Original citation:

Permanent WRAP URL:
http://wrap.warwick.ac.uk/95228

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY NC SA 4.0) and may be reused according to the conditions of the license. For more details see: https://creativecommons.org/licenses/by-nc-sa/4.0/

A note on versions:
The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: wrap@warwick.ac.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 ($\geq 20 \mu g$ 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 $\mu g$ 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 (Perlay 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 $\mu g$ 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...
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-Estevez 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., 2015). FIT kits were sent by post for participants to complete at home and returned by mail to the screening hubs.

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 \(\mu g\) 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.

Colonooscopy (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.

**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 FOB 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 \(\geq 20\ \mu g\) 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/respondent 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).

**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 $\mu$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 $\mu$g g$^{-1}$ gave a similar positive rate to the gFOBT and 180 $\mu$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 $\sim 160$ $\mu$g g$^{-1}$. It is anticipated that Wales will adopt a threshold of 150 $\mu$g g$^{-1}$ and Scotland 80 $\mu$g g$^{-1}$. Results for thresholds between 30 and 180 $\mu$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 allowing performance to be compared directly (Steyerberg, 2009).

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 $\geq 20$ $\mu$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 $\mu$g g$^{-1}$ to 20 854 $\mu$g g$^{-1}$ (other studies have reported a similarly high result (Auge et al, 2014)), with a median result of 55.6 $\mu$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 ($D=-2.7601$, $P=0.006$).

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 $\mu$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 $\mu$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 $\mu$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 $\mu$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 –
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

Figure 1. Study flow diagram.
Table 1. Diagnostic outcome by age and sex (n = 1810)

<table>
<thead>
<tr>
<th>Diagnostic outcome</th>
<th>Age category, female</th>
<th></th>
<th>Age category, male</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤ 60</td>
<td>61–65</td>
<td>66–70</td>
<td>71–75</td>
</tr>
<tr>
<td>Abnormal</td>
<td>29</td>
<td>90</td>
<td>93</td>
<td>70</td>
</tr>
<tr>
<td>Cancer</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>High-risk adenoma</td>
<td>7</td>
<td>12</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>Intermediate-risk adenoma</td>
<td>13</td>
<td>23</td>
<td>45</td>
<td>18</td>
</tr>
<tr>
<td>Low-risk adenoma</td>
<td>18</td>
<td>67</td>
<td>73</td>
<td>45</td>
</tr>
<tr>
<td>Normal (no abnormalities found)</td>
<td>23</td>
<td>45</td>
<td>57</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>242</td>
<td>301</td>
<td>185</td>
</tr>
</tbody>
</table>

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

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

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

\[
P = \frac{e^{-(4.439 + 0.360 \times \text{FIT result} + 1) + 0.020 \times \text{age at episode start} + 0.559 \times \text{sex (male)} + 0.000 \times \text{first time invitee} + 0.820 \times \text{previous non-responder} + 0.308 \times \text{previous responder}}{1 + e^{-(4.439 + 0.360 \times \text{FIT result} + 1) + 0.020 \times \text{age at episode start} + 0.559 \times \text{sex (male)} + 0.000 \times \text{first time invitee} + 0.820 \times \text{previous non-responder} + 0.308 \times \text{previous responder}}
\]

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 (2160.04) 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 \( \mu \)g 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 \( \mu \)g 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: 0.669 (0.659–0.672); 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).
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).

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 internationally. 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.

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 adjusted screening algorithm detected 13 more 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%).

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

<table>
<thead>
<tr>
<th>Threshold gHb</th>
<th>FIT</th>
<th>Risk-adjusted</th>
<th>Neural network</th>
<th>FIT</th>
<th>Risk-adjusted</th>
<th>Neural network</th>
</tr>
</thead>
<tbody>
<tr>
<td>160 μg g⁻¹ faeces</td>
<td>169</td>
<td>37 – Cancer</td>
<td>182</td>
<td>193</td>
<td>206</td>
<td>193</td>
</tr>
<tr>
<td>37 – High-risk adenoma</td>
<td>37 – Cancer</td>
<td>132</td>
<td>132</td>
<td>140</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>66 – Intermediate-risk adenoma</td>
<td>62 – Intermediate-risk adenoma</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Threshold gHb</th>
<th>FIT</th>
<th>Risk-adjusted</th>
<th>Neural network</th>
<th>FIT</th>
<th>Risk-adjusted</th>
<th>Neural network</th>
</tr>
</thead>
<tbody>
<tr>
<td>160 μg g⁻¹ faeces</td>
<td>380</td>
<td>36 – Cancer</td>
<td>367</td>
<td>356</td>
<td>1055</td>
<td>1068</td>
</tr>
<tr>
<td>36 – Cancer</td>
<td>131</td>
<td>118</td>
<td>202</td>
<td>220</td>
<td>221</td>
<td></td>
</tr>
<tr>
<td>148 – High-risk adenoma</td>
<td>60 – High-risk adenoma</td>
<td>44</td>
<td>43</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>196 – Intermediate-risk adenoma</td>
<td>200</td>
<td>202</td>
<td>220</td>
<td>221</td>
<td>223</td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** FIT – faecal immunochemical test; gHb – 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%.
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.


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