

Keywords: colorectal cancer screening; faecal immunochemical test; prediction model; test performance; risk-based screening

Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model

Jennifer Anne Cooper¹, Nick Parsons¹, Chris Stinton¹, Christopher Mathews², Steve Smith³, Stephen P Halloran^{4,5}, Sue Moss² and Sian Taylor-Phillips^{*,1}

¹Division of Health Sciences, Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK; ²Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK; ³NHS Bowel Cancer Screening Midlands and North West Programme Hub, Rugby, UK; ⁴Public Health England, London, UK and ⁵University of Surrey, Guildford, UK

Background: The faecal immunochemical test (FIT) is replacing the guaiac faecal occult blood test in colorectal cancer screening. Increased uptake and FIT positivity will challenge colonoscopy services. We developed a risk prediction model combining routine screening data with FIT concentration to improve the accuracy of screening referrals.

Methods: Multivariate analysis used complete cases of those with a positive FIT ($\geq 20 \mu\text{g g}^{-1}$) and diagnostic outcome ($n = 1810$; 549 cancers and advanced adenomas). Logistic regression was used to develop a risk prediction model using the FIT result and screening data: age, sex and previous screening history. The model was developed further using a feedforward neural network. Model performance was assessed by discrimination and calibration, and test accuracy was investigated using clinical sensitivity, specificity and receiver operating characteristic curves.

Results: Discrimination improved from 0.628 with just FIT to 0.659 with the risk-adjusted model ($P = 0.01$). Calibration using the Hosmer–Lemeshow test was 0.90 for the risk-adjusted model. The sensitivity improved from 30.78% to 33.15% at similar specificity (FIT threshold of $160 \mu\text{g g}^{-1}$). The neural network further improved model performance and test accuracy.

Conclusions: Combining routinely available risk predictors with the FIT improves the clinical sensitivity of the FIT with an increase in the diagnostic yield of high-risk adenomas.

Colorectal cancer (CRC) is the third most common cancer in men and the second most common in women (Ferlay *et al*, 2015). In England, people aged 60 to 74 years are offered a biennial guaiac faecal occult blood test (gFOBT) as part of the NHS Bowel Cancer Screening Programme (BCSP). A newer test, the faecal immunochemical test (FIT), has been shown to have superior analytical performance and enhanced clinical performance compared with gFOBT (van Rossum *et al*, 2008; Allison *et al*, 2014; Launois *et al*, 2014).

In England, a 6-month pilot study was initiated by the NHS BCSP to assess uptake and acceptability, as well as diagnostic

performance (Moss *et al*, 2016). A marked improvement in uptake was observed compared with the gFOBT (66.4 vs 59.3%). This improved uptake, combined with higher positivity, increases demand on a finite colonoscopy service. A suggested annual increase of 290 000 additional participants could place impossible demands on the service (Moss *et al*, 2016). FIT thresholds between 150 and $180 \mu\text{g Hb g}^{-1}$ faeces were considered by the pilot to ensure demands on colonoscopy were within the available capacity (Moss and Mathews, 2015; Moss *et al*, 2016).

Another approach that could improve effective colonoscopy use, test accuracy and consequently health outcomes is personalised

*Correspondence: Dr S Taylor-Phillips; E-mail: s.taylor-phillips@warwick.ac.uk

Received 4 June 2017; revised 21 September 2017; accepted 29 September 2017; published online 2 November 2017



risk-based CRC screening (Auge *et al*, 2014; Cooper *et al*, 2016; Moss *et al*, 2016). A few studies have developed risk prediction models that combine the FIT concentration with other risk indicators for use in screening referral decisions (Omata *et al*, 2011; Tao *et al*, 2012; Auge *et al*, 2014; Stegeman *et al*, 2014; Yen *et al*, 2014; Aniwan *et al*, 2015; Otero-Estévez *et al*, 2015). Stegeman *et al* (2014) combined FIT with risk factors obtained from a lifestyle questionnaire in a logistic regression model and found improved sensitivity at a similar specificity.

Previous studies have required additional testing or lifestyle questionnaires to obtain predictor information for the model. Sending additional documents such as questionnaires have been shown to significantly lower screening uptake (Watson *et al*, 2013). A more efficient approach is for the prediction model to utilise screening data routinely available as an electronic record, thus reducing participant burden, enhancing data accuracy and completeness.

Although logistic regression is typically used in medical research for prediction modelling, other machine learning algorithms such as artificial neural networks could perform better in certain medical scenarios (Sargent, 2001; Dreiseitl and Ohno-Machado, 2002). The real advantage of a neural network is in their flexibility to model complex nonlinear relationships between predictors and outcome. They can also provide absolute risk probabilities for use in decision-making.

The aim of this study was to develop a risk prediction model combining routinely available predictors from the NHS Bowel Cancer Screening System (BCSS) with individual FIT results to determine whether model performance and test accuracy are improved in an average risk English screening population. An artificial neural network model was also investigated to determine if this improved predictive power further.

MATERIALS AND METHODS

Since this study develops a risk prediction model and assesses test accuracy, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and Standards for Reporting of Diagnostic Accuracy Studies (STARD) statements have been followed when reporting this study (Bossuyt *et al*, 2003; Collins *et al*, 2015).

Study population and data source. The NHS BCSP performed a comparative study to determine the acceptability and accuracy of the FIT compared with the gFOBT (Moss *et al*, 2016). The study involved two out of the five regional screening hubs in England: (i) the Midlands and North West Hub and (ii) the Southern Hub. Between 7 April and 10 October 2014, 1 126 087 individuals were invited to complete a gFOBT and 40 930 invited to complete a FIT (1 out of 28 screening invitations). The pilot analysed data from participants aged 59 to 75 years old and is discussed in further detail elsewhere (Moss *et al*, 2016). This analysis is limited to complete cases (i.e. participants with complete data records) and those who had a FIT result of $\geq 20 \mu\text{g g}^{-1}$ (the threshold chosen for the pilot) and a definitive colonoscopy outcome ($n = 1810$).

The data used for both the FIT pilot and this study were held on the BCSS, which contains routine information on the screening pathway for all participants. These data were anonymised and provided by the Health and Social Care Information Centre (HSCIC) – now NHS Digital – through the Office for Data Release. The data were extracted by the HSCIC on the 10 March 2016. Ethical approval was obtained from the University of Warwick Biomedical and Scientific Research Ethics Committee (Reference Number REGO-2015-1575). For the sample population analysed, FIT kits were distributed between 15 April 2014 and 19 November 2014. Completed kits were received at the lab between 22 April

2014 and 5 March 2015 and examined between 25 April 2014 and 9 March 2015.

Routinely available predictors. The routinely available predictors recorded on the BCSS that were investigated included age, sex, Index of Multiple Deprivation (IMD) score and previous screening history (i.e. whether someone was a previous non-responder/responder to gFOBT screening compared with a first time invitee at baseline). Age at the start of the screening episode for the pilot was provided by NHS Digital. Social deprivation was measured using the IMD score, which is derived using the English Indices of Deprivation 2010 based on participant postcode (Department for Communities and Local Government, 2011).

FIT concentration (Index test). The OC-SENSOR FIT was measured using the OC-SENSOR Diana analyser (Eiken Chemical Co. Ltd, Tokyo, Japan, supplied by Mast Diagnostics, Bootle, UK). The FIT units were converted from ng Hb ml^{-1} buffer to $\mu\text{g Hb g}^{-1}$ faeces as recommended by the World Endoscopy Organisation (Fraser *et al*, 2012). FIT kits were sent by post for participants to complete at home and returned by mail to the screening hubs.

Colonoscopy (diagnostic test). Subjects with a positive test were offered a specialist screening practitioner appointment within 14 days of a positive FIT test and, if appropriate, referred for a colonoscopy assessment within 14 days of this appointment (alternative investigations are arranged if the colonoscopy is inappropriate, for example, CT scan or flexible sigmoidoscopy) (Department of Health, 2014). Colonoscopies were performed using the quality assurance guidelines for colonoscopy published by the NHS Cancer Screening Programmes (NHS BCSP, 2011).

Model outcome. The binary model outcome was CRC or advanced adenoma (combined) detected at colonoscopy after a positive FIT referral. Advanced adenomas were those classified as either high risk or intermediate risk, as these have potential if left untreated to develop into CRC (Winawer and Zauber, 2002; Brenner *et al*, 2007). An abnormal diagnostic test outcome indicates that an abnormality has been detected, but not polyps or cancer (e.g. haemorrhoids). The NHS BCSS uses an algorithm to record the diagnosis of an individual based on the guidelines for CRC screening and surveillance (Supplementary Table S1) (Cairns *et al*, 2010). Where there was more than one diagnostic outcome recorded for an individual, the 'greatest risk' scenario was used.

Statistical analysis. All data were analysed in RStudio Version 0.99.903 (driven by R version 3.3.1) on a Windows 7 computer (R Core Team, 2014). Two models were tested using logistic regression, with a binary response variable of cancer/advanced adenoma status: (i) FIT concentration only as a predictor and (ii) FIT concentration and routine data. The risk-adjusted model (ii) was then extended further using a feed forward neural network.

Routinely available risk predictors recorded on the BCSS were selected based on previous studies (Auge *et al*, 2014; Stegeman *et al*, 2014) and the information available from the data extract. The risk-adjusted model was built by adding all the routinely available risk factors into a single multivariable logistic regression model and then using backwards elimination to remove non-significant variables with a P -value > 0.1 . To address overfitting, 10-fold cross-validation was used during model development (Moons *et al*, 2015). Cross-validation involves partitioning the data sample into distinct subsets, performing the analysis on one subset (training data), and validating the analysis on the other subset (validation data). All possible pairwise interactions were investigated, none of which were significant at the 5% level.

All continuous variables were kept as such (i.e. not dichotomised/categorised) as recommended by the TRIPOD guidelines (Altman and Royston, 2006; Royston *et al*, 2006; Moons *et al*, 2015).

The log of the FIT concentration was used for analysis. Age was not formally significant in the model but was retained *a priori* in a minimum model due to clinical importance (McDonald *et al*, 2012; Auge *et al*, 2014; Stegeman *et al*, 2014). Screening history was coded as a factor (either a previous responder, previous non-responder compared with first time invitee at baseline). This was determined using two variables recorded on the BCSS; sequence number and type of episode.

Model performance was assessed using calibration and discrimination. Calibration (the agreement between observed outcomes and predictions) was determined using the Hosmer–Lemeshow statistic and calibration plots of predicted risk *vs* observed risk for deciles of participants (Steyerberg, 2009). Discrimination (the ability of the test to distinguish between those with and without the outcome) was assessed using the c-statistic (the area under the receiver operating characteristic (ROC) curve). The likelihood ratio test was used to determine whether the risk-based model had a significantly better fit than the model with just the FIT alone. Overall model performance was assessed using Nagelkerke's R^2 (Nagelkerke, 1991).

The ROC curves were plotted for the risk-adjusted FIT model and FIT only to compare test accuracy across different thresholds. Individuals were then sorted by predicted probability and the number of referrals kept the same between using the FIT alone and using risk-adjusted FIT. Two by two tables were produced to determine the sensitivity and specificity for thresholds between 150 and 180 $\mu\text{g g}^{-1}$ (and the equivalent risk threshold) for both models. These thresholds were selected based on previous work from the FIT pilot (Moss *et al*, 2016). A threshold of 150 $\mu\text{g g}^{-1}$ gave a similar positivity rate to the gFOBT and 180 $\mu\text{g g}^{-1}$ a similar referral rate (Moss and Mathews, 2015; Moss *et al*, 2016). A recommended threshold for the NHS Bowel Cancer Screening programme based on colonoscopy capacity is $\sim 160 \mu\text{g g}^{-1}$. It is anticipated that Wales will adopt a threshold of 150 $\mu\text{g g}^{-1}$ and Scotland 80 $\mu\text{g g}^{-1}$. Results for thresholds between 30 and 180 $\mu\text{g g}^{-1}$ are presented in the Supplementary Material.

An alternative, and possibly a better performing model to the conventional logistic regression is a feedforward artificial neural network (ANN). This model is highly flexible and, unlike logistic regression, does not require the strong assumption of linearity for combinations of variables allowing more complex nonlinear relationships between predictors and the response variable (Tu, 1996). For neural network development, the package 'nnet' in R was used for analysis purposes (Venables, 2002).

A multilayer ANN model with an input layer (consisting of the same predictors as the logistic regression model), a single hidden layer and an output layer with a single node was developed (Tu, 1996). Model fitting proceeded in a similar manner to that described for the logistic regression model using cross-validation, allowing performance to be compared directly (Steyerberg, 2009). The continuous variables (including log of FIT) were standardised using Gaussian normalisation as this approach produced lower cross-validated deviances. Networks were pruned to improve generalisation by dropping out weights with the lowest magnitude and assessing the change in cross-validated deviance (Ripley, 2007). A range of values of the weight decay regularisation term were also tested to give the lowest SSE (sum of squared errors). The final optimised neural network model was then compared with the logistic regression model by assessing model performance and test accuracy.

RESULTS

Study population. From a total of 40 930 individuals who were sent a FIT kit, 27 066 (66.13%) adequately participated (those who

had a definitive positive or negative result) and from this 2117 (7.82%) had a FIT result $\geq 20 \mu\text{g g}^{-1}$, which was classed as positive. From this group, 1818 (85.88%) had a definitive outcome recorded, this is a similar proportion of those undergoing further investigation as reported in other studies (Logan *et al*, 2011). Where a diagnostic appointment was made and an individual did not attend, this was classified as 'Not attended' and where an appointment was cancelled the outcome was classified as 'Cancelled' (Supplementary Table S2). Complete cases gave the final sample of $n = 1810$ after removing eight records without an IMD score (Figure 1 for Study flow diagram).

Seventy-two cancers, 214 high-risk adenomas, 262 intermediate-risk adenomas and 466 low-risk adenomas were detected in the study group. This gave 549 cases with a positive outcome (cancer and advanced adenomas) and 1261 participants with a negative or low-risk outcome. The mean age of this group was 66.54 years (Table 1 for outcome by age and sex). The FIT result ranged from 20 $\mu\text{g g}^{-1}$ to 20 854 $\mu\text{g g}^{-1}$ (other studies have reported a similarly high result (Auge *et al*, 2014)), with a median result of 55.6 $\mu\text{g g}^{-1}$. There were 912 individuals served by the Midlands hub and 898 by the Southern hub. The FIT concentration increased relative to the severity of the outcome Supplementary Figure S1.

Logistic regression model (complete cases used for analysis $n = 1810$). Backwards elimination identified that the FIT result, sex and previous screening history were significantly associated with CRC and advanced adenoma (the final logistic regression model is shown in Table 2). The odds of CRC and advanced adenoma increase as the FIT result increases (OR: 1.434; CI: 1.309–1.573), for males (OR: 1.749; CI: 1.415–2.166) and for previous non-responders (OR: 2.271; CI: 1.422–3.667). Age was found to not be statistically significant but was retained in the model because of clinical importance (OR: 1.020; CI: 0.889–2.112). IMD was removed from the model (OR: 0.997; CI: 0.990–1.004, $P = 0.457$).

Discrimination and calibration. The ROC curves for both models are presented in Figure 2. The AUC for the FIT only model was 0.63 (95% CI: 0.60–0.66) compared with 0.66 for the risk-adjusted model, indicating improved discrimination (95% CI: 0.63–0.69). The AUCs were significantly different ($D = -2.7601$, $P\text{-value} = 0.006$).

The calibration plots of observed risk against predicted risk are given for both models in Supplementary Figure S2. The calibration for the risk-adjusted model based on the Hosmer–Lemeshow statistic was 0.898 *vs* 0.481 for the FIT. Small P -values and points that are far from the line of equality in the calibration plot indicate a poor fit.

Test accuracy. Test accuracy is presented in a 2 by 2 table for a threshold of 160 $\mu\text{g g}^{-1}$ (Table 3). At all investigated thresholds, the sensitivity and specificity of risk adjusted FIT was greater than FIT alone (see Supplementary Table S3). At a threshold of 160 $\mu\text{g g}^{-1}$ (keeping the referral rate the same gives an equivalent risk threshold of 0.389 for the risk-adjusted model), the FIT has a sensitivity of 30.78 *vs* 33.15% for the risk-adjusted model and a specificity of 83.66 *vs* 84.69%.

The risk-adjusted model for this sample population leads to the detection of 13 additional advanced adenomas and the same number of cancers (17 more high-risk adenomas, 4 less intermediate-risk adenomas) when compared with the FIT only at an equivalent threshold of 160 $\mu\text{g g}^{-1}$. The severity profiles of the detected lesions are shown in Table 3 (further thresholds presented in Supplementary Tables S4–S6).

Presenting the results by sex (Table 4) shows the risk model at 160 $\mu\text{g g}^{-1}$ recalls more men and fewer women, increases detection in men but decreases detection in women when compared to the FIT result alone. The FIT result alone recalled 225 men (115 TP –

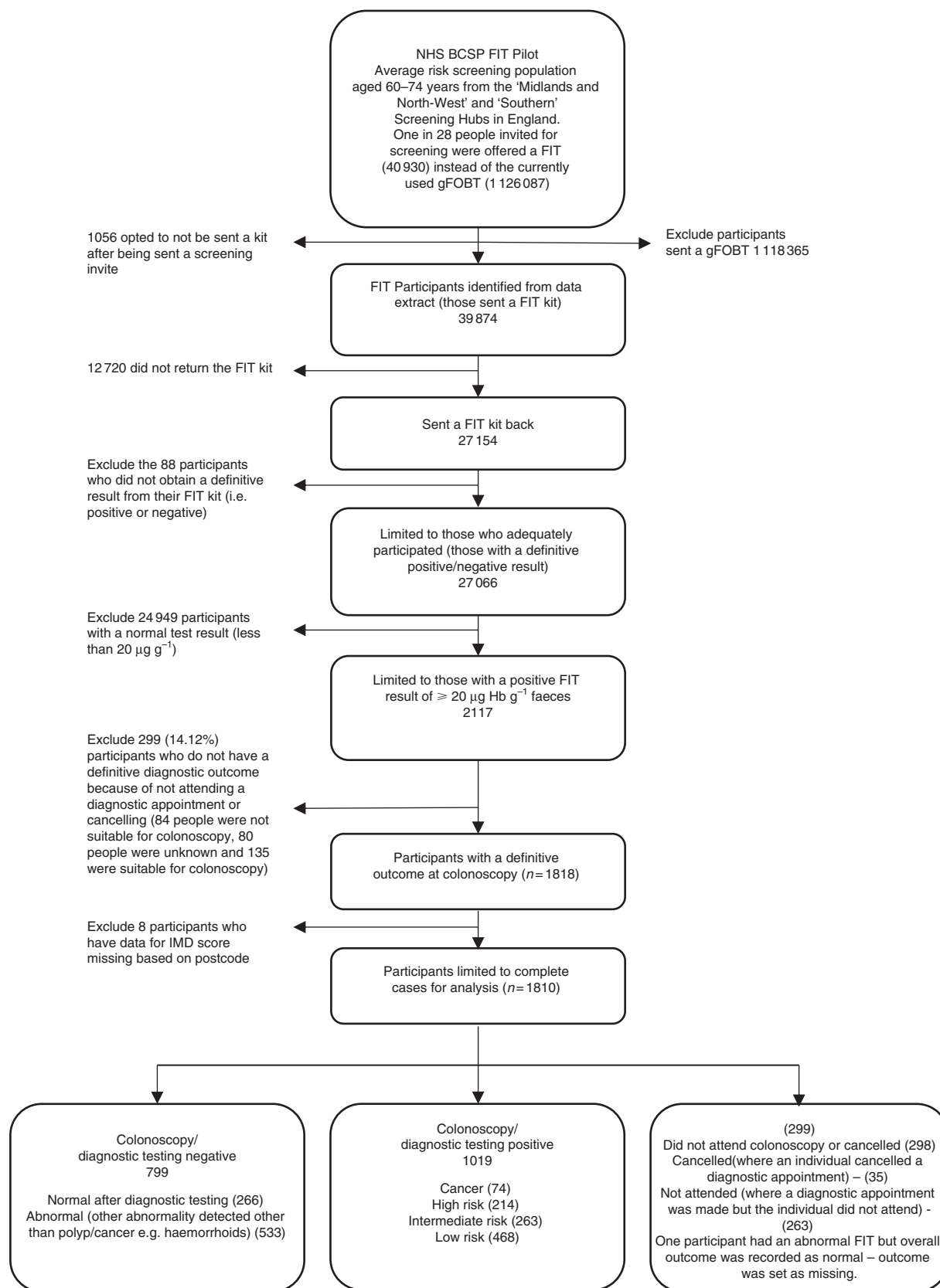


Figure 1. Study flow diagram.

true positives, 110 FP – false positives), of which 115 had cancer or advanced adenoma (51.11%) and 150 women (54 TP, 96 FP) where 54 (36%) had cancer or advanced adenoma. The logistic regression model recalled 314 men (156 TP, 158 FP), of which 156 (49.68%)

had cancer or advanced adenoma, and 61 women (26 TP, 35 FP), of which 26 (42.62%) had cancer or advanced adenoma. Results by sex are shown at different thresholds in Supplementary Tables S7–S13. Supplementary Table S14 also gives the cancer and advanced

Table 1. Diagnostic outcome by age and sex (n = 1810)

Diagnostic outcome	Age category, female					Age category, male				
	≤60	61–65	66–70	71–75	Total	≤60	61–65	66–70	71–75	Total
Abnormal	29	90	93	70	282	28	86	85	50	249
Cancer	2	5	6	10	23	4	15	21	10	50
High-risk adenoma	7	12	27	23	69	17	37	58	33	145
Intermediate-risk adenoma	13	23	45	18	99	20	53	52	38	163
Low-risk adenoma	18	67	73	45	203	33	84	90	56	263
Normal (no abnormalities found)	23	45	57	19	144	13	39	36	32	120
Total	92	242	301	185	820	115	314	342	219	990

Table 2. Final multiple logistic regression model (FIT combined with risk indicators)

Coefficients	Estimate	Std. error	Pr (> z)	OR (95% CI)
Intercept	- 4.439	0.934	< 0.001	0.012 (0.002–0.073)
log(FIT result + 1)	0.360	0.047	< 0.001	1.434 (1.309–1.573)
Age at episode start	0.020	0.015	0.171	1.020 (0.991–1.050)
Sex (male)	0.559	0.109	< 0.001	1.749 (1.415–2.166)
First time invitee	0.000	—	—	—
Previous non-responder (compared with first time screen)	0.820	0.241	0.001	2.271 (1.422–3.667)
Previous responder (compared with first time screen)	0.308	0.220	0.162	1.361 (0.889–2.112)

Abbreviations: CI = confidence interval; FIT = faecal immunochemical test; OR = odds ratio; Pr = probability. Risk equation:

$$P = \frac{e^{-4.44+0.360x_1+0.02x_2+0.56x_3+0.82x_4+0.31x_5}}{1+e^{-4.44+0.360x_1+0.02x_2+0.56x_3+0.82x_4+0.31x_5}}$$

where *P* is the probability; *x*₁ the log(FIT result + 1); *x*₂ the age at episode start; *x*₃ the sex (male compared with female at baseline); *x*₄ the previous non-responder (compared with first time screen); *x*₅ the previous responder (compared with first time screen). Null deviance – 2221.4 on 1809 degrees of freedom; residual deviance – 2103.0 on 1804 degrees of freedom; AIC – 2115; number of Fisher scoring iterations – 4; Nagelkerke’s *R*² – 0.09 (risk-adjusted model) and 0.05 (FIT only); events per variable/parameter – 91.5 (549 events/5 parameters + 1 for the parameter representing the constant).

adenoma detection rates for each sex and screening history subgroup.

Neural network. A network with five input nodes, three hidden layer nodes and one output node gave the lowest cross-validated deviance (2103.04) and was selected to develop further. A weight decay of 0.01 gave the smallest SSE (346.0445). The model with the lowest cross-validated deviance (2077.694) after pruning is shown graphically in Supplementary Figure S3 and includes the FIT result, age, sex and previous screening history. Supplementary Figure S4 shows the risk equation for the final neural network and Supplementary Table S15 gives the corresponding weight connection values.

The AUC for the neural network was higher than the equivalent logistic regression model: 0.69 (95% CI: 0.66–0.71). An ROC test between the logistic regression model and the neural network shows that the AUC is statistically significantly different (*D* = - 3.5057, *P*-value = 0.0005). ROC curves of all three models are given in Figure 2. Calibration for the neural network gave a similar result (0.8924) to the logistic regression model (0.8977). Patient profiles are presented for both the logistic regression and neural network risk models in Supplementary Table S16.

At all investigated thresholds, the sensitivity and specificity of the neural network was greater than the equivalent logistic regression model. For 160 μg g⁻¹ the sensitivity of the neural network was 35.15% and the specificity 85.57%. Applying the neural network at a threshold of 160 μg g⁻¹ leads to 24 more advanced adenomas being detected and the same number of cancers (30 more high-risk adenomas and 6 less intermediate-risk adenomas) compared with FIT only (Tables 3 and 4).

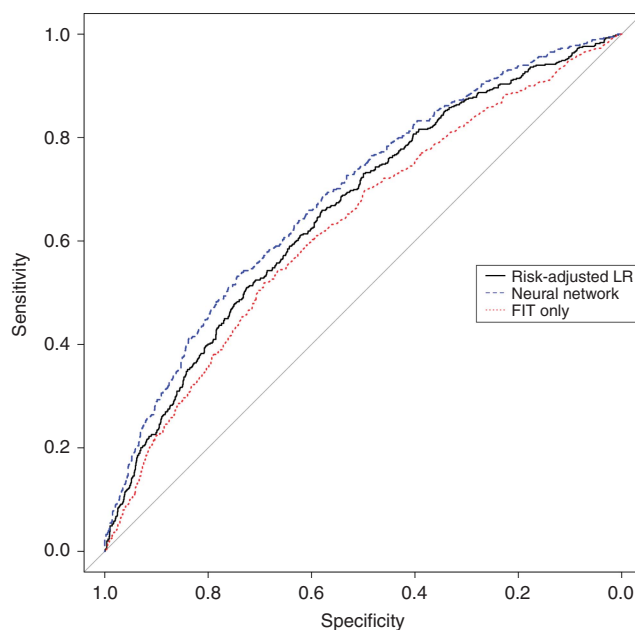


Figure 2. ROC curves for FIT only compared with the risk-adjusted FIT and neural network models. Area under the curve (AUC) (95% CI) for the Neural Network Model: 0.686 (0.659–0.712); AUC (95% CI) for the Risk-adjusted Logistic Regression Model: 0.659 (0.632–0.686); AUC (95% CI) for the FIT only: 0.628 (0.600–0.656).

Table 3. Two by two table for FIT only, the risk-adjusted logistic regression model and the neural network

160 µg Hb g ⁻¹ faeces threshold	Diagnostic positive			Diagnostic negative			Total
	FIT	Risk-adjusted	Neural network	FIT	Risk-adjusted	Neural network	
FIT/risk positive	169 37 – Cancer 66 – High-risk adenoma 66 – Intermediate-risk adenoma	182 37 – Cancer 83 – High-risk adenoma 62 – Intermediate-risk adenoma	193 37 – Cancer 96 – High-risk adenoma 60 – Intermediate-risk adenoma	206 70 – Abnormal 92 – Low-risk adenoma 44 – Normal (no abnormalities found)	193 69 – Abnormal 81 – Low-risk adenoma 43 – Normal (no abnormalities found)	182 62 – Abnormal 79 – Low-risk adenoma 41 – Normal (no abnormalities found)	375
FIT/risk negative	380 36 – Cancer 148 – High-risk adenoma 196 – Intermediate-risk adenoma	367 36 – Cancer 131 – High-risk adenoma 200 – Intermediate-risk adenoma	356 36 – Cancer 118 – High-risk adenoma 202 – Intermediate-risk adenoma	1055 396 – Abnormal 439 – Low-risk adenoma 220 – Normal (no abnormalities found)	1068 397 – Abnormal 450 – Low-risk adenoma 221 – Normal (no abnormalities found)	1079 404 – Abnormal 452 – Low-risk adenoma 223 – Normal (no abnormalities found)	1435
Total	549			1261			1810

Abbreviations: FIT = faecal immunochemical test; Hb = haemoglobin. A threshold of 160 µg Hb g⁻¹ faeces was used for the FIT, which is equivalent to a risk threshold of 0.389 for the risk-adjusted model and 0.407 for the neural network. Profiles of outcome severity are also given. FIT only: Sensitivity 30.78%, specificity 83.66%, PPV 45.07%, NPV 73.52%, FIT positivity 20.72%, cancer detection rate 9.34%. Risk adjusted: Sensitivity 33.15%, specificity 84.69%, PPV 48.53%, NPV 74.42%, FIT positivity 20.72%, cancer detection rate 10.60%. Neural network: Sensitivity 35.15%, specificity 85.57%, PPV 51.47%, NPV 75.19%, FIT positivity 20.72%, cancer detection rate 10.66%.

At 160 µg g⁻¹ compared with the logistic regression model, the neural network increases the number of cancers and advanced adenomas detected for women equalising the difference seen between the sexes and also halves the number of FP results for women compared with FIT only. The neural network recalled 279 men (146 TP, 133 FP), of which 146 (52.33%) had cancer or advanced adenoma, and 96 women (47 TP, 49 FP), of which 47 (48.96%) had cancer or advanced adenoma. The neural network improves the percentage of cancers/advanced adenomas detected in those recalled for further diagnostic tests (PPV – positive predictive value).

DISCUSSION

This study has demonstrated that including routinely available risk predictors in the screening algorithm alongside the FIT can improve both model performance and test accuracy. The risk-adjusted screening algorithm detected 13 more advanced adenomas and the same number of cancers when keeping the referral rate constant at a FIT threshold of 160 µg Hb g⁻¹ faeces. Based on the results from this data, for every 1 000 000 people invited to screening, we estimate 318 additional advanced adenomas (4447/1 000 000) would be detected compared with FIT only (4129/1 000 000). Although this approach would require external validation, the figures give the relative performance of this risk-based approach. The algorithm mainly improves detection in men compared with women.

By extending the model using more complex methods, the neural network was shown to improve model performance and test accuracy further with the detection of 24 more advanced adenomas (FIT threshold 160 µg g⁻¹). For every 1 000 000 people invited to screening, we estimate 586 additional advanced adenomas would be detected compared with FIT (4715/1 000 000). This modelling approach also equalised the difference in cancers/advanced adenomas detected between men and women seen with the logistic regression model. Although the neural network recalls fewer women, the PPV is increased compared with the other models and is similar between the sexes (men – 52.33; women – 48.96%).

Strengths of the study include the quality of data since this was collected for the FIT pilot comparative study, which was implemented within a live screening programme. In addition, routine data were used to develop the risk prediction model meaning no additional data collection, reducing costs and the burden on screening participants. The test thresholds analysed were those that were identified in the FIT pilot as well as other internationally used thresholds to aid comparison of a risk-adjusted approach.

Limitations of the study include the lack of follow-up data for participants with a result of <20 µg g⁻¹. Ideally, follow-up data for participants sent the FIT would be obtained from cancer registries (National Cancer Intelligence Network, or Office for National Statistics data). A follow-up period of 2 years would allow the clinical identification of existing cancers. Not all individuals had a diagnostic result if they cancelled or did not attend the appointment and this could cause potential selection bias if non-healthy participants tend to not have follow-up colonoscopy. The pattern of attendance for diagnostic investigation seen in this study is, however, similar to that seen in the screening programme in general (Logan *et al*, 2011). By selecting those with a result of ≥20 µg g⁻¹ and limiting to those with a definitive diagnostic outcome, the selected groups are at higher risk of CRC than the general screening population. This approach can lead to partial verification bias and inflated test accuracy measures (de Groot *et al*, 2011; Naaktgeboren *et al*, 2016). However, the results provided in this study give relative performance of a risk-adjusted approach vs a regular screening approach.

Part of the increase in detection for the FIT in the pilot was due to increased uptake of this test compared with the gFOBT (66.4 vs 59.3%) (Moss *et al*, 2015); this study assumes the same uptake seen with the pilot. In subsequent FIT screening rounds, there could be a change in the uptake whereby non-responders to gFOBT are more likely to respond to the FIT, whereas non-responders to FIT may be less likely to respond to the next FIT. This could affect future detection rates and subsequently model performance. However, data from four rounds of a biennial FIT screening programme in the Netherlands showed that uptake increased from 60 to 63%, and the same could be expected with this new test (van der Vlugt *et al*, 2017).

Table 4. Two by two table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex

160 µg Hb g ⁻¹ faeces threshold		Diagnostic positive						Diagnostic negative						Total	
		FIT only		Risk-adjusted		Neural network		FIT only		Risk-adjusted		Neural network			
		Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female		
FIT/risk positive	Total	115	54	156	26	146	47	Total	110	96	158	35	133	49	375
	Cancer	27	10	29	8	27	10	Low-risk adenoma	41	29	60	9	49	13	
	High-risk adenoma	45	21	72	11	71	25	Abnormal	51	41	66	15	59	20	
	Intermediate-risk adenoma	43	23	55	7	48	12	Normal (no abnormalities found)	18	26	32	11	25	16	
FIT/risk negative	Total	243	137	202	165	212	144	Total	522	533	474	594	499	580	1435
	Cancer	23	13	21	15	23	13	Low-risk adenoma	222	174	203	194	214	190	
	High risk adenoma	100	48	73	58	74	44	Abnormal	198	241	183	267	190	262	
	Intermediate risk adenoma	120	76	108	92	115	87	Normal (no abnormalities found)	102	118	88	133	95	128	
Total		549							1261						1810

Abbreviations: FIT = faecal immunochemical test; Hb = haemoglobin. A threshold of 160 µg Hb g⁻¹ faeces was used for the FIT, which is equivalent to a risk threshold of 0.389 for the risk-adjusted model and 0.407 for the neural network. Profiles of outcome severity are also given.

Other studies that have investigated the added value of risk factors combined with the FIT include a study in the Netherlands, which combined the following risk predictors: total calcium intake, family history, age and FIT result (OC-Sensor) (Stegeman *et al*, 2014). The AUC ROC improved from 0.69 to 0.76 compared with an improvement of 0.63 to 0.66 reported in this research. This study obtained its additional data using a questionnaire, which would rely on a response with a potential negative impact on uptake, whereas our study used routine data.

Stratification of risk using a logistic regression model combining age and sex with the FIT result has been investigated by Auge *et al* (2014). CRC risk was stratified into 16 categories and 3 risk levels based on the positive predictive value. The authors suggest that this stratified approach could be used to prioritise higher risk individuals for colonoscopy. By categorising risk, however, we lose individual information as the probabilities become standardised for all individuals in one group (Moons *et al*, 2015). Our study gives an absolute risk prediction for each individual, providing a personalised and potentially more accurate approach to screening.

This study utilised the data recorded routinely on the BCSS to develop a risk prediction model, which could be implemented in practice without additional data collection. Although the performance of the neural network was better than the logistic regression model, the interpretation of neural networks is more complex and for this reason they are not routinely used in clinical practice (Dayhoff and DeLeo, 2001; Sargent, 2001). Both models, on the other hand, give the absolute risk prediction for each individual and this can be used to make clinical decisions regarding screening referral by setting an appropriate ‘risk threshold’. In addition, if further predictors are investigated in the future, nonlinear predictors and model interactions may be better captured with a neural network or other machine-learning algorithm.

Based on the results of this study, a risk-adjusted approach could be implemented at the point of screening to decide which participants are at greatest risk for more targeted colonoscopy referral. Before application of a risk-adjusted approach, external validation of the model would be required to assess performance also enabling a more accurate risk positivity threshold to be derived. The algorithm led to greater detection in males compared to females, which depending on screening programme aims will need greater investigation (e.g. using separate models for each sex). Likewise, the detection rate seen between responders/non-

responders/first-time invitees will need consideration in future risk models by dissecting previous screening history in greater detail.

Model performance metrics including Nagelkerke’s *R*², AUC and the deviance suggest that the prediction of cancer/advanced adenomas at colonoscopy is not fully explained or captured by predictors used in the model. Future research should therefore focus on the investigation of additional predictors from the BCSS to improve predictive performance. Additional predictors from the BCSS could include flexible sigmoidoscopy participation and previous colonoscopy results, the outcomes of which affect future risk. Previous FIT results could also be monitored over time as the Hb concentration relates to the detection of adenomas in future screening rounds (Digby *et al*, 2016). Spot positivity of previous gFOBTs could also be investigated while transitioning over to the FIT (Geraghty *et al*, 2014). Lifestyle factors have also been shown to have a significant effect on the risk of CRC (diet, alcohol, physical inactivity and being overweight) (Parkin *et al*, 2011). Although this information is not currently included on the BCSS, other sources such as electronic health records or questionnaires could be used to obtain this information.

As the NHS BCSP prepares to transition to the FIT in 2018, these initial investigations have shown that further exploration of the BCSS for additional predictors which could be included in the screening algorithm may help to improve test accuracy and make more effective use of an expensive and severely limited colonoscopy resource.

ACKNOWLEDGEMENTS

Jennifer Cooper, Sian Taylor-Phillips and Chris Stinton are supported by the NIHR CLAHRC West Midlands initiative. This paper presents independent research and the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- Allison JE, Fraser CG, Halloran SP, Young GP (2014) Population screening for colorectal cancer means getting FIT: the past, present, and future of colorectal cancer screening using the fecal immunochemical test for hemoglobin (FIT). *Gut Liver* **8**(2): 117–130.
- Altman DG, Royston P (2006) The cost of dichotomising continuous variables. *BMJ* **332**(7549): 1080–1080.
- Aniwan S, Rerknimitr R, Kongkam P, Wisedopas N, Ponuthai Y, Chaithongrat S, Kullavanijaya P (2015) A combination of clinical risk stratification and fecal immunochemical test results to prioritize colonoscopy screening in asymptomatic participants. *Gastrointest Endosc* **81**(3): 719–727.
- Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, Buron A, Lopez-Ceron M, Bessa X, Serradesanferm A, Piraces M, Macia F, Guayta R, Filella X, Molina R, Jimenez W, Castells A (2014) Risk stratification for advanced colorectal neoplasia according to fecal hemoglobin concentration in a Colorectal Cancer Screening Program. *Gastroenterology* **147**: 628–636.e1.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG (2003) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* **138**(1): W1–12.
- Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U (2007) Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. *Gut* **56**(11): 1585–1589.
- Cairns SR, Scholefield JH, Steele RJ, Dunlop MG, Thomas HJ, Evans GD, Eaden JA, Rutter MD, Atkin WP, Saunders BP, Lucassen A, Jenkins P, Fairclough PD, Woodhouse CR (2010) Guidelines for colorectal cancer screening and surveillance in moderate and high risk groups (update from 2002). *Gut* **59**(5): 666–689.
- Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* **13**: 1.
- Cooper JA, Moss SM, Smith S, Seaman HE, Taylor-Phillips S, Parsons N, Halloran SP (2016) FIT for the future: a case for risk-based colorectal cancer screening using the faecal immunochemical test. *Colorectal Dis* **18**(7): 650–653.
- Dayhoff JE, DeLeo JM (2001) Artificial neural networks: opening the black box. *Cancer* **91**(8, Suppl): 1615–1635.
- de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, Moons KG (2011) Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ (Clin Res Ed)* **343**: d4770.
- Department for Communities and Local Government (2011) *The English Indices of Deprivation 2010*. Department for Communities and Local Government. Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010>.
- Department of Health (2014) *NHS Public Health Functions Agreement 2015–2016*. Service Specification No. 26, Bowel Cancer Screening Programme. Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010>.
- Digby J, Fraser CG, Carey FA, Diament RH, Balsitis M, Steele RJ (2016) Faecal haemoglobin concentration is related to detection of advanced colorectal neoplasia in the next screening round. *J Med Screen* **24**: 62–68.
- Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* **35**(5–6): 352–359.
- Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**(5): E359–E386.
- Fraser CG, Allison JE, Halloran SP, Young GP (2012) A proposal to standardize reporting units for fecal immunochemical tests for hemoglobin. *J Natl Cancer Inst* **104**(11): 810–814.
- Geraghty J, Butler P, Seaman H, Snowball J, Sarkar S, Blanks R, Halloran S, Bodger K, Rees CJ (2014) Optimising faecal occult blood screening: retrospective analysis of NHS Bowel Cancer Screening data to improve the screening algorithm. *Br J Cancer* **111**(11): 2156–2162.
- Launois R, Le Moine JG, Uzzan B, Fiestas Navarrete LI, Benamouzig R (2014) Systematic review and bivariate/HSROC random-effect meta-analysis of immunochemical and guaiac-based fecal occult blood tests for colorectal cancer screening. *Eur J Gastroenterol Hepatol* **26**(9): 978–989.
- Logan RFA, Patnick J, Nickerson C, Coleman L, Rutter MD, von Wagner C (2011) Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. *Gut* **61**: 1439–1446.
- McDonald PJ, Strachan JA, Digby J, Steele RJ, Fraser CG (2012) Faecal haemoglobin concentrations by gender and age: implications for population-based screening for colorectal cancer. *Clin Chem Lab Med* **50**(5): 935–940.
- Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* **162**(1): W1–73.
- Moss S, Mathews C (2015) *NHS Bowel Cancer Screening Programmes: Evaluation of Pilot of Faecal Immunochemical Test: Final Report*. National Screening Committee Website, Centre for Cancer Prevention, Wolfson Institute, Queen Mary University of London: London, UK (QMUL).
- Moss S, Mathews C, Day TJ, Smith S, Seaman HE, Snowball J, Halloran SP (2016) Increased uptake and improved outcomes of bowel cancer screening with a faecal immunochemical test: results from a pilot study within the national screening programme in England. *Gut* **66**: 1631–1644.
- Naaktgeboren CA, de Groot JAH, Rutjes AWS, Bossuyt PMM, Reitsma JB, Moons KGM (2016) Anticipating missing reference standard data when planning diagnostic accuracy studies. *BMJ* **352**: i402.
- Nagelkerke N (1991) A note on a general definition of the coefficient of determination. *Biometrika* **78**: 691–692.
- NHS BCSP (2011) Quality assurance guidelines for colonoscopy. NHS BCSP Publication No. 6 February 2011. By *NHS Cancer Screening Programmes*, Chilton A, Rutter M (eds). Available at: <https://www.gov.uk/government/publications/bowel-cancer-screening-colonoscopy-quality-assurance>.
- Omata F, Shintani A, Isozaki M, Masuda K, Fujita Y, Fukui T (2011) Diagnostic performance of quantitative fecal immunochemical test and multivariate prediction model for colorectal neoplasms in asymptomatic individuals. *Eur J Gastroenterol Hepatol* **23**(11): 1036–1041.
- Otero-Estévez O, De Chiara L, Rodríguez-Berrocal FJ, Páez De La Cadena M, Cubiella J, Castro I, Gonzalez-Mao C, Hernandez V, Martínez-Zorzano VS (2015) Serum sCD26 for colorectal cancer screening in family-risk individuals: comparison with faecal immunochemical test. *Br J Cancer* **112**(2): 375–381.
- Parkin DM, Boyd L, Walker LC (2011) 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *Br J Cancer* **105**(S2): S77–S81.
- R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.
- Ripley BD (2007) *Pattern Recognition and Neural Networks*. Cambridge University Press: Cambridge, UK.
- Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* **25**(1): 127–141.
- Sargent DJ (2001) Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* **91**(8, Suppl): 1636–1642.
- Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, Kuipers EJ, Fockens P, Kraaijenhagen RA, Bossuyt PM (2014) Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. *Gut* **63**(3): 466–471.
- Steyerberg EW (2009) *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York, NY, USA.
- Tao S, Haug U, Kuhn K, Brenner H (2012) Comparison and combination of blood-based inflammatory markers with faecal occult blood tests for non-invasive colorectal cancer screening. *Br J Cancer* **106**(8): 1424–1430.
- Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* **49**(11): 1225–1231.
- van der Vlugt M, Grobbee EJ, Bossuyt PMM, Bongers E, Spijker W, Kuipers EJ, Lansdorp-Vogelaar I, Essink-Bot M-L, Spaander MCW, Dekker E (2017) Adherence to colorectal cancer screening: four rounds of faecal immunochemical test-based screening. *Br J Cancer* **116**(1): 44–49.
- van Rossum LG, van Rijn AF, Laheij RJ, van Oijen MG, Fockens P, van Krieken HH, Verbeek AL, Jansen JB, Dekker E (2008) Random comparison of guaiac and immunochemical fecal occult blood tests for colorectal cancer in a screening population. *Gastroenterology* **135**(1): 82–90.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*. 4th edn. Springer: New York, NY, USA.

Watson J, Shaw K, Macgregor M, Smith S, Halloran S, Patnick J, Beral V, Green J (2013) Use of research questionnaires in the NHS Bowel Cancer Screening Programme in England: impact on screening uptake. *J Med Screen* **20**(4): 192–197.

Winawer SJ, Zauber AG (2002) The advanced adenoma as the primary target of screening. *Gastrointest Endosc Clin N Am* **12**(1): 1–9v.

Yen AM, Chen SL, Chiu SY, Fann JC, Wang PE, Lin SC, Chen YD, Liao CS, Yeh YP, Lee YC, Chiu HM, Chen HH (2014) A new insight into fecal

hemoglobin concentration-dependent predictor for colorectal neoplasia. *Int J Cancer* **135**(5): 1203–1212.



This work is licensed under the Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>

© The Author(s) named above 2018

Supplementary Information accompanies this paper on British Journal of Cancer website (<http://www.nature.com/bjc>)