# Development of Risk Prediction Models Combining Routine EHR Data for Colorectal Cancer Screening Referral Decisions

**Jennifer A Cooper**[1], Prof. Tom Marshall[2], Dr. Ronan Ryan[2], Dr. Nick Parsons[1], Dr. Chris Stinton[1], Dr. Sian Taylor-Phillips[1]

[1]Department of Health Sciences, Gibbet Hill Campus, The University of Warwick, Coventry, CV4 7AL, UK

[2]Institute of Applied Health Research, University of Birmingham, West Midlands, B15 2TT, UK

**MEMTAB Theme**: eHealth and mHealth for diagnosis, prognosis and monitoring

## Background

- Risk prediction models incorporating both faecal occult blood test (**FOBT**) results and other colorectal cancer (**CRC**) risk factors have demonstrated increased sensitivity than FOBT alone.[1,2]
- Electronic primary care records have a rich level of data including; symptoms, diagnoses, lab test results and prescriptions which may add a further dimension to a risk based prediction model.
- The BCSS receives data for its participants from the NHS Spine which houses demographic information drawn from GP records. There is capacity to draw further data from the Spine/GP records to improve screening referral decisions (**Figure 1**).
- **This study aimed to determine:**

(i) the availability of GP data for key predictors of CRC in the screening population.
(ii) whether this additional information can inform more accurate screening referral decisions in future risk prediction models.

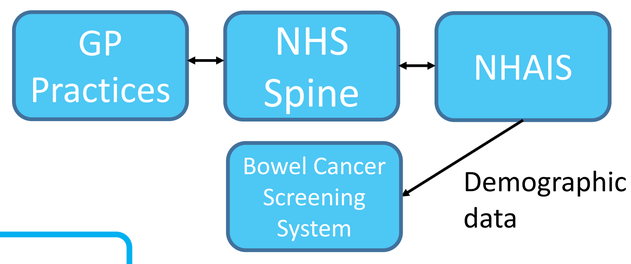


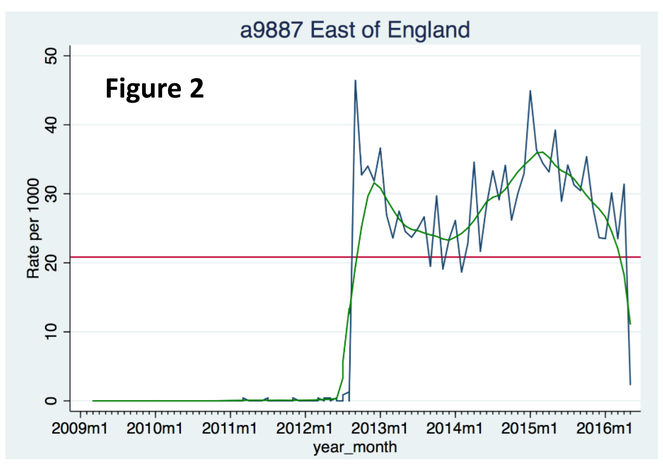

**Figure 1**: Data schematic of the BCSS and interlinks with other data services. **BCSS**: Bowel Cancer Screening System.

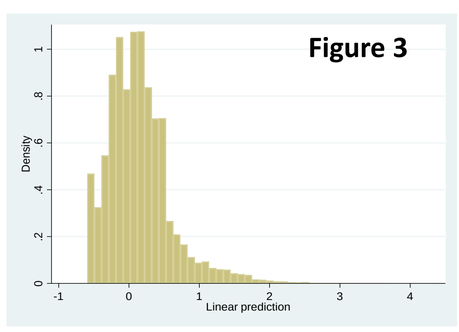
## Methods

- **Data Collection** - Large database of electronic primary care records (The Health Improvement Network, **THIN**).
- Cohort derived in part by identifying an Acceptable Electronic BCSP (**AEB**) date for each practice in England for participants aged 60-74.
- **Risk Factor/Predictor Information** – >30 clinical predictors were extracted using Read code, drug code lists and AHD strategies. Combined outcome was CRC and polyps.
- Completeness of variables and univariable analysis was investigated for this cohort.
- **Statistical Analysis –** A prediction model combining the FOBT with other clinical predictors was developed using Cox Regression and multivariable fractional polynomials with backwards elimination ('mfp' function in Stata, p=0.05).
- For internal validation, optimism adjusted performance metrics were determined using bootstrapping. Absolute risk predictions generated by estimating the baseline survival.
- Analysis was repeated for negative FOBT patients to assess whether other predictors could still warrant screening referral despite a negative result.

## Results

- **Descriptive analysis** - The screening cohort derived from THIN gave 292,168 patients and 360 practices. There were 6362 positive FOBTs and 285,806 negative FOBTs (2.2% positivity), 53.3% female, mean age 66.4.
- AEB date used for practice eligibility, data quality assurance & to define patient start dates. Example plot for a practice is shown in **Figure 2**.
- **Data completeness**: Reported symptoms (100%), smoking status (99.4%) and alcohol consumption (78%). The least complete factors included: lab results (platelet count, MCV, and haemoglobin at around 45%), and ferritin at 8.6%.
- **Univariable analysis** for >30 clinical variables identified screening factors had the strongest association for CRC. For example, previous positive FOBTs HR 5.0 (CI:4.2-6.1) and rectal bleeding had a HR of 3.1 (2.5-3.9).



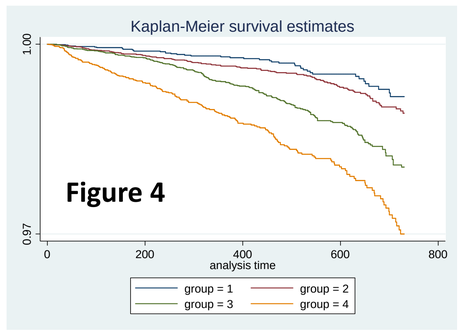




**Figure 2**: Rate of electronic BCSP notifications received. Blue line: Actual rate of BCSP notifications. Red line: Expected rate of notifications. Green line: LOWESS to assess the overall trend. AEB date: 1st September 2012.
**Figure 3**: Distribution of linear predictor (Mean:0.15, SD:0.45).
**Figure 4**: Separation of Kaplan-Meier curves for 4 risk groups based on the linear predictor.

## Results

- **Multivariable model** - The Cox Regression model including FOBT results (n=98,303, 1197 events) had 13 predictors and 2 interactions.
- The model for negative FOBT patients (n=95,792, 587 events) is given in **Table 1**.
- Predictors retained in both models: age, sex, smoking, MCV, family history GI cancer, previous polyps, abdominal pain, diarrhoea, flatulence & change in bowel habit.
- **Model performance** - **Table 2** gives the optimism adjusted performance metrics.
- **Results for the negative FOBT model**: Distribution of the LP (**Figure 3**), discrimination also assessed by analyzing separation between KM curves for 4 risk groups (**Figure 4**).
- Baseline survival after shrinkage (0.988) was estimated to give absolute risk probabilities for each individual (**Equation 1**) (Van Houwelingen's heuristic shrinkage).
- A Calibration plot is given for deciles of risk (**Figure 5**).
- Gompertz parametric model best fit for AIC (7497.7), cumulative hazard and KM plots.

| Table 2: Statistic | Optimism adjusted performance |
|---|---|
| **Model including FOBT results** | |
| C statistic | 0.850 |
| c-slope | 0.991 |
| D statistic | 2.298 |
| $R^2$ | 0.558 |
| **Model for negative FOBT patients** | |
| C statistic | 0.650 |
| c-slope | 0.944 |
| D statistic | 0.836 |
| $R^2$ | 0.144 |

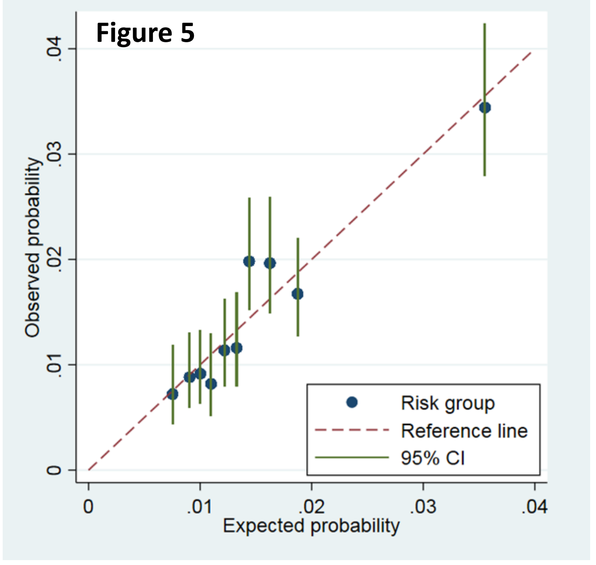



**Figure 5:** Calibration plot for negative FOBT model.
**Table 1**: Cox Regression model for negative FOBT patients (age centred -66.97), d.f. 39, EPV = 15.1.
**Table 2**: Optimism adjusted performance for both models.
**Equation 1**: Equation for absolute risk prediction.

| Table 1: Variable | Obs Coefficient (95% CI) | P Value (P>z) |
|---|---|---|
| **Smoking Status (baseline never smoked):** | | |
| Ex-smoker | 0.285 (0.080-0.491) | 0.006 |
| Current smoker | 0.516 (0.223-0.810) | 0.001 |
| **IBS** | 0.258 (0.014-0.502) | 0.039 |
| **Previous polyps** | 1.225 (0.965-1.484 | 0.000 |
| **Flatulence** | 0.953 (-0.037-1.944) | 0.059 |
| **Weight loss** | 0.867 (0.195-1.539) | 0.011 |
| **MCV<80fL (baseline ≥80fL)** | 0.877 (0.306-1.447) | 0.003 |
| **Family history GI cancer** | 0.603 (0.117-1.089) | 0.015 |
| **Abdominal pain** | 0.365 (0.117-0.612) | 0.004 |
| **Diarrhoea** | 0.572 (0.264-0.880) | 0.000 |
| **Sex (baseline female)** | -0.323 (-0.475- -0.172) | 0.000 |
| **Age at FOBT (continuous)** | 0.034 (0.015-0.053) | 0.000 |
| **Change in bowel habit** | 0.793 (0.257-1.328) | 0.004 |

| Equation 1: Survival and Event Probability | High Risk Example |
|---|---|
| **Survival Probability:** $S(2) = S_0(2)^{exp(LP)}$ | **Survival Probability:** $0.821 = 0.988^{exp(2.833)}$ |
| **Event Probability** $P = 1 - S(2)$ | **Event Probability:** 0.179 = 1-0.821 |

## Conclusions

- This research has identified predictors which could be considered for inclusion in a future risk adjusted screening model.
- Additional data could be drawn onto the BCSS to contribute to a referral algorithm
- Calculating individual risk can help referral decisions as well as patients and screening practitioners make a more informed choice.
- **Strengths:** Large sample size, internal validation to quantify optimism, development of the AEB date for data quality assurance.
- **Limitations** Sample size reduced due to complete variables (not MAR), recording of cancer could be enhanced using cancer registry data, relies on recorded symptoms.
- **Future Research** Similar analyses with the FIT & SNOMED codes. Inclusion of more lab test data (to detect intermittent & low level bleeding/longitudinal results), use of flexible parametric models.

**1.** Cooper, J. A., et al. Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model." Br J Cancer. 2018;118(2): 285-293.
**2.** Cooper, J. A., et al. FIT for the future: a case for risk-based colorectal cancer screening using the faecal immunochemical test. Colorectal Disease. 2016;18(7):650-3.

@RJenniferCooper

E-Portfolio Research Page

3 Prevention and detection of diseases

## NIHR SPARC Award and NIHR CLAHRC WM

*This research was supported by a NIHR Infrastructure Short Placement Award for Research Collaboration hosted at the University of Birmingham. This research was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care West Midlands (NIHR CLAHRC WM). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.*

NHS National Institute for Health Research