	88/525 (17%)	84/243 (35%)	<0·0001	0·75 (0·72–0·79)	48/349 (14%)	49/171 (51%)	<0·0001	0·77 (0·73–0·82)	102/132 (77%)	77/107 (72%)	0·0146	0·51 (0·46–0·57)
≤6 months	8/525 (2%)	43/243 (18%)	..	..	4/349 (1%)	30/171 (88%)	..	..	3/132 (2%)	12/107 (11%)	..	..
7 to <12 months	8/525 (2%)	16/243 (7%)	..	..	3/349 (1%)	16/171 (84%)	..	..	3/132 (2%)	3/107 (3%)	..	..
1 to <2 years	26/525 (5%)	25/243 (10%)	..	..	13/349 (4%)	19/171 (59%)	..	..	2/132 (2%)	4/107 (4%)	..	..
2 to <5 years	52/525 (10%)	25/243 (10%)	..	..	34/349 (10%)	20/171 (37%)	..	..	11/132 (8%)	4/107 (4%)	..	..
5 to <10 years	79/525 (15%)	23/243 (9%)	..	..	59/349 (17%)	15/171 (2%)	..	..	5/132 (4%)	3/107 (3%)	..	..
10 to <20 years	123/525 (23%)	13/243 (5%)	..	..	87/349 (25%)	16/171 (16%)	..	..	0	3/107 (3%)	..	..
≥20 years	141/525 (27%)	14/243 (6%)	..	..	101/349 (29%)	6/171 (6%)	..	..	6/132 (5%)	1/107 (1%)	..	..
Time since heartburn started												
Never	40/525 (8%)	13/243 (5%)	<0·0001	0·75 (0·72–0·79)	28/349 (8%)	11/170 (6%)	<0·0001	0·77 (0·73–0·81)	121/138 (88%)	77/107 (72%)	0·0292	0·57 (0·52–0·62)
≤6 months	4/525 (<1%)	52/243 (21%)	..	..	2/349 (1%)	37/170 (22%)	..	..	3/138 (2%)	12/107 (11%)	..	..
7 to <12 months	7/525 (1%)	23/243 (9%)	..	..	4/349 (1%)	15/170 (9%)	..	..	1/138 (1%)	2/107 (2%)	..	..
1 to <2 years	15/525 (3%)	34/243 (14%)	..	..	12/349 (3%)	25/170 (15%)	..	..	1/138 (1%)	4/107 (4%)	..	..
2 to <5 years	45/525 (9%)	35/243 (14%)	..	..	33/349 (9%)	28/170 (16%)	..	..	5/138 (4%)	4/107 (4%)	..	..
5 to <10 years	90/525 (17%)	37/243 (15%)	..	..	56/349 (16%)	19/170 (11%)	..	..	1/138 (1%)	2/107 (2%)	..	..
10 to <20 years	141/525 (27%)	25/243 (10%)	..	..	87/349 (25%)	25/170 (15%)	..	..	1/138 (1%)	3/107 (3%)	..	..
≥20 years	183/525 (35%)	24/243 (10%)	..	..	127/349 (36%)	10/170 (6%)	..	..	5/138 (4%)	3/107 (3%)	..	..

Data are n (%), n/N (%), or mean (SD), p value, or AUC with 95% CI in parentheses. p values were calculated using the χ^2 test or Student's *t* test and AUCs are calculated for each dataset using the pROC package, which created a logistic regression model for each feature. AUC=area under the receiver operator curve. *Once or twice a week.

Table 1: **Demographic and symptom characteristics in the three datasets, by the presence or absence of Barrett's oesophagus**

	Information gain	Remain in model after correlation-based feature selection	Regression coefficients in final model to predict Barrett's oesophagus*	Odds ratio for Barrett's oesophagus
Taking antireflux medication	0·192	Yes	2·033	7·639 (yes)
Sex	0·133	Yes	1·592	4·901 (male)
Waist circumference	0·107	Yes	0·035	1·035
Duration of heartburn†	0·095	Yes	0·132	1·142
Frequency of stomach pain	0·085	Yes	–0·836	0·433
Duration of acidic taste†	0·074	Yes	0·297	1·345
Age	0·065	Yes	0·034	1·035
Frequency of heartburn	0·062	No	..	..
Ethnicity	0·060	No	..	..
Weight	0·060	No	..	..
Height	0·051	No	..	..
Frequency of sleep disruption	0·049	No	..	..
Body-mass index	0·040	No	..	..
Amount of alcohol drunk at age 30 years	0·036	No	..	..
Frequency of acidic taste	0·031	No	..	..
Education level	0·018	No	..	..
Number of cigarettes smoked	0·016	Yes	0·045	1·046
Ever smoked	0·014	No	..	..
Amount of alcohol drunk currently	0·011	No	..	..

These features offered more than minimal information gain to predict a diagnosis of Barrett's oesophagus. The number of features was reduced by assessing for correlated feature selection. The final eight features were fed into the analytical model. The intercept for the regression equation is –5·031. *To three significant figures. †Years since started.

Table 2: **Ranked features in the BEST2 training dataset**

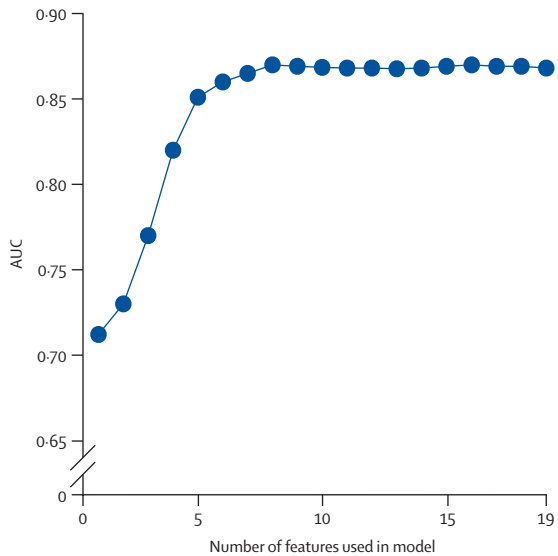

Figure 2: **Performance of the model using the BEST2 training dataset**
Increasing the number of features strengthens the model to a plateau point that is reached around eight features. The model AUC remains unaffected when up to a total of 19 features are added. AUC=area under the receiver operator curve.

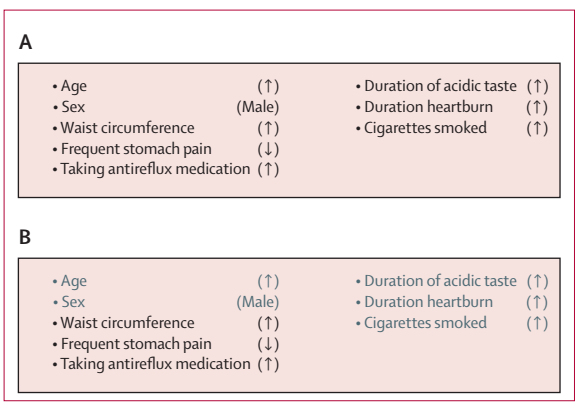

Figure 3: **Risk prediction model panels for Barrett's oesophagus**
(A) The eight features selected by correlation-based feature selection for the BEST2 training dataset, and the direction of association with presence of Barrett's oesophagus. (B) Of the eight features identified, those that are still associated using the correlation-based feature selection model using the reconstructed datasets, excluding potential age, sex, race and symptom duration biases, are shown in black, with those no longer associated in grey. Arrows show the direction of association, with an arrow pointing up indicating an increased likelihood of Barrett's oesophagus.

were also more likely to be male and smokers, had more central obesity, took more antireflux medication, and had less frequent stomach pain. Addtionally, those with Barrett's oseophagus had experienced acidic taste and heartburn for significantly longer than those without.

In the case-control BEST2 training dataset, all cases had a confirmed diagnosis of Barrett's oesophagus. We selected features with a non-negligible information gain. In line with previous work,[36] we used a threshold of 0·01 (ie, above a negligible zero value) to select features that would positively affect the model. Features with a weaker correlation to disease were removed. A total of 19 features were selected (table 2). We sorted these features from highest to lowest information gain correlation with Barrett's oesophagus and considered subsets with the top k features ranging between 1 and 24. We selected the eight features with the highest information gain and found no significant increase in the AUC (p value of the moving average of the next 10 points compared with the original values being 0·7; figure 2). This finding is consistent with the concept that adding features, even those with strong correlation to Barrett's oesophagus (table 2), does not necessarily improve model performance.

We developed multivariable models using correlation-based feature selection based on the entire 24 common features. Correlation-based feature selection selected eight features as independent predictors of Barrett's oesophagus (age, sex, waist circumference, stomach pain, taking antireflux medication, duration of heartburn, duration of acidic taste in the mouth, and smoking; figure 3A). These features were not the same as the top eight features identified with information gain analysis (table 2).

The prediction model was based on the features selected via the correlation-based feature selection analysis (figure 3A). Once we had our small panel of features, we tested the five different machine learning methods and found that logistic regression yielded the best median

AUC, and so we elected to use this model (figure 4). Furthermore, this model is most readily understandable to a medical audience making it easy to convert into a usable tool in clinical practice.

We used the testing dataset to provide an upper estimate of the model's predictive ability (table 2). Using the BEST2 testing dataset, the AUC was 0·87 (95% CI 0·84–0·90) and, for a sensitivity arbitrarily set at 90%, the specificity was 68%.

We validated this model using the BEST2 testing dataset. The model reproduced well, with an AUC of 0·86 (95% CI 0·83–0·89), with sensitivity set at 90%, and specificity of 65%. The model was finally tested on the independent validation BOOST dataset. Here the model achieved an AUC of 0·81 (95% CI 0·74–0·84), with a set sensitivity of 90%, and a specificity of 58%. This three-stage development process led to a stable, reproducible model.

For completeness, we also present the accuracy, recall, precision, and F-measure results of the training model applied to the BEST2 training dataset and the external validation model on the BOOST dataset (table 3). The results were in a relatively narrow range (eg, accuracy 76·88–84·51% and F-measure 0·77–0·84) with the lowest values being recorded when validating the BEST2 model on the BOOST data. These results are consistent with the AUC results.

We repeated our analyses using reconstructed databases to remove potential biases. Reconstructing the cohorts reduced the BEST2 training dataset from 776 to 394 patients; the BEST2 testing dataset from 523 to 297 patients; and the BOOST external validation dataset from 398 to 162 patients (table 4). We used the same workflow to create a new model. We determined the new correlation-based feature selection variables (figure 3B). The same features remain apart from age, sex, and symptom duration. No new features entered the correlation-based feature selection analysis. As for the initial analyses, we selected features with non-negligible information gain, and selected a total of seven features. We then built multivariable models based on correlation-based feature selection on these seven features. Three were selected as independent predictors of Barrett's oesophagus (waist circumference, frequent stomach pain, and taking antireflux medication; figure 3B). The overall accuracies are lower than the original eight features but a clear difference remains between patients with and without Barrett's oesophagus. The initial model had an AUC of 0·84 (95% CI 0·79–0·88; sensitivity 90%, specificity 43%), which decreased to 0·78 (95% CI 0·72–0·84; sensitivity 90%, specificity 41%) after testing internally, and to 0·77 (0·64–0·81; sensitivity 90%, specificity 37%) after external validation.

Most features identified through both iterations of the model are readily understandable such as age, male sex, longer duration of symptoms, taking antireflux medications, and central obesity (ie, waist circumference). However, the feature of lower frequency of stomach pain appears counterintuitive.

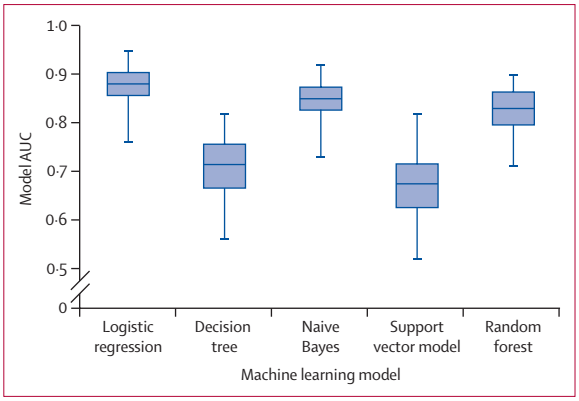

Figure 4: **Comparison of model's AUC with different machine learning classification algorithms**
Box plots show AUCs and 95% CIs. AUCs when using the BEST2 training dataset with 13 features. AUC=areas under the receiver operator curve.

	Accuracy	Recall	Precision	F-measure
BEST2 testing dataset	84·51	0·85	0·84	0·84
BEST2 validated on BOOST	76·88	0·77	0·77	0·77

The first line are these measures when evaluating the model developed with the BEST2 training dataset and tested on the BEST2 testing dataaset. The second line shows the results of these measures for the BEST2 model after external validation on the BOOST dataset.

Table 3: **Extended metrics for evaluating the machine learning application, by dataset**

Discussion

We have shown that a panel with eight features, including detailed stomach and chest symptoms, can identify the presence of Barrett's oesophagus with high sensitivity and specificity in a case-control population. The currently used system for identifying patients with Barrett's oesophagus, or those at risk of oesophageal adenocarcinoma, is flawed because it is based on symptoms that trigger expensive and unpleasant invasive tests. Simple triaging of individuals might be possible on the basis of predictive panels that include variables that are widely available or easy to obtain. Work on the QResearch database has shown the usefulness of this approach to predict oesophageal cancer.[37] This approach is slowly being incorporated into general practice but it has not yet been robustly confirmed to detect the premalignant phenotype of Barrett's oesophagus, potentially because Barrett's oesophagus is frequently asymptomatic and takes many years to develop into cancer. Nevertheless, this condition needs to be recognised because of the success of early intervention in preventing oesophageal adenocarcinoma with its dire prognosis.[38]

In our study, we specifically did not include patients with ultrashort Barrett's oesophagus (ie, Prague classification of <C1 or <M3). Differences exist between UK and US guidelines on follow-up for this low-risk group and our aim was to create a prediction tool that avoided this

	BEST2 training dataset (n=394)				BEST2 testing dataset (n=297)				BOOST validation dataset (n=162)			
	Barrett's oesophagus present	Barrett's oesophagus absent	p value (χ^2)	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value (χ^2)	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value (χ^2)	AUC
n	296 (75%)	98 (25%)	..	..	227 (76%)	70 (24%)	..	..	87 (54%)	75 (46%)	..	..
Waist circumference, cm	100·66 (13·17)	93·03 (12·51)	<0·0001	0·66 (0·59–0·72)	100·19 (12·97)	95·55 (13·02)	0·00926	0·60 (0·52–0·68)	91·03 (8·31)	87·09 (9·38)	0·00588	0·62 (0·53–0·71)
Taking antireflux medication												
No	14/296 (5%)	43/95 (45%)	<0·0001	0·70 (0·65–0·75)	17/226 (8%)	29/70 (41%)	<0·0001	0·67 (0·61–0·73)	7/86 (8%)	24/74 (32%)	<0·0001	0·62 (0·56–0·68)
Yes	282/296 (95%)	52/95 (55%)	..	..	209/226 (92%)	41/70 (59%)	..	..	79/86 (92%)	50/74 (68%)	..	..
Stomach pain frequency												
Never	217/296 (73%)	35/95 (37%)	<0·0001	0·71 (0·64–0·76)	158/225 (70%)	38/70 (54%)	0·00867	0·60 (0·54–0·66)	59/85 (69%)	32/69 (46%)	0·00018	0·68 (0·60–0·75)
Occasionally*	48/296 (16%)	28/95 (29%)	..	..	47/225 (21%)	15/70 (21%)	..	..	11/85 (13%)	3/69 (4%)	..	..
Weekly	20/296 (7%)	9/95 (9%)	..	..	10/225 (4%)	6/70 (9%)	..	..	8/85 (9%)	13/69 (19%)	..	..
Daily	11/296 (4%)	23/95 (24%)	..	..	10/225 (4%)	11/70 (16%)	..	..	7/85 (8%)	21/69 (30%)	..	..

Data are n (%), n/N (%), or mean (SD), p value, or AUC with 95% CI in parentheses. p values were calculated using the χ^2 test and AUCs as calculated using the reconstructed model. Where data differ between groups it is due to missing data. Total percentages might not equal 100% due to rounding. AUC=area under the receiver operator curve. *Once or twice a month.

Table 4: **Demographic and symptom characteristics of reconstructed dataset, by presence or absence of Barrett's oesophagus**

ambiguity. Although the methods we used are generally applicable and should be considered for prediction of other diseases, we focused on Barrett's oesophagus as an example of how a tool could be used by primary-care physicians to better target people for formal screening. Patient age and sex, together with medication and smoking history, are routinely captured in primary care systems. Additionally asking about duration of heartburn and acidic taste, frequency of stomach pain, and measuring waist circumference should be simple for physicians. Alternatively, a patient could do a self-assessment using a web-based app and generate a personalised risk profile for having Barrett's oesophagus. Precise cutoffs between patients and controls will need to be defined once this risk prediction panel is tested prospectively in a primary care population in which the prevalence of Barrett's oesophagus is lower than in our cohorts. For a particular AUC, the sensitivity chosen for use in clinical practice can be altered depending on the clinical question. Whereas, if triaging for cancer in symptomatic individuals would require a sensitivity of 95% or greater, missing a diagnosis of Barrett's oesophagus might not be so critical, and a sensitivity of even lower than 90% might be adequate. Indeed, machine learning might offer a way to create accurate predictive panels to prescreen for many other diseases and could be tuned to achieve the desired sensitivity depending on the importance of the disease in question.

Reflux duration is strongly correlated with cancer risk and is longer in patients with Barrett's oesophagus. In our panel, use of antireflux medicines was a strong predictor of Barrett's oesophagus. Metabolic obesity characteristically presents with truncal obesity and is also a risk factor for Barrett's oesophagus,[39] which explains why our model predicted patients with Barrett's oesophagus to have greater waist circumferences. Waist circumference is not routinely collected, but is an easy measurement to collect, particularly for patients who wish to self-triage. A clear correlation exists between waist circumference and body-mass index (BMI), which is routinely collected. Our method identified the most important independent predictors of Barrett's oesophagus. In routine practice, replacing waist circumference with BMI might be more practical but the model would then need to be reworked. Another finding that initially appears counterintuitive is the negative correlation between Barrett's oesophagus and frequency of stomach pain; however, on further investigation this correlation makes sense. Most patients with oesophageal adenocarcinoma are not identified before cancer develops despite many of them having Barrett's oesophagus. Indeed, 40% of patients with oesophageal adenocarcinoma have not previously had symptomatic reflux and many probably had Barrett's oesophagus.[40] Therefore, Barrett's oesophagus has been hypothesised to not be associated with severity of reflux symptoms;[41] which fits with the model determined from our data.

Our panel of features differs from the QResearch database work for oesophageal cancer.[42] The QResearch panel includes dysphagia, appetite loss, weight loss, and anaemia as predictors for cancer and does not include duration of symptoms or central obesity data. These differences reflect the different realities of Barrett's oesophagus and oesophageal adenocarcinoma.

Previous works have identified risk factor panels, including multiple biomarkers, such as leptin and interleukin levels, or data from genome-wide association studies, which are not easily available, and others included only a few symptoms.[10,43] For those in which the risk factor panels were larger, several key differences exist between our analyses and these previous works. We confirmed the importance of older age,[13,43,44] male sex,[11,12,14] gastro-oesophageal reflux disease,[10,11,13,16,44,45] smoking,[17,18,43,44] and central obesity;[43,44,46] however, we found that many of these risk factors were cross-correlated in our data analysis. We overcame the challenge of panels failing external validation through a combination of univariate and multivariable feature selection techniques that yielded a stable panel. The results are better than previous panels with sensitivities of 70–80% and specificities of 50–60% or AUCs of 0·7 or lower.[10,11,43,44] By contrast, our panel validates between completely different datasets with an AUC of at least 0·81 when only considering eight risk factors. This predictive panel of risk factors might be adequate to be used as a triaging tool in clinical practice for Barrett's oesophagus.

Three recent studies support our risk prediction panel. Xie and colleagues followed-up 63 000 patients for 20 years in Norway for risk of developing oesophageal adenocarcinoma and they constructed a model based on a very similar risk panel to ours.[47] Their data were taken from a patient cohort without the level of symptom granularity we achieved by using data from cohorts in which patients were interviewed. The AUC of their model to identify 15-year risk of oesophageal adenocarcinoma was 0·84 (95% CI 0·76–0·91) but it did not attempt to identify patients with Barrett's oesophagus.[47] Similarly, Kunzmann and colleagues examined 355 034 individuals from the UK Biobank for risk of developing oesophageal adenocarcinoma. Their panel including age, sex, smoking, BMI, and history of oesophageal conditions or treatments and they identified individuals who would later develop oesophageal adenocarcinoma with an AUC of 0·80 (95% CI 0·77–0·82).[48] Once again, their study did not specifically aim to identify Barrett's oesophagus, although the features are remarkably similar to those we identified, suggesting that many patients they identified might have undiagnosed Barrett's oesophagus. We found one study that targeted sporadic Barrett's oesophagus alone that was undertaken in a small Australian cohort in which their choice of risk factors was determined by complex deduction; however, this approach did lead to a tool with an AUC of 0·82 (95% CI 0·78–0·87).[19] This tool was later validated in an independent dataset.[19,49] One additional feature of that model was hypertension, which was not identified as an independently important feature in our model even though we queried for it, raising the question of the stability of their model.

Because our aim was to create a tool for prescreening, we intentionally used the BEST2 and BOOST datasets, which had a higher incidence of Barrett's oesophagus than the general population. Generally, an open challenge to machine learning is how to properly identify important so-called minority categories, such as Barrett's oesophagus. Because Barrett's oesophagus is relatively uncommon, with a prevalence as low as 2% found in Mexico,[50] one could create extremely accurate models by assuming no individuals have Barrett's oesophagus. In the BEST2 and BOOST datasets, this issue was mitigated by use of a targeted collection of suspected at-risk individuals, which led to a distribution of Barrett's oesophagus that is much higher than that in the general population. Several methods exist to computationally rebalance the data beyond or in addition to this approach. The most common approach is undersampling, whereby existing records belonging to a prevalent category are intentionally removed to create a different ratio between the classes. Here, the relatively high number of patients with Barrett's oesophagus could be adjusted by randomly removing some of the patients. Alternatively, oversampling could be used, whereby individuals without Barrett's oesophagus are added to generate a new balance between the target patients. One popular example of this approach is the synthetic minority oversampling technique,[51] which synthetically adds artificial cases to the minority class. Another approach would be to apply a ratio of controls to known cases to train the model with a prevalence that more closely aligns with the real-world setting.

The advantage of using datasets that inherently have higher distributions of patients with Barrett's oesophagus is that our data are non-synthetic and thus more likely to be effective as a screening tool; although, one could argue that undertaking this study in a cohort with a prevalence of Barrett's oesohpagus that is similar to that of the general population might yield different results. However, further studies are needed to confirm this hypothesis and to study any potential effects of false positives or false negatives generated in a real-world setting. To this end, we propose that our algorithm should be applied to the data generated from the BEST3 study, which is a pragmatic, multisite, cluster-randomised controlled trial set in primary care centres in England, UK, where the prevalence of Barrett's oesophagus is representative of the general UK population and in which the same questions have been asked as in BEST2 and BOOST.[52] We are also undertaking another prospective study (ISRCTN 11921553) to test this hypothesis independently in a second population that more closely aligns with the general population prevalence of the disease.

The methods used to apply the machine learning analysis also present a challenge. Many researchers do

both univariate and multivariable analysis of each dataset independently, which often leads to selecting similar features in both datasets. We have previously used this approach ourselves. We made very small changes in our definitions of Barrett's oesophagus (with or without intestinal metaplasia), each of which was associated with different risk factors being important in the ensuing algorithms. These differences stem from a lack of so-called stability in the features that each model independently selected;[8] too many features, even those with relatively high prediction value, often reduces the model's power.

One current solution to both these challenges is effective feature selection. We approached this challenge by identifying which features add information. This approach is called information gain, a univariate approach. In our previous work,[8] we used a threshold of 0·1 within χ^2 with one degree of freedom to select eight features in the dataset. An advantage to using feature selection to determine important features is that they are based on a filter approach to selection, which is undertaken without any connection to a specific learning algorithm. Similarly, no human bias is involved. We incorporated this approach as one step in our current analysis.

Our results show stability across the BEST2 and BOOST datasets. Although each of these datasets was collected independently, their collection methodologies and definitions were similar enough for effective comparison. This study shows that such analyses are possible if stable features are identified that are not influenced by random artifacts in the data collection process.[53]

We considered using other multivariate feature selection algorithms including least absolute shrinkage and section operator (LASSO).[54] LASSO is one type of feature selection that is embedded in logistic regression because its feature analysis is inherently linked to this machine learning method. It has a similar limitation to the support vector machine recursive feature selection (RFE-SVM) approach.[55] Both approaches are limited to only one algorithm, in the case of RFE-SVM, the support vector machine algorithm that we also considered. Because we aimed to consider a variety of machine learning methods, we preferred using information gain and correlation-based feature selection, which are filter methods and can be used without any connection to a specific machine learning prediction model,[24] thus facilitating improved medical understanding.[35]

We also considered correlations between features, which often exist in medical datasets. We used the multivariable correlation-based feature selection algorithm to do this. We reasoned that features selected by correlation-based feature selection should be more stable than other approaches. This hypothesis is borne out by the high AUC of the predictive model and its stability against the independent validation cohort.

Having created our dataset, we considered possible biases and sought to minimise these by reconstructing the cohorts to avoid any age, sex, or race bias; however, we found that our model remains robust.

The risk prediction panels we generated are easy to use in practice. Theoretically, people could enter their symptoms into a smartphone app and receive an immediate risk factor analysis. These data could then be uploaded to a central database (eg, in the cloud) that would be updated after that person sees their medical professional.

Our study had several limitations. Because both datasets were collected from at-risk individuals, the dataset was enriched for patients with Barrett's oesophagus. Additionally, patients attending for symptom assessment are more symptomatic than those undergoing surveillance endoscopy. Nevertheless, all the patients with Barrett's oesophagus undergoing surveillance would have presented initially with symptoms. Notably, many individuals with Barrett's oesophagus have no symptoms and so this risk prediction panel is unlikely to work for these people. Nonetheless, given the robustness of the models generated, the predictive panel produced here could be of benefit to rapidly triage symptomatic patients for minimally invasive screening tools, such as the cytosponge test, because many symptomatic individuals currently undergo no testing at all.[56]

Further prospective data collection is needed using a cohort study design in a primary care setting where the prevalence of Barrett's oesophagus will be much lower to confirm the validity of our findings and to establish the final best risk prediction model parameters.

Contributors

AR contributed to study conception and design, analysis and interpretation of data, drafting of the manuscript, and statistical analysis. DGG contributed to study conception and design, acquisition of data, analysis and interpretation of data, drafting of the manuscript, and critical revision of the manuscript for important intellectual content. SJ, JA, DH, AW, SJL, and members of the BEST2 study group contributed to acquisition of data. SSS and OFA contributed to drafting of the manuscript and critical revision of the manuscript for important intellectual content. MN, MRJ, and AW contributed to analysis and interpretation of data and critical revision of the manuscript for important intellectual content. EMH, MB-Z, and UN contributed to analysis and interpretation of data, statistical analysis, and technical or material support. RCF and PS contributed to study conception and design, critical revision of the manuscript for important intellectual content, and technical or material support. LBL contributed to study conception and design, acquisition of data, analysis and interpretation of data, drafting of the manuscript, obtained funding, and study supervision.

Declaration of interests

The cytosponge device was designed by RCF and her research team in 2009–10; in 2013, the MRC licensed the technology to Covidien GI Solutions, now part of Medtronic. Medtronic have had no influence on the design, conduct, or analysis of this study. RCF is a named inventor on patents pertaining to the cytosponge and related assays and has not received any financial benefits to date from this device. All other authors declare no competing interests.

Data sharing

We will make provision for and consider data sharing requests from bona fide researchers who must abide by the following principles: data will be collected with an appropriately high level of quality assurance, data will be held securely with appropriate documentation, data will not be put into the public domain or otherwise shared without explicit ethical review or legal obligation, and they will aim to use any data generated to the maximum

public good. Data will be made available from 1 year after publication of this Article and will be de-identified. These datasets are governed by data usage policies specified by the data controller (University College London, University of Cambridge, Cancer Research UK [CRUK]). People wishing to access the data will need to ensure that their use fulfils the requirements of the data controllers. If a conflict exists that severely restricts the analyses that can be undertaken, we would endeavour to support outside researchers by hosting them as visiting workers in our team so that they can access the data. We are committed to complying with CRUK's Data Sharing and Preservation Policy. Applications are subject to review by LBL, PS, and RCF and should be directed to l.lovat@ucl.ac.uk.

Acknowledgments

This research was funded by the Charles Wolfson Charitable Trust and Guts UK and supported by the National Institute for Health Research University College London (UCL) Hospitals Biomedical Research Centre. This work was also supported by the Cancer Research UK Experimental Cancer Medicine Centre at UCL and the Wellcome/EPSRC Centre for Interventional and Surgical Sciences at UCL (203145Z/16/Z). BEST2 was funded by Cancer Research UK (12088 and 16893).

References

1 Brown KF, Rumgay H, Dunlop C, et al. The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. *Br J Cancer* 2018; **118:** 1130–41.

2 Lagergren J. Adenocarcinoma of oesophagus: what exactly is the size of the problem and who is at risk? *Gut* 2005; **54** (suppl 1): i1–5.

3 Hvid-Jensen F, Pedersen L, Drewes AM, Sørensen HT, Funch-Jensen P. Incidence of adenocarcinoma among patients with Barrett's esophagus. *N Engl J Med* 2011; **365:** 1375–83.

4 Ross-Innes CS, Debiram-Beecham I, O'Donovan M, et al. Evaluation of a minimally invasive cell sampling device coupled with assessment of trefoil factor 3 expression for diagnosing Barrett's esophagus: a multi-center case-control study. *PLoS Med* 2015; **12:** e1001780.

5 Park J, Cho YK, Kim JH. Current and future use of esophageal capsule endoscopy. *Clin Endosc* 2018; **51:** 317–22.

6 Fitzgerald RC, di Pietro M, Ragunath K, et al. British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. *Gut* 2014; **63:** 7–42.

7 Alexandre L, Broughton T, Loke Y, Beales ILP. Meta-analysis: risk of esophageal adenocarcinoma with medications which relax the lower esophageal sphincter. *Dis Esophagus* 2012; **25:** 535–44.

8 Liu X, Wong A, Kadri SRSR, et al. Gastro-esophageal reflux disease symptoms and demographic factors as a pre-screening tool for Barrett's esophagus. *PLoS One* 2014; **9:** e94163.

9 Wang P, Li Y, Reddy CK. Machine learning for survival analysis: a survey. *ACM Comput Surv* 2019; **51:** 1.

10 Locke GR, Zinsmeister AR, Talley NJ. Can symptoms predict endoscopic findings in GERD? *Gastrointest Endosc* 2003; **58:** 661–70.

11 Gerson LB, Edson R, Lavori PW, Triadafilopoulos G. Use of a simple symptom questionnaire to predict Barrett's esophagus in patients with symptoms of gastroesophageal reflux. *Am J Gastroenterol* 2001; **96:** 2005–12.

12 Ford AC, Forman D, Reynolds PD, Cooper BT, Moayyedi P. Ethnicity, gender, and socioeconomic status as risk factors for esophagitis and Barrett's esophagus. *Am J Epidemiol* 2005; **162:** 454–60.

13 Eloubeidi MA, Provenzale D. Clinical and demographic predictors of Barrett's esophagus among patients with gastroesophageal reflux disease: a multivariable analysis in veterans. *J Clin Gastroenterol* 2001; **33:** 306–09.

14 Ward EM, Wolfsen HC, Achem SR, et al. Barrett's esophagus is common in older men and women undergoing screening colonoscopy regardless of reflux symptoms. *Am J Gastroenterol* 2006; **101:** 12–17.

15 Thukkani N, Sonnenberg A. The influence of environmental risk factors in hospitalization for gastro-oesophageal reflux disease-related diagnoses in the United States. *Aliment Pharmacol Ther* 2010; **31:** 852–61.

16 Anderson LA, Watson RGP, Murphy SJ, et al. Risk factors for Barrett's oesophagus and oesophageal adenocarcinoma: results from the FINBAR study. *World J Gastroenterol* 2007; **13:** 1585–94.

17 Johansson J, Håkansson HO, Mellblom L, et al. Risk factors for Barrett's oesophagus: a population-based approach. *Scand J Gastroenterol* 2007; **42:** 148–56.

18 Steevens J, Schouten LJ, Driessen ALC, et al. A prospective cohort study on overweight, smoking, alcohol consumption, and risk of Barrett's esophagus. *Cancer Epidemiol Biomarkers Prev* 2011; **20:** 345–58.

19 Ireland CJ, Fielder AL, Thompson SK, Laws TA, Watson DI, Esterman A. Development of a risk prediction model for Barrett's esophagus in an Australian population. *Dis Esophagus* 2017; **30:** 1–8.

20 Sun X, Elston RC, Barnholtz-Sloan JS, et al. Predicting Barrett's esophagus in families: an esophagus translational research Network (BETRNet) model fitting clinical data to a familial paradigm. *Cancer Epidemiol Biomarkers Prev* 2016; **25:** 727–35.

21 Ross-Innes CS, Chettouh H, Achilleos A, et al. Risk stratification of Barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study. *Lancet Gastroenterol Hepatol* 2017; **2:** 23–31.

22 Lipman G, Bisschops R, Sehgal V, et al. Systematic assessment with I-SCAN magnification endoscopy and acetic acid improves dysplasia detection in patients with Barrett's esophagus. *Endoscopy* 2017; **49:** 1219–28.

23 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol* 2015; **67:** 1142–51.

24 Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; **23:** 2507–17.

25 Nie F, Xiang S, Jia Y, Zhang C, Yan S. Trace ratio criterion for feature selection. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. Menlo Park, CA: The AAAI Press, 2008: 671–76.

26 Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak* 2011; **11:** 51.

27 Moturu ST, Johnson WG, Liu H. Predicting future high-cost patients: a real-world risk modeling application. In: Proceedings - 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007. Fremont, CA: Institute of Electrical and Electronics Engineers, 2007: 202–08.

28 Maroco J, Silva D, Rodrigues A, Guerreiro M, Sanatana I, de Mendonça A. Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic. *BMC Res Notes* 2011; **4:** 299.

29 Langley NR, Dudzik B, Cloutier A. A decision tree for nonmetric sex assessment from the skull. *J Forensic Sci* 2018; **63:** 31–37.

30 Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol* 2017; **69:** 2657–64.

31 Zhang W, Zeng F, Wu X, Zhang X, Jiang R. A comparative study of ensemble learning approaches in the classification of breast cancer metastasis. In: Proceedings - 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing. Shanghai: Institute of Electrical and Electronics Engineers, 2009: 242–45.

32 Deo RC. Machine learning in medicine. *Circulation* 2015; **132:** 1920–30.

33 Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Med Inform Decis Mak* 2005; **5:** 3.

34 Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: a scoping review. *PLoS One* 2019; **14:** e0212356.

35 Rosenfeld A, Richardson A. Explainability in human–agent systems. *Auton Agent Multi Agent Syst* 2019; **6:** 673–705.

36 Jiang X, Jao J, Neapolitan R. Learning predictive interactions using information gain and Bayesian network scoring. *PLoS One* 2015; **10:** e0143247.

37 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:** e1–10.

38 Haidry RJ, Lipman G, Banks MR, et al. Comparing outcome of radiofrequency ablation in Barrett's with high grade dysplasia and intramucosal carcinoma: a prospective multicenter UK registry. *Endoscopy* 2015; **47:** 980–87.

39 Di Caro S, Cheung WH, Fini L, et al. Role of body composition and metabolic profile in Barrett's oesophagus and progression to cancer. *Eur J Gastroenterol Hepatol* 2016; **28:** 251–60.

40 Lagergren J, Bergström R, Lindgren A, Nyrén O. Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. *N Engl J Med* 1999; **340:** 825–31.

41 Nason KS, Wichienkuer PP, Awais O, et al. Gastroesophageal reflux disease symptom severity, proton pump inhibitor use, and esophageal carcinogenesis. *Arch Surg* 2011; **146:** 851–58.

42 Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* 2015; **5:** e007825.

43 Thrift AP, Garcia JM, El-Serag HB. A multibiomarker risk score helps predict risk for Barrett's esophagus. *Clin Gastroenterol Hepatol* 2014; **12:** 1267–71.

44 Rubenstein JH, Morgenstern H, Appelman H, et al. Prediction of Barrett's esophagus among men. *Am J Gastroenterol* 2013; **108:** 353–62.

45 Thrift AP, Kendall BJ, Pandeya N, Vaughan TL, Whiteman DC. A clinical risk prediction model for Barrett esophagus. *Cancer Prev Res (Phila)* 2012; **5:** 1115–23.

46 Kubo A, Cook MB, Shaheen NJ, et al. Sex-specific associations between body mass index, waist circumference and the risk of Barrett's oesophagus: a pooled analysis from the international BEACON consortium. *Gut* 2013; **62:** 1684–91.

47 Xie S-H, Ness-Jensen E, Medefelt N, Lagergren J. Assessing the feasibility of targeted screening for esophageal adenocarcinoma based on individual risk assessment in a population-based cohort study in Norway (The HUNT Study). *Am J Gastroenterol* 2018; **113:** 829–35.

48 Kunzmann AT, Thrift AP, Cardwell CR, et al. Model for identifying individuals at risk for esophageal adenocarcinoma. *Clin Gastroenterol Hepatol* 2018; **16:** 1229–1236.

49 Ireland CJ, Gordon AL, Thompson SK, et al. Validation of a risk prediction model for Barrett's esophagus in an Australian population. *Clin Exp Gastroenterol* 2018; **11:** 135–42.

50 Herrera Elizondo JL, Monreal Robles R, García Compean D, et al. Prevalence of Barrett's esophagus: an observational study from a gastroenterology clinic. *Rev Gastroenterol Mex* 2017; **82:** 296–300.

51 Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017; **18:** 559–63.

52 Offman J, Muldrew B, O'Donovan M, et al. Barrett's oESophagus trial 3 (BEST3): study protocol for a randomised controlled trial comparing the Cytosponge-TFF3 test with usual care to facilitate the diagnosis of oesophageal pre-cancer in primary care patients with chronic acid reflux. *BMC Cancer* 2018; **18:** 784.

53 Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl Inf Syst* 2007; **12:** 95–116.

54 Park T, Casella G. The Bayesian LASSO. *J Am Stat Assoc* 2008; **103:** 681–86.

55 Zhang X, Lu X, Shi Q, et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 2006; **7:** 197.

56 Offman J, Fitzgerald RC. Alternatives to traditional per oral endoscopy for screening. *Gastrointest Endosc Clin N Am* 2017; **3:** 379–96.