A comparison of the faecal haemoglobin concentrations and diagnostic accuracy in patients suspected with colorectal cancer and serious bowel disease as reported on four different faecal immunochemical test systems

Benton, Sally C.; Piggott, Carolyn; Zahoor, Zahida; O'Driscoll, Shane; Fraser, Callum; D'Souza, Nigel

Published in:
Clinical Chemistry and Laboratory Medicine (CCLM)

DOI:
10.1515/cclm-2021-1248

Publication date:
2022

Sally C. Benton*, Carolyn Piggott, Zahida Zahoor, Shane O’Driscoll, Callum G. Fraser, Nigel D’Souza, Michelle Chen, Theo Georgiou Delisle and Muti Abulafi

A comparison of the faecal haemoglobin concentrations and diagnostic accuracy in patients suspected with colorectal cancer and serious bowel disease as reported on four different faecal immunochemical test systems

https://doi.org/10.1515/cclm-2021-1248
Received November 29, 2021; accepted May 17, 2022;
published online June 1, 2022

Abstract

Objectives: Faecal immunochemical tests for haemoglobin (FIT) are used in colorectal cancer (CRC) screening programmes and to triage patients presenting with symptoms suggestive of CRC for further bowel investigations. There are a number of quantitative FIT analytical systems available. Currently, there is no harmonisation or standardisation of FIT methods. The aim of the study was to assess the comparability of numerical faecal haemoglobin concentrations (f-Hb) obtained with four quantitative FIT systems and the diagnostic accuracy at different f-Hb thresholds.

Methods: A subgroup of the National Institute for Health and Care Excellence (NICE) FIT study, a multicentre, prospective diagnostic accuracy study were sent four FIT specimen collection devices from four different FIT systems or two FIT devices for one FIT system. Faecal samples were examined and analysis of results carried out to assess difference between methods at thresholds of limit of detection (LoD), 10 µg haemoglobin/g faeces (µg/g) and 100 µg/g.

Results: 233 patients returned specimen collection devices for examination on four different systems; 189 patients returned two FIT kits for one system. At a threshold of 100 µg/g the sensitivity is the same for all methods. At lower thresholds of LoD and 10 µg/g differences were observed between systems in terms of patients who would be referred and diagnostic accuracies.

Conclusions: The lack of standardisation or harmonisation of FIT means that differences are observed in f-Hb generated on different systems. Further work is required to understand the clinical impact of these differences and to minimise them.

Keywords: colorectal cancer; faecal haemoglobin; faecal immunochemical test; significant bowel disease; symptomatic patients.

Introduction

Faecal immunochemical tests for haemoglobin (FIT) are used in colorectal cancer (CRC) screening programmes

The NICE FIT Steering Group: Oliver Warren (Chelsea and Westminster Hospital NHS Foundation Trust); Saidyousuf Ahmadi, Carlene Parchment, Arun Shanmuganandan (Croydon University Hospital); Nicholas West (Epsom and St Helier University Hospitals); Toni Mitchell, Stephen Sah and Nick Jackson (Hammersmith Medical Research); Alistair Myers (Hillingdon Hospitals NHS Foundation Trust); Paul Ziprin (Imperial College Healthcare NHS Trust); Ian Bloom (Kingston Hospital NHS Foundation Trust); Stan Kaye (Royal Marsden Partners); Andy Ramwell (St George’s University Hospitals NHS Foundation Trust); John T Jenkins (St Mark’s Hospital); Kevin Monahan (West Middlesex University Hospital).

***Corresponding author: Sally C. Benton**, Clinical Biochemistry, Royal Surrey County Hospital, Berkshire and Surrey Pathology Services, Guildford, Surrey, UK; and NHS Bowel Cancer Screening South of England Hub, Berkshire and Surrey Pathology Services, 20 Priestley Road, Surrey Research Park, GU2 7YS, Guildford, Surrey, UK, Phone: +44 1483 409850, E-mail: sally.benton@nhs.net. https://orcid.org/0000-0001-9230-9088

Carolyn Piggott, Zahida Zahoor and Shane O’Driscoll, NHS Bowel Cancer Screening South of England Hub, Berkshire and Surrey Pathology Services, Guildford, Surrey, UK. https://orcid.org/0000-0002-6343-6202 (C. Piggott). https://orcid.org/0000-0002-7487-7097 (Z. Zahoor). https://orcid.org/0000-0002-8693-3658 (S. O’Driscoll)

Callum G. Fraser, Centre for Research into Cancer Prevention and Screening, Population Health and Genomics, School of Medicine, University of Dundee, Scotland, UK. https://orcid.org/0000-0002-1333-7994

Nigel D’Souza, Theo Georgiou Delisle and Muti Abulafi, Department of Colorectal Surgery, Croydon University Hospital, Croydon, UK

Michelle Chen, Research and Development, RM Partners, the West London Cancer Alliance, London, UK

around the world to select those participants at highest risk of CRC for further bowel investigation, usually colonoscopy. More recently, FIT have been recommended to triage patients presenting with lower bowel symptoms for further investigation [1] and has other potential uses in the diagnosis and monitoring of CRC [2].

There are a large number of FIT systems available [3]. These include qualitative and quantitative methods, some of which can be used at home, others in clinics, and several more appropriate for use in laboratories. There is currently no harmonisation or standardisation of FIT methods, and no single primary reference material exists as yet, meaning that results from different methods are unlikely to give the same result [4–6].

In CRC screening programmes, whilst having harmonised results between FIT systems would be beneficial, the differences can be managed, at least in part, because thresholds for positivity can be set according to required referral rates, which are dependent on colonoscopy capacity [7]. With the increasing use of FIT in the assessment of patients presenting with lower bowel symptoms, particularly since the beginning of the COVID-19 pandemic, where clearly defined thresholds have been set for referral, for example in Scotland [8], an understanding of the differences between FIT systems is important and standardisation or harmonisation extremely desirable.

Globally, four commonly used quantitative FIT analytical systems are used. All have acceptable analytical performance characteristics; however, the lack of standardisation suggests that results obtained between systems will be different [9]. There are currently very limited data available to understand the differences obtained on real samples from patients between systems both from the f-Hb obtained and the impact of these differences on diagnostic accuracy when different FIT systems are used as a tool for triage of symptomatic patients [5].

In this study, we have assessed the comparability of f-Hb obtained on four quantitative FIT analytical systems and the diagnostic accuracy, at different thresholds, of these. To act as a control and to confirm if any differences observed between the four systems are due to pre-analytical sampling variation we have assessed the comparability of two samples taken for the same FIT system.

Materials and methods

Study design

The study was a sub-study of the National Institute for Health and Care Excellence (NICE) FIT study [10] and was designed to meet the Standards for Reporting of Diagnostic Accuracy Studies (STARD) Guidelines [11]. Ethics and study approval were granted from the UK Health Research Authority (IRAS 218404).

Hospital sites that had demonstrated good compliance with the NICE FIT study protocol were approached and asked to participate. Eight hospitals agreed to participate and recruit patients.

Patient selection

Patients were eligible for inclusion if they were referred from primary care with "high risk" symptoms of suspected CRC meeting NICE criteria [1] and were investigated by colonoscopy. Patients were identified by the central NICE FIT study team or local cancer research network (CRN) team after they had been booked for colonoscopy and contacted in person, by post or telephone, and invited to participate in the study. Participation was voluntary and non-participation did not affect the patient pathway. The total number of eligible patients at each site and the total number invited to participate were not recorded.

Patients who agreed to participate were randomised and allocated to one of two sampling arms (cohort 1 or cohort 2). Cohort 1 received four different FIT sample collection devices (1. Extol Hemo Auto-MC Collection Picker, Minaris Medical Co. Ltd, Tokyo, Japan; 2. OC-Auto Sampling Bottle 3, Eiken Chemical Co. Ltd, Tokyo, Japan; 3. SENTiFIT pierceTube, Sentinel Diagnostics SpA, Milan, Italy; 4. Specimen Collection Container A, Alfresa Pharma Corp., Osaka, Japan). Cohort 2 received two Extel Hemo Auto-MC Collection Picker devices. Each cohort received the relevant manufacturers' instructions to collect the samples into the devices and it was stated that the devices did not need to be used in a specific order. The Extel Hemo Auto-MC Collection Picker: HM-JACKarc system was selected for cohort two since it was the analytical system used for the main NICE FIT study.

Patients were asked to collect their samples from a single bowel motion before commencing their bowel preparation for colonoscopy, and to write the date of the sample on the return packets, before returning immediately in the Royal Mail 1st class pre-paid envelope provided.

Patients were excluded if they did not have a complete colonoscopy (unless due to CRC), or were triaged to another investigation (e.g., flexible sigmoidoscopy or computerised tomography (CT) colonography), or if they withdrew consent. In addition, patients were excluded if they did not return suitable FIT samples due to incorrect faecal collection, if samples were undated, had multiple sample dates, were received more than seven days after the sampling date, had excess faeces on the outside of the device, or the devices had leaked or were damaged in transit.

FIT sample analysis

FIT analysis was performed at a single laboratory where staff were blinded to patient clinical information. If samples were not analysed on the day of receipt they were refrigerated (4–6 °C) for up to seven days. Samples were not always analysed on the four different analysers simultaneously. A portion of the SENTiFIT pierce Tube samples (type 3) were frozen for up to 76 days prior to analysis. This was due to technical problems with the FIT system. To support whether these samples could be used in the analysis, we carried out the Mann Whitney U-test to compare results in the frozen vs. non-frozen group. In addition, to determine whether storage of specimens at minus 20 °C would affect the results in routine practice, we calculated the percentage of samples ≥10 µg/g, the most critical threshold in current clinical decision-making for all four FIT systems when no samples were frozen and when samples were frozen were examined.

One of each of the FIT systems was used to analyse the relevant samples and four research scientists performed the analyses. All analyses were overseen by a Health and Care Professions Council registered biomedical scientist and regular quality control and quality assurance procedures carried out, the laboratory being accredited to International Organization for Standardization (ISO) standards (UKAS 15189). On the day of analysis, the samples were brought to room temperature, mixed by inversion and analysed once using the relevant FIT system (Extel Hemo Auto-MC Collection Picker: HM-JACKarc; OC-Auto Sampling Bottle 3: OC-Sensor PLEDIA; SENTiFIT pierceTube/FOB Gold Wide: SENTiFIT 270; Specimen Collection Container A: NS-Prime). For f-Hb results that were above the upper limit of the measurement range (ULMR) of the methods (OC-Sensor PLEDIA: 10–200 µg/g, NS-Prime: 10–240 µg/g, SENTiFIT 270/FOB Gold Wide: 3–170 µg/g), the samples were automatically diluted by the analyser and re-examined: the dilutions applied were OC-Sensor Pledia: 1/15 and 1/250, NS-Prime: 1/100, and SENTiFIT 270: 1/10) using the appropriate FIT system diluent. The limits of detection (LoD) used in this study are based on internal analytical studies carried out according to published guidance [12]. These values have previously been reported [9].

Colonoscopy data

CRN teams at the participating hospitals collated the colonoscopy and histology data for any lesions removed or biopsies performed. These data was verified by local senior registrars.

Data analysis

Patient colonoscopy results were categorised into the most severe diagnosis; CRC, higher risk adenoma (HRA), inflammatory bowel disease (IBD), other disease of less clinical significance (lower risk adenoma, haemorrhoids, diverticular disease), or normal. HRA were defined as any polyp with high-grade dysplasia or any polyp over 10 mm in size with low-grade dysplasia, or sessile serrated lesion in the right colon [13]. CRC, HRA and IBD were combined to form the significant bowel disease (SBD) category.

For each FIT system in each cohort, the number and proportion of patients referred at thresholds of the limit of detection (LoD) of the respective FIT system 10 µg/g and 100 µg/g were calculated.

Inter-assay percentage agreements at thresholds of 10 µg/g and 100 µg/g were calculated and assessed with Cohen's kappa coefficient [14]. Because the LoD is different for each FIT system, these were not included as part of this analysis.

Data for sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and their 95% confidence intervals were calculated for f-Hb thresholds of the LoD, 10 µg/g and 100 µg/g. Receiver operating characteristic (ROC) curves were plotted for f-Hb using Analyse-it (Analyse-It Software Ltd, Leeds, UK, Software version: Method Validation edition) on a Windows 10 platform. Areas under the curve (AUC) were compared using the DeLong et al. test [15].

For comparison of f-Hb obtained on the different FIT systems, a Bland-Altman difference plot was constructed for both cohorts and the mean and median for each system calculated: f-Hb below the LoD and above the ULMR for each FIT system were excluded as were any patients without a f-Hb result on all four FIT systems; for example, if a result below the LoD or above the ULMR was generated on one or more samples, this set of samples was excluded from the data analysis.

Results

In cohort 1, 291 patients returned a set of four completed FIT system devices: 30 sets of devices were excluded and analysis not performed because they were either damaged in transit, had an illogical sample date, more than one sample date, no sample date, sample date more than 14 days from receipt in the laboratory date, or had been collected after a colonoscopy; 28 were excluded after faecal sample analysis because colonoscopy data was unavailable. Thus, 233 patients were included in the final analysis; 7 patients (3.0%) had a final diagnosis of CRC and 18 patients (7.7%) had SBD, which includes those with CRC.

For the Sentifit samples that were frozen for longer than the manufacturers stability claim, we carried out the Mann Whitney U-test to compare results in the frozen vs. non-frozen group and there was no significant difference (p=0.56). In addition when samples were not frozen, the percentage with f-Hb ≥10 µg/g, the most critical f-Hb threshold for assessment of patients with symptoms was 24.3%. When frozen samples examined by SENTiFIT were studied separately, and all four systems compared, 25.5% had f-Hb ≥10 µg/g. These frozen samples have been included in the analysis.

In cohort 2, 189 patients returned two devices for examination on the HM-JACKarc system. 16 patients were excluded before faecal sample analysis and 21 patients were excluded after analysis for the same reasons listed as for cohort 1 above. 152 patients were included in the final analysis. 1 patient (0.7%) had a final diagnosis of CRC and 15 patients (9.7%) had SBD, which includes the patient with CRC.

Cohort 1

The numbers and percentages of patients with results above the FIT system LoD and greater than 10 µg/g and 100 µg/g or above are shown in Table 1. At a threshold of 10 µg/g approximately 50% of patients would have a positive (i.e., above the threshold) f-Hb result if examined on the FOB Gold Wide® as compared to the other three FIT systems.

Tables 2A and 2B show the Cohen's kappa coefficient at thresholds of 10 µg/g and 100 µg/g respectively. At a threshold of 10 µg/g, a substantial agreement was observed between HM-JACKarc, OC-Sensor PLEDIA and NS-Prime and a moderate agreement of SENTiFIT 270/FOB Gold

Wide® with f-Hb generated by the other three FIT systems. At a threshold of 100 µg/g, an “almost perfect” agreement was observed between the results obtained from all systems except SENTiFIT 270/FOB Gold Wide® and NS-Prime for which a “substantial” agreement was observed.

Table 1: Numbers (percentage of total) of patients with results above each threshold.

FIT system	HM-JACKarc	OC-Sensor PLEDIA	NS-Prime	SENTiFIT 270/ FOB Gold Wide
Results >LoD	62 (27.0%)	100 (43.0%)	119 (51.0%)	30 (13.0%)
Results ≥10 µg/g	39 (17.0%)	37 (16.0%)	42 (18.0%)	19 (8.0%)
Results >100 µg/g	10 (4.0%)	12 (5.0%)	10 (4.0%)	11 (5.0%)

LoD, limit of detection: HM-JACKarc=2 µg/g; OC-Sensor PLEDIA=1 µg/g; NS-Prime=3 µg/g; SENTiFIT 270/Gold Wide=2 µg/g.

Table 2A: Cohen’s kappa coefficient at a threshold of 10 µg/g.

FIT system 1	FIT system 2	Cohen’s kappa	Interpretation
HM-JACKarc	OC-Sensor PLEDIA	0.7686	Substantial agreement
HM-JACKarc	NS-Prime	0.746	Substantial agreement
HM-JACKarc	SENTiFIT 270/ FOB Gold Wide	0.497	Moderate agreement
NS-Prime	OC-Sensor PLEDIA	0.711	Substantial agreement
SENTiFIT 270/FOB Gold Wide	OC-Sensor PLEDIA	0.48	Moderate agreement
SENTiFIT 270/FOB Gold Wide	NS-Prime	0.575	Moderate agreement

Table 2B: Cohens kappa coefficient at a threshold of 100 µg/g.

FIT system 1	FIT system 2	Cohen’s kappa	Interpretation
HM-JACKarc	OC-Sensor PLEDIA	0.8960.905	Almost perfect agreement
HM-JACKarc	NS-Prime	0.896	Almost perfect agreement
HM-JACKarc	SENTiFIT 270/ FOB Gold Wide	0.850	Almost perfect agreement
NS-Prime	OC-Sensor PLEDIA	0.904	Almost perfect agreement
SENTiFIT 270/ FOB Gold Wide	OC-Sensor PLEDIA	0.863	Almost perfect agreement
SENTiFIT 270/ FOB Gold Wide	NS-Prime	0.751	Substantial agreement

The indices of diagnostic accuracy of f-Hb analysis on each of the four FIT systems for CRC at the LoD, 10 µg/g and 100 µg/g are summarised in Table 3A and similar data for SBD is in Table 3B.

The AUC for each FIT system for CRC and SBD are shown in Table 4.

For CRC, statistical analysis demonstrated that the AUC for NS-Prime is significantly different from 0.5 ($p<0.05$). For HM-JACKarc, OC-Sensor PLEDIA, and FOB Gold Wide the AUC is statistically equal to 0.5 and so there is no discrimination at the 5% significance level ($p>0.05$).

Statistical analysis demonstrates that, for SBD, the results from the HM-JACKarc, OC-Sensor PLEDIA, and SENTiFIT270/FOB Gold Wide, the AUC were significantly different from 0.5 with $p<0.05$. The AUC for NS-Prime showed no statistical discrimination at the 5% significance level ($p=0.11$).

At a 5% significance level, the De Long method for comparison of tests showed that the AUC for both CRC and SBD show equivalence to one another ($p<0.05$ for all comparisons).

Table 5 shows the f-Hb obtained on each FIT system for the 18 patients diagnosed with CRC and SBD. Four of the seven patients with CRC had results >100 µg/g on all FIT systems; one of seven had f-Hb <LoD on all methods; one of seven had <LoD on all methods except NS-Prime and one of seven had >10 µg/g on OC Sensor PLEDIA and NS-Prime but <LoD on HM-JACKarc and SENTiFIT 270/FOB Gold Wide.

After excluding f-Hb outside the LoD and ULMR there were only 14 paired data points. These results were plotted in a Bland-Altman plot (Figure 1A). The median values of the 14 data points for the HM-JACKarc, OC-Sensor PLEDIA, NS-Prime and SENTiFIT 270/FOB Gold Wide respectively were 16 µg/g, 14 µg/g, 17 µg/g and 5 µg/g (Figure 1B).

Cohort 2

There was only one CRC detected in cohort 2 and thus insufficient data for statistical analysis regarding CRC: the two samples obtained on this patient were concordant and both had f-Hb above the ULMR; 15 patients had SBD. Applying thresholds of LoD (2 µg/g), 10 µg/g and 100 µg/g, HM-JACK arc 1 had 52 (34.0%) 44 (29.0%) and 27 (18.0%) f-Hb results above these thresholds, respectively, and HM-JACKarc 2 had 29 (19.0%) 10 (7.0%) and 12 (8.0%).

For SBD, the Cohen’s kappa at 10 µ/g was 0.737 and at 100 µg/g was 0.717 which both demonstrate substantial agreement.

Table 3A: Sensitivity, specificity, positive predictive value and negative predictive value for CRC at limit of detection (LoD), 10 μg/g and 100 μg/g on the four FIT systems.

Method	≥Threshold	% Sensitivity (95% CI)	% Specificity (95% CI)	Positive predictive value, % (95% CI)	Negative predictive value, % (95% CI)
HM-JACKarc	LoD (2 μg/g)	57.1 (25.1–84.2)	68.1 (61.8–73.9)	5.3 (2.8–9.8)	98.1 (95.6–99.2)
	10 μg/g	57.1 (25.1–84.2)	84.5 (79.2–88.6)	10.3 (5.3–18.9)	98.5 (96.4–99.3)
	100 μg/g	57.1 (25.1–84.2)	97.3 (94.3–98.8)	40.0 (19.4–64.8)	98.7 (96.9–99.4)
OC Sensor PLEDIA	LoD (1 μg/g)	71.4 (35.9–91.8)	45.6 (39.2–52.1)	3.9 (2.4–6.2)	98.1 (94.1–99.4)
	10 μg/g	71.4 (35.9–91.8)	85.8 (80.7–89.8)	13.5 (8.1–21.6)	99.0 (96.8–99.7)
	100 μg/g	57.1 (25.1–84.2)	96.5 (93.2–98.2)	33.3 (16.4–56.0)	98.6 (96.9–99.4)
NS-Prime	LoD (3 μg/g)	85.7 (48.7–97.4)	31.9 (26.1–38.2)	3.8 (2.8–5.1)	98.6 (92.1–99.8)
	10 μg/g	71.4 (35.9–91.8)	83.6 (78.2–87.9)	11.9 (7.2–19.0)	99.0 (96.7–99.7)
	100 μg/g	57.1 (25.1–84.2)	97.3 (94.3–98.8)	40.0 (19.4–64.8)	98.7 (96.9–99.4)
SENTiFIT 270/FOB Gold Wide	LoD (2 μg/g)	57.1 (25.1–84.2)	82.3 (76.8–86.7)	9.1 (4.7–16.8)	98.4 (96.3–99.3)
	10 μg/g	57.1 (25.1–84.2)	93.4 (89.3–95.9)	21.1 (10.6–37.4)	98.6 (96.8–99.4)
	100 μg/g	57.1 (25.1–84.2)	96.9 (93.7–98.5)	36.4 (17.8–60.1)	98.6 (96.9–99.4)

Table 3B: Sensitivity, specificity, positive predictive value and negative predictive value for serious bowel disease at limit of detection (LoD), 10 μg/g and 100 μg/g on four FIT systems.

FIT system	≥Threshold	% Sensitivity (95% CI)	% Specificity (95% CI)	Positive predictive value, % (95% CI)	Negative predictive value, % (95% CI)
HM-JACKarc	LoD (2 μg/g)	61.1 (38.6–79.7)	69.8 (63.3–75.5)	14.5 (10.0–20.5)	95.5 (92.3–97.5)
	10 μg/g	55.6 (33.7–75.4)	86.5 (81.3–90.4)	25.6 (16.8–37.0)	95.9 (93.3–97.5)
	100 μg/g	38.9 (20.3–61.4)	98.6 (96.0–99.5)	70.0 (39.7–89.2)	95.1 (93.0–96.5)
OC-Sensor PLEDIA	LoD (1 μg/g)	66.7 (43.8–83.7)	46.0 (39.5–52.7)	9.4 (6.8–12.8)	94.3 (89.4–97.0)
	10 μg/g	50.0 (29.0–71.0)	87.0 (81.8–90.8)	24.3 (15.3–36.4)	95.4 (92.9–97.1)
	100 μg/g	38.9 (20.3–61.4)	97.7 (94.7–99.0)	58.3 (33.1–79.9)	95.0 (93.0–96.5)
NS-Prime	LoD (3 μg/g)	72.2 (49.1–87.5)	31.6 (25.8–38.1)	8.1 (6.2–10.7)	93.2 (86.3–96.7)
	10 μg/g	50.0 (29.0–71.0)	84.7 (79.2–88.9)	21.4 (13.5–32.3)	95.3 (92.7–97.0)
	100	38.9 (20.3–61.4)	98.6 (96.0–99.5)	70.0 (39.7–89.2)	95.1 (93.0–96.5)
SENTiFIT 270/FOB Gold Wide	LoD (2 μg/g)	44.4 (24.6–66.3)	83.3 (77.7–87.7)	18.2 (10.9–28.8)	94.7 (92.2–96.5)
	10 μg/g	44.4 (24.6–66.3)	94.9 (91.1–97.1)	42.1 (25.1–61.2)	95.3 (93.1–96.9)
	100 μg/g	33.3 (16.3–56.3)	97.7 (94.7–99.0)	54.6 (28.9–78.0)	94.6 (92.7–96.0)

Table 4: AUC for each of the FIT systems for colorectal cancer (CRC) and significant bowel disease (SBD) in cohort.

FIT system	AUC for CRC (95% CI)	AUC for SBD (95% CI)
HM-JACKarc	0.71 (0.44–0.98)	0.71 (0.56–0.87)
OC-Sensor PLEDIA	0.76 (0.49–1.03)	0.68 (0.52–0.85)
NS-Prime	0.79 (0.53–1.06)	0.65 (0.47–0.83)
SENTiFIT 270/FOB Gold Wide	0.73 (0.48–0.97)	0.67 (0.53–0.82)

The indices of diagnostic accuracy of f-Hb analysis on each of the samples for SBD at the LoD, 10 μg/g and 100 μg/g are summarised in Table 6.

There was an AUC for SBD of 0.70 (CI: 0.54–0.86) for HM-JACKarc 1 and 0.69 (CI: 0.54–0.85) for HM-JACKarc 2. Both results were significantly different from 0.50 with a $p<0.05$. The De long method for comparison of tests demonstrate that the AUC were equivalent ($p<0.05$).

After excluding f-Hb results outside the LoD and ULMR there were 26 paired data points. These results were plotted

Table 5: f-Hb concentrations from each analyser for each patient with colorectal cancer (CRC) and significant bowel disease (SBD).

2	>8,001	11,929	12,380	>1701	CRC
3	1,008	1978	2,267	469	CRC
4	188	417	1,015	132	CRC
5	<2	<1	<3	<2	CRC
6	<2	<1	5	<2	CRC
7	<2	13	18	<2	CRC
8	4,046	16,868	3,087	1701	Adenoma with high grade dysplasia
9	<2	<1	<3	<2	Adenoma with high grade dysplasia
10	<2	<1	<3	<2	Adenoma with high grade dysplasia
11	<2	<1	<3	<2	Sessile right side
12	4,368	6,089	3,415	630	Adenoma low grade dysplasia >10 mm
13	37	2	3	<2	Adenoma low grade dysplasia >10 mm
14	18	7	4	<2	Adenoma low grade dysplasia >10 mm
15	<2	4	<3	<2	Adenoma low grade dysplasia >10 mm
16	2	<1	5	<2	Adenoma low grade dysplasia >10 mm
17	2,496	840	743	43	Inflammatory bowel disease
18	70	10	53	20	Inflammatory bowel disease

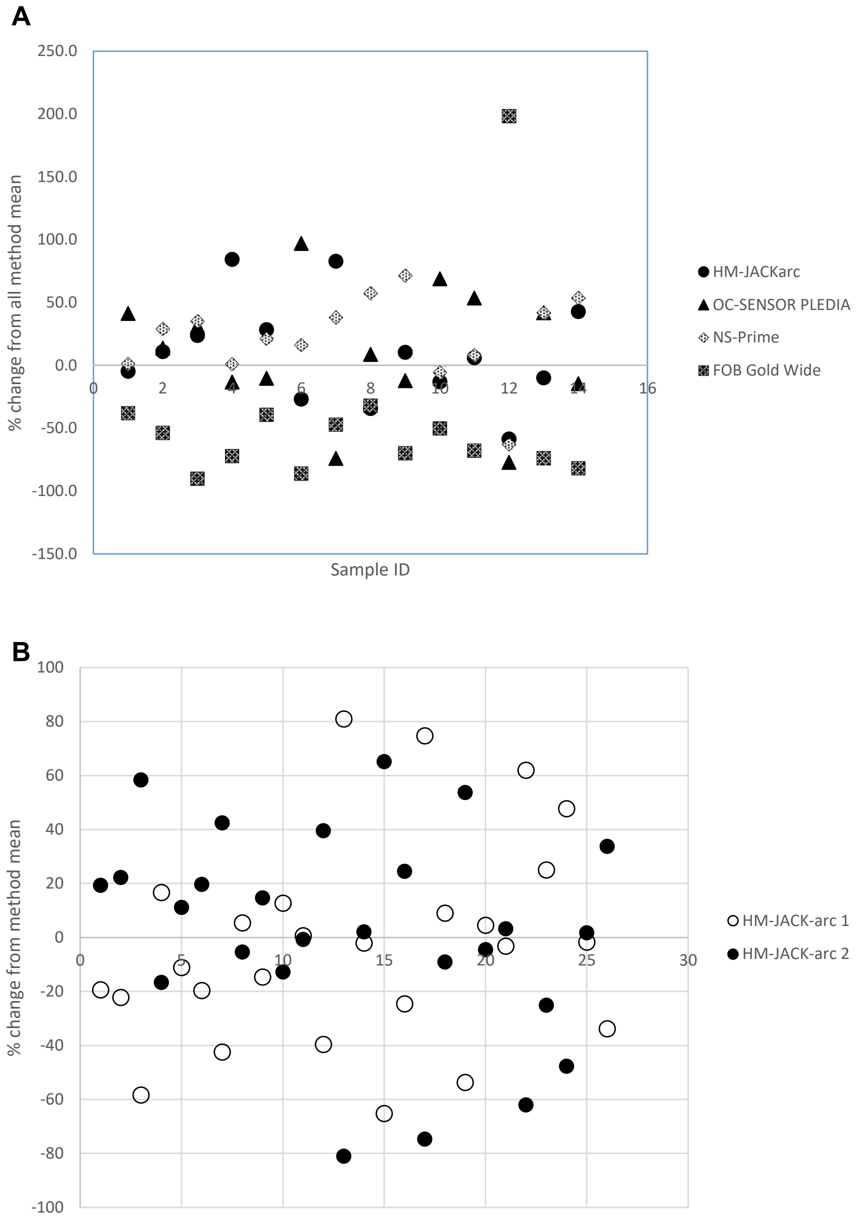

Figure 1: Bland-Altman difference plot showing the percentage change from the all method mean of the (A) 14 patients who had all four samples with results within the measurement range of the systems (B) 26 patients who had two samples with faecal haemoglobin concentrations within the measurement range.

Table 6: Sensitivity, specificity, positive predictive value and negative predictive value for serious bowel disease at limit of detection (LoD) 10 µg/g and 100 µg/g on two separate samples tested on the same HM-JACKarc analytical system.

FIT system	≥Threshold, µg/g	Sensitivity, %	Specificity, %	Positive predictive value, %	Negative predictive value, %
HM-JACKarc (1)	LoD (2 µg/g)	66.7 (41.7–84.8)	61.3 (53.0–69.1)	15.9 (11.1–22.2)	94.4 (89.0–97.2)
	10 µg/g	46.7 (24.8–69.9)	85.4 (78.5–90.3)	25.9 (15.1–40.8)	93.6 (90.1–95.9)
	100 µg/g	26.7 (10.9–52.0)	95.6 (90.8–98.0)	40.0 (17.5–67.7)	92.3 (89.8–94.2)
HM-JACKarc (2)	LoD (2 µg/g)	60.0 (35.8–80.2)	64.2 (55.9–71.8)	15.5 (10.3–22.7)	93.6 (88.6–96.5)
	10 µg/g	46.7 (24.8–69.9)	83.9 (76.9–89.2)	24.1 (14.1–38.1)	93.5 (89.9–95.9)
	100 µg/g	33.3 (15.2–58.3)	94.9 (89.8–97.5)	41.7 (20.5–66.4)	92.9 (90.1–94.9)

in a Bland-Altman plot (Figure 1B). The median values of the 26 data points for the HM-JACKarc 1 and HM-JACKarc 2 groups were 28 µg/g and 27 µg/g respectively.

Discussion

The lack of standardisation or harmonisation of FIT methods has raised questions as to whether the number of patients presenting with suspected CRC symptoms referred for further investigation, and hence CRC detection rate, will vary on the FIT system used. To obtain data on this is challenging;

- f-Hb is unstable once the bowel motion has been passed and so samples must be collected by patients directly into manufacturer specific devices. FIT devices can only be analysed on the corresponding manufacturer's system.
- Due to the heterogenic nature of passed faeces, collecting samples from the same bowel motion in to separate devices will mean that, because different parts of the faeces are being sampled, the f-Hb is unlikely to be identical.

In this study, four samples collected from the same bowel motion from 233 patients were received to examine whether the indices of diagnostic accuracy of four FIT systems were equivalent and whether any bias in f-Hb concentration was observed between the systems. To our knowledge, this is the first study to include all four routinely used quantitative FIT analytical systems in a comparison study using samples from patients.

In cohort 1, the results demonstrated that a large difference would be observed in referral patterns depending on which FIT system was used at thresholds of LoD and 10 µg/g. At a threshold of 10 µg/g, fewer than half the number of referrals would be made if the SENTiFIT 270/FOB Gold Wide system was employed as compared to the other methods and dramatically fewer at the LoD. This indicates that the calibration for the SENTiFIT 270/FOB Gold Wide gives lower f-Hb results than the other three systems. This observation is supported when the numerical f-Hb results are compared using a Bland Altman Difference plot (Figure 1).

Translating this to accurate diagnosis of SBD, in such patients, the SENTiFIT 270/FOB Gold Wide, along with the HM-JACKarc would not have resulted in the referral of one patient with CRC at a threshold of 10 µg/g that the f-Hb from NS-Prime and OC-Sensor PLEDIA would have referred. It is widely accepted, as recently documented in detail [2] that f-Hb is not the sole criterion for referral but is considered along with presenting symptoms and other findings such as results of the full blood count. f-Hb concentrations are positively associated with severity of disease [16] and, at a higher threshold of 100 µg/g, the sensitivity for both CRC and SBD was the same for all methods. Three CRC and seven HRA (Table 6) would have been missed by all four systems.

A challenge in the comparison of FIT systems is the large pre-analytical variation that is inherent in the test [4]. To try to account for the impact of this, we invited a second cohort of individuals to return two FIT devices for a single method (HM-JACKarc). In cohort 2 there was not perfect agreement between two results from the same sample on the FIT systems, although substantial agreement was observed. As such, we could consider a Cohen's kappa of substantial agreement acceptable in terms of the diagnostic accuracy and that anything less than this could be considered variation outside that explicable by pre-analytical and analytical variation. In this case, we would consider the moderate agreement observed when comparing the SENTiFIT 270/FOB Gold Wide method to be a significant outcome deviation compared to the other three systems.

FIT is a diagnostic adjunct, and there will be clinical indications for referral for further investigation despite the f-Hb. Nonetheless, an individual f-Hb result will determine further management and, as such, the risks and benefits of the differences between FIT systems need to be thoroughly explored.

In a similar study that included 732 symptomatic patients who returned both HM-JACKarc and OC-Sensor PLEDIA devices [5], HM-JACKarc was reported to have a lower sensitivity for CRC and overall lower mean f-Hb for HM-JACKarc as compared to OC-Sensor. In our study, the median f-Hb for HM-JACKarc was higher than OC-Sensor PLEDIA and the mean very similar (26 µg/g vs. 27 µg/g). A recent meta-analysis [17] reported that "FIT brand" (HM-JACKarc or OC-Sensor) was also a significant predictor of heterogeneity of specificity. However, this study concluded that, whilst the meta-regression analysis suggested statistically significant difference between HM-JACKarc and OC-Sensor at a threshold of 10 µg/g that these are clinically irrelevant and could be explained by different study strategies. They also considered that, because approximately 90% of CRC detected have f-Hb above 100 µg/g, it is unlikely that further information will show clinically significant differences between FIT systems.

The main limitation of our study is the small sample size in both cohorts. This was an opportunistic sub-study appended to a much larger clinical study and the focus of the recruitment was to the main clinical study. The numbers of patients with CRC and SBD were small and the assumption was made that patients followed the instructions for sample collection and did collect the relevant number of samples from the same bowel motion. The limited data and hence wide CI make conclusive interpretation of the AUC difficult.

In this very small cohort, it is difficult to be conclusive about the clinical impact and significance of the differences observed between the FIT systems. In cohort 1, the results demonstrated that a large difference was observed in referral patterns dependent on which FIT system was used at thresholds of LoD and 10 µg/g. The results obtained on the SENT-i-FIT system are numerically different at both the LoD and >10 µg/g along with the actual numeric result of some patients with SBD. Data on cohort 2 demonstrates that variation does occur even when exactly the same FIT system is employed.

The lack of standardisation or harmonisation of FIT means that the differences observed are to be expected and it highlights the urgency with which these improvements are required. We have shared the results of our study with the suppliers of all four systems studied and trust that beneficial quality improvements, as highlighted in our study, will lead to the desirable further harmonisation between systems as recommended [18].

Acknowledgments: The study supported by Croydon University Hospital and RM Partners in the design and conduct of the trial. Patients were recruited nationally through the NIHR Clinical Research Network, Principal Investigators and R&D teams at each site. Alpha Labs Ltd supported the study prior to funding with FIT kits and production of patient information. We thank Professor Stephen Duffy for his guidance with some of the statistical methodology.

Research funding: This study is a sub-study of the NICE FIT study which was funded by an NHS England award to RM Partners, the West London Cancer Alliance hosted by The Royal Marsden NHS Foundation Trust. The study was supported by the National Institute for Health Research Clinical Research Network Portfolio. Alpha Labs Ltd, MAST Diagnostics, Sysmex and Abbott/ Alfresa supported the study by providing FIT kits and reagents without charge. The study funders had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Author contributions: SCB drafted the manuscript. ZZ and CP carried out statistical analysis. CP, SOD and ZZ carried out laboratory analysis of samples. CGF provided guidance on statistical analysis. All authors provided significant input to the study, reviewed and revised drafts of the manuscript, and approved the submitted version. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Informed consent was obtained from all individuals included in this study.

Ethical approval: Research involving human subjects complied with all relevant national regulations, institutional policies and is in accordance with the tenets of the Helsinki Declaration (as revised in 2013). The study was approved by the National Research Ethics Service Committee, London- South East. Full title: NICE Guidelines and The Faecal Immunochemical Test Study. REC No: 16/LO/2,174. IRAS project ID 218404.

Principal investigators: Mark Austin (Brighton and Sussex University Hospitals NHS Trust); Gemma Faulkner (Bolton NHS Foundation Trust); John Stebbing (Royal Surrey NHS Foundation Trust); Jonathan Epstein (Salford Royal NHS Foundation Trust).

Patient consent statement: No identifiable patient data is present in the manuscript, Permission to reproduce material from other sources, No material was reproduced from other sources.

Permission to reproduce material from other sources: No material was reproduced from other sources.

Clinical trial registration: ISRCTN49676259.

References

1. National Institute for Health and Care Excellence. Suspected cancer: recognition and referral. In: NICE Guideline (NG12). London: National Institute for Health and Care Excellence; 2017.
2. Pin Vieito N, Puga M, Fernandez-de-Castro D, Cubiella J. Faecal immunochemical test outside colorectal cancer screening? World J Gastroenterol 2021;27:6415–29.
3. Daly JM, Xu Y, Levy BT. Which fecal immunochemical test should I choose? J Prim Care Community Health 2017;8:264–77.
4. Benton SC, Symonds E, Djedovic N, Jones S, Deprez L, Kocna P, et al. Faecal immunochemical tests for haemoglobin: analytical challenges and potential solutions. Clin Chim Acta 2021;517: 60–5.
5. Chapman CJ, Banerjea A, Humes DJ, Allen J, Oliver S, Ford A, et al. Choice of faecal immunochemical test matters: comparison of OC-Sensor and HM-JACKarc in the assessment of patients at high risk of colorectal cancer. Clin Chem Lab Med 2020;59:721–8.
6. Clark G, Strachan JA, Carey FA, Godfrey T, Irvine A, McPherson A, et al. Transition to quantitative faecal immunochemical testing from guaiac faecal occult blood testing in a fully rolled-out population-based national bowel screening programme. Gut 2021;70:106–13.
7. Moss S, Mathews C, Day TJ, Smith S, Seaman HE, Snowball J. 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 2017;66:1631–44.
8. Scottish Government. Guidance for the use of FIT in the prioritization of patients with colorectal symptoms now and in the recovery period after COVID; 2020. Available from: https://www.nhsggc.org.uk/media/261915/clinical-guidance-quantitative-faecal-immunochemical-testing-004.pdf.
9. Piggott C, Carroll MRR, John C, O'Driscoll S, Benton SC. Analytical evaluation of four faecal immunochemistry tests for haemoglobin. Clin Chem Lab Med 2020;59:173–8.
10. D'Souza N, Delisle TG, Chen M, Benton SC, Abulafi M, Warren O, et al. Faecal immunochemical testing in symptomatic patients to prioritize investigation: diagnostic accuracy from NICE FIT Study. Br J Surg 2021;23:804–10.
11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351: h5527.
12. Fraser CG, Benton SC. Detection capability of quantitative faecal immunochemical tests for haemoglobin (FIT) and reporting of low faecal haemoglobin concentrations. Clin Chem Lab Med 2019;57: 611–6.
13. D'Souza N, Georgiou Delisle T, Chen M, Benton S, Abulafi M. Faecal immunochemical test is superior to symptoms in predicting pathology in patients with suspected colorectal cancer symptoms referred on a 2WW pathway: a diagnostic accuracy study. Gut 2020;70:1130–8.
14. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Soc Adm Phar 2013;9:330–8.
15. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
16. Digby J, Cleary S, Gray L, Datt P, Goudie DR, Steele RJC, et al. Faecal haemoglobin can define risk of colorectal neoplasia at surveillance colonoscopy in patients at increased risk of colorectal cancer. U Eur Gastroenterol J 2020;8:559–66.
17. Pin-Vieito N, Tejido-Sandoval C, de Vicente-Bielza N, Sánchez-Gómez C, Cubiella J. Faecal immunochemical tests safely enhance rational use of resources during the assessment of suspected symptomatic colorectal cancer in primary care: systematic review and meta-analysis. Gut 2022;71:950–60.
18. Benton SC, Symonds E, Djedovic N, Jones S, Deprez L, Kocna P, et al. Faecal immunochemical tests for haemoglobin: analytical challenges and potential solutions. Clin Chim Acta 2021;517: 60–5.