Machine learning models combining the faecal immunochemical test with routinely collected data may reduce colonoscopy demand in 35,812 pan-risk patients

Andres Tamm[1,2], Brian Shine[4], Tim James[4], Jaimie Withers[5,6], Hizni Salih[5,6], Theresa Noble[5,6], James E. East[5,7], Eva Morris[2,3], Jim Davies[2,5,8], Brian D. Nicholson[1]

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK, [2]Big Data Institute, University of Oxford, UK, [3]Nuffield Department of Population Health, University of Oxford, Oxford, UK, [4]Department of Clinical Biochemistry, John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK, [5]NIHR Oxford Biomedical Research Centre, Oxford, UK, [6]NIHR Health Informatics Collaborative, Oxford University Hospitals NHS Foundation Trust, Oxford, UK, [7]Translational Gastroenterology and Liver Unit, Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK, [8]Department of Computer Science, University of Oxford, UK

Background & Aims

- The Faecal Immunochemical Test (FIT) is used in primary care for triaging patients with suspected colorectal cancer (CRC). NICE recommends referring patients with FIT ≥ 10 µg/g [1], yet only about one in eleven FIT positives have CRC [2].
- Previous attempts to improve the precision of FIT by combining it with demographics and bloods have been restricted to traditional statistical methods and a limited number of variables [3]. This study applies machine learning (ML) models on broader sets of predictor variables.
- This poster represents an updated analysis of Oxford University Hospitals data, focussing on the more recent pan-risk patient population (patients with both low- and high-risk symptoms), who were tested with higher-quality buffered kits, to reflect current FIT test use. These are initial results on a limited set of models and will be updated subsequently.

Methods

- FITs requested by GPs (Jan 2017-May 2025) were extracted from the Oxford University Hospitals (OUH) Clinical Datawarehouse (CDW) along with linked data.
- Adults with at least one record for core bloods (haemoglobin, mean cell volume, platelets, white cells), and 180-day follow-up for CRC were included.
- Predictors were grouped as:
  - (1) FIT, age, sex, and common bloods (23 variables);
  - (2) Set 1 plus 180-day blood test time series slopes (46 variables);
  - (3) Sets 1-2 plus clinical symptoms, additional bloods and slopes, diagnoses, procedures and prescriptions (661 variables).
- ML models (**ridge regression**, **explainable boosting machine [EBM], gradient boosted decision trees [XGBoost]**) were trained on these sets.
- Models were evaluated by the percentage reduction in the number of referrals compared to FIT ≥10 µg/g, at thresholds that capture the same number of cancers as FIT (estimated from training).
- The most recent pan-risk (low/high risk) subset of patients with buffered FITs was divided into five cross-validation folds **(Fig 1),** adding older data into training folds only, to reflect current FIT use.
- The models were simultaneously optimised for high test set performance and low optimism.

Results

- 69,840 individuals (865 CRC) were included; 35,812 pan-risk patients (399 CRC) formed the primary subset. Of these, 6,039 (16.9%) had FIT ≥10 µg/g.
- At the threshold of ≥10 µg/g, FIT had 92.5% sensitivity, 84.0% specificity, 6.1% positive predictive value, and 99.9% negative predictive value.
- EBM performed the best across the three ML models: with all predictors (Set 3), mean c-statistic was 0.95 and mean average precision 0.28 (FIT alone: c-statistic 0.92, average precision 0.13, **Fig 2**). Average precision can be thought of as the average positive predictive value (PPV) across all possible classification thresholds.
- All ML models had notably higher average precision than FIT alone. However, it seems that higher average precision came mostly from risk score thresholds where the sensitivity of the model was significantly lower (<70%) than sensitivity of FIT ≥10 µg/g (**Fig 3**).
- On Set 3 (full set of predictor variables), EBM produced a mean reduction in referrals of 23.8% (SD: 6.2%), while on average missing 3% cancers compared to FIT ≥10 µg/g. The average reduction in referrals as a point estimate was higher for EBM than for ridge regression or XGBoost (**Fig 3**)**.** Reduction in referrals was computed as the percent reduction in the number of patients that test positive at a model risk score threshold compared to FIT ≥10 µg/g, when the model threshold is selected to capture the same percentage of cancers as FIT ≥10 µg/g on training data. The threshold was selected on training data to mimic real-world usage. The metric is called 'reduction in referrals', as patients with a positive test would be referred to further cancer investigations, usually colonoscopy.
- The top 10 most predictive variables in the EBM model were FIT, age, sex, change in bowel habit, albumin, platelets, serum ferritin, mean cell volume, mean cell haemoglobin, and index of multiple deprivation.

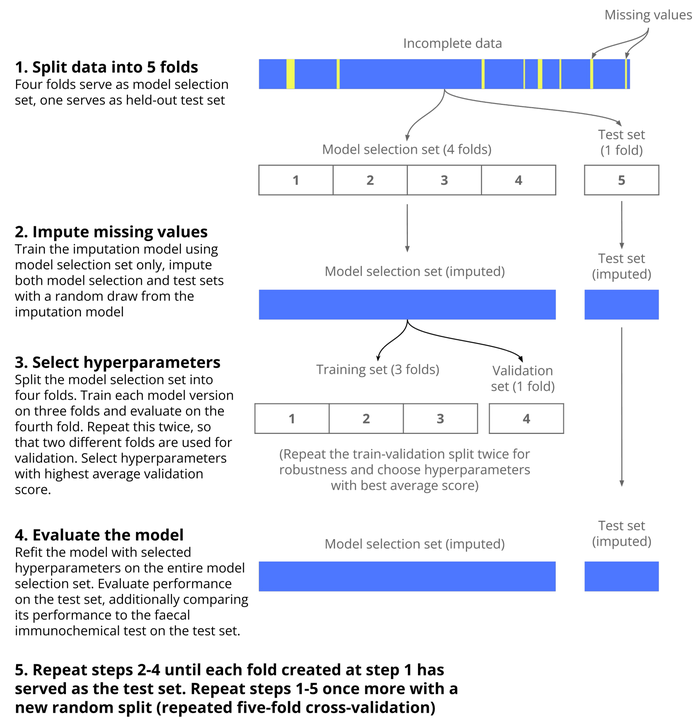

Fig 1. Model selection and validation using cross-validation with an inner hyperparameter tuning loop.

Overall discrimination

- Models have higher average PPV (precision) than FIT

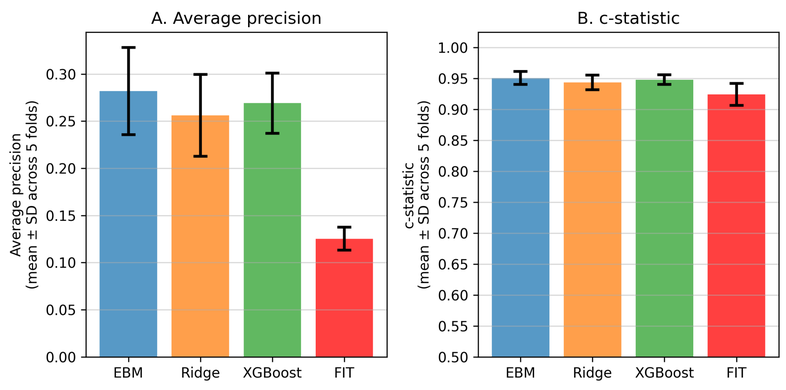

Fig 2. Average precision and c-statistic (mean ± SD across five cross-validation folds) for the machine learning models and the FIT test. The models were trained on the full set of predictor variables (Set 3).

Sensitivity-PPV trade-off

- Models have notably higher PPV at the lower-sensitivity range (sensitivity < 70%): this probably drives the higher average precision

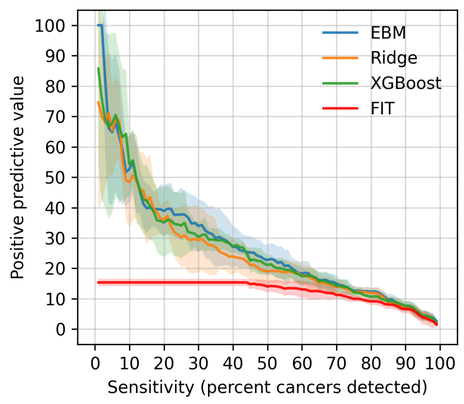

Fig 3. Precision-recall curve (sensitivity vs PPV) for the machine learning models and the FIT test. At each level of sensitivity, mean ± SD across five cross-validation folds is shown. The models were trained on the full set of predictor variables (Set 3).

Potential reduction in referrals

- The models would potentially lead to less patients testing positive than at FIT ≥10 µg/g, while potentially delaying diagnosis in about 3% cancers compared to FIT (see discussion).

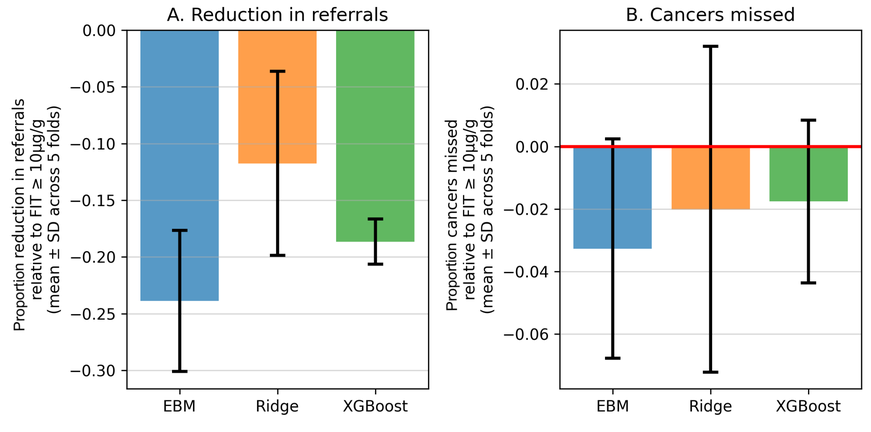

Fig 4. Proportion reduction in referrals and proportion cancers missed on held-out data (mean ± SD across five cross-validation folds) at a risk score threshold that captured the same number of cancers as FIT ≥10 µg/g on training data. The models were trained on the full set of predictor variables (Set 3). Proportions, not percentages, are shown.

Discussion & Conclusions

- ML models using routinely collected data across pan-risk patients may reduce colonoscopy demand by about 20% compared to FIT alone, with a risk of delaying diagnosis in about 3% of cancers.
- **Strengths.** This analysis includes nearly all patients with FIT in Oxfordshire since the introduction of FIT in 2017 and up to May 2025. We included a rich set of predictor variables, capturing the potential health status of patients, and employed powerful ML models that can automatically detect interactions and find cancer signals from such heterogeneous data.
- **Limitations.** The number of FIT-positive (≥10 µg/g) cancers missed by the model may be higher than 3%, as the total missed cancers may be offset by detecting additional FIT-negative cancers. Model hyperparameters were evaluated on a random sample of training data, which included a mix of buffered and non-buffered FIT kits and may have thus been suboptimal. Risks of missing pre-cancerous lesions remain unquantified. Except for blood test time series slopes, the models use cross-sectional information only.
- **Next steps.** A broader set of ML models are being trained on more data splits, using a more robust pipeline (assigning lower importance weights to non-buffered data and ensuring that hyperparameters are only evaluated on higher-quality buffered kits). Larger multi-hospital datasets, such as from the new Thames Valley and Surrey Secure Data Environment, are needed to further analyse the potential of ML models.

Acknowledgements

This work uses data provided by patients and collected by the NHS.

References

1. National Institute for Health and Care Excellence (NICE). Quantitative faecal immunochemical testing to guide colorectal cancer pathway referral in primary care. Diagnostics guidance [DG56]. https://www.nice.org.uk/guidance/dg56 (2023).
2. Booth R, Carten R, D'Souza N, Westwood M, Kleijnen J, Abulafi M. Role of the faecal immunochemical test in patients with risk-stratified suspected colorectal cancer symptoms: A systematic review and meta-analysis to inform the ACPGBI/BSG guidelines. The Lancet Regional Health–Europe. 2022 Dec 1;23.
3. Hampton JS, Kenny RP, Rees CJ, Hamilton W, Eastaugh C, Richmond C, Sharp L. The performance of FIT-based and other risk prediction models for colorectal neoplasia in symptomatic patients: a systematic review. EClinicalMedicine. 2023 Oct 1;64.

UNIVERSITY OF OXFORD
NUFFIELD DEPARTMENT OF PRIMARY CARE HEALTH SCIENCES
Medical Sciences Division