Population-based colorectal cancer screening by fecal immunochemical testing over multiple rounds
van der Vlugt, M.

Citation for published version (APA):
A randomised comparison of two faecal immunochemical tests in population-based colorectal cancer screening

Esmée J. Grobbee, Manon van der Vlugt, Hanneke J. van Vuuren, Ann K. Stroobants, Marco W. Mundt, Wolfert J. Spijker, Evelien J.C. Bongers, Ernst J. Kuipers, Iris Lansdorp-Vogelaar, Patrick M. Bossuyt, Evelien Dekker, Manon C.W. Spaander

Gut, 2016
ABSTRACT

Objective

Colorectal cancer screening programmes are implemented worldwide; many are based on faecal immunochemical testing (FIT). The aim of this study was to evaluate two frequently used FITs on participation, usability, positivity rate and diagnostic yield in population-based FIT screening.

Design

Comparison of two FITs was performed in a fourth round population-based FIT screening cohort. Randomly selected individuals aged 50-74 were invited for FIT screening and were randomly allocated to receive an OC-Sensor (Eiken, Japan) or FOB-Gold (Sentinel, Italy) test (March-December 2014). A cut-off of 10 µg haemoglobin (Hb)/g faeces (ie, 50 ng Hb/mL buffer for OC-Sensor and 59 ng Hb for FOB-Gold) was used for both FITs.

Results

In total, 19,291 eligible invitees were included (median age 61, IQR 57-67; 48% males): 9,669 invitees received OC-Sensor and 9,622 FOB-Gold; both tests were returned by 63% of invitees (p=0.96). Tests were non-analysable in 0.7% of participants using OC-Sensor vs 2.0% using FOB-Gold (p<0.001). Positivity rate was 7.9% for OC-Sensor, and 6.5% for FOB-Gold (p=0.002). There was no significant difference in diagnostic yield of advanced neoplasia (1.4% for OC-Sensor vs 1.2% for FOB-Gold; p=0.15) or positive predictive value (PPV; 31% vs 32%; p=0.80). When comparing both tests at the same positivity rate instead of cut-off, they yielded similar PPV and detection rates.

Conclusions

The OC-Sensor and FOB-Gold were equally acceptable to a screening population. However, FOB-Gold was prone to more non-analysable tests. Comparison between FIT brands is usually done at the same Hb stool concentration. Our findings imply that for a fair comparison on diagnostic yield between FIT’s positivity rate rather than Hb concentration should be used.
INTRODUCTION

Colorectal cancer (CRC) is one of the major causes of death in the Western World [1-3]. Population-based CRC screening aims to detect CRC and its precursors in an earlier phase, thereby reducing CRC morbidity and mortality [4]. Of the currently available screening tests, faecal occult blood tests (FOBTs) and sigmoidoscopy are the only strategies that have been proven in prospective randomised controlled trials to reduce CRC-related mortality [5]. The evidence favouring faecal immunochemical testing (FIT) over guaiac faecal occult blood testing (gFOBT) is substantial [6-11]. FIT is more likely to detect haemoglobin (Hb) from the lower GI tract than most gFOBTs. FIT is also easier to use resulting in higher participation rates. In addition, FIT enables quantitative measuring faecal Hb, allowing the use of different cut-off concentrations, and can be analysed by automation.

Many countries have implemented FIT-based CRC screening programmes or are about to do so [12]. When implementing a national screening programme, choosing the appropriate test is of key importance. FIT can be based on dry sampling, using cards, or wet sampling, using tubes with buffers [8]. The most frequently used FITs involve wet sampling (ie, storage and transport of faeces in a wet preservative), and different FIT brands in this category are available. These various FIT brands often have different sampling tubes and buffer volumes, resulting in different expressions of Hb concentration that are not interchangeable [13]. Such differences complicate direct comparison of FITs. It has been proposed to standardise quantitative FIT results in μg Hb per gram faeces, allowing diagnostic test accuracy to be compared more easily between FITs using the same standardised cut-off [14]. At present, there is no evidence favouring a specific FIT brand [15].

As small differences in test characteristics can have major effects on a population level, it is important to further assess brand-related differences. In the Netherlands, several pilot studies have been performed using the OC-Sensor (Eiken, Japan) [16-18]. However, the recently started nationwide programme has selected the FOB-Gold test (Sentinel, Italy) through a European bid. Few comparative data are available for these two tests. To the best of our knowledge, no study has investigated the OC-Sensor and FOB-Gold head to head in a screening setting. Such a comparison is relevant since both tests are among the most widely used FITs. We therefore aimed to compare the OC-Sensor and FOB-Gold with regard to participation rate, usability, positivity rate (PR) and diagnostic yield, using a standardised cut-off of faecal Hb concentration.

METHODS

Study population/study design

Details about the design of the ongoing population-based CRC screening pilot programme have been described previously [7,16,17,19,20]. This trial was registered at http://www.trialregister.nl (identifier NTR5385). In short, from June 2006 the demographic data of all individuals between 50 and 74 years living in a selected region in the southwest and northwest of the Netherlands were obtained from municipal population registers. As there was no CRC screening programme at the time of the trial the target population was screening-naïve when first approached.

This study was part of a dynamic (ie, including all subjects included in any of the first three rounds as well as individuals who moved into the target areas and those who had reached the target
age) cohort study. Individuals who had moved out of the selected area or passed the upper-age limit were not reinvited. Invitations for this fourth screening round were done similar to previous rounds. This implied that for the Southern region, random samples were taken from the target population by a computer-generated algorithm (Tenalea, Amsterdam, The Netherlands). In the Northern region random samples of selected postal code areas were taken. For this fourth round of screening both cohorts were combined. Invitations were sent out between March 2014 and December 2014. Individuals with a history of IBD or CRC, as well as those who had undergone a colonoscopy in the past 2 years, had an estimated life expectancy of <5 years, or were unable to give informed consent were excluded from the study.

**Intervention and randomisation; two types of FIT-screening**

All invitees were randomly allocated to receive either the OC-Sensor (Eiken, Japan) or the FOB-Gold (Sentinel, Italy) sorted according to household without stratification. Randomisation was performed before invitation in a 1:1 ratio. People in the same household were allocated to the same test, to avoid confusion in handling the FIT. The FIT kit was sent by mail with the instructions to collect a single sample of one bowel movement using a collection device probe. The test result was considered positive when the Hb concentration in the FIT sample was $\geq 10$ µg Hb/g faeces. This cut-off corresponds to 50 ng Hb/mL for the OC-Sensor and 59 ng Hb/mL for the FOB-Gold. In case of an non-analysable test, a new FIT was sent to the participant.

**FIT analysis**

The OC-Sensor device collected 10 mg faeces with a serrated probe attached to the cap in 2.0 mL preservative buffer. The FOB-Gold pierceable tube collected 10 mg faeces with a serrated probe attached to the cap in 1.7 mL preservative buffer. Participants were asked to dip the probe four times in the faeces and to reinsert the probe into the respective device. Participants were asked to post the faeces samples within 24 hours after collection and keep the sample in the refrigerator. Together with the FIT participants signed an informed consent form with the date of sample collection. Participants returned the FIT and permission form at ambient temperature by freepost to two laboratories (Laboratory Clinic Chemistry, Academic Medical Centre, Amsterdam or Gastroenterology & Hepatology laboratory, Erasmus Medical Centre, Rotterdam, the Netherlands).

At arrival in the laboratory, the FIT sample was screened for collection date and presence of permission form and administrated in ICOLON IT database. ICOLON database was developed and owned by the regional organisations for Population Screening South-West and Mid-West Netherlands. After arrival at the laboratory OC-Sensor FIT sample was stored at -20°C, FOB-Gold FIT sample was stored at 4°C (median 3 days, range 1-6 days) until analysis. Both FIT samples were stored according to the manufacturer’s recommendations. FITs, which were inappropriately used or non-analysable, were marked in ICOLON, and participants received a new FIT. All other specimens received in the laboratory were analysed. If a specimen was received in the laboratory >7 days from date of sample collection and the test result was $< 10$ µg Hb/g faeces, the participant received a new
FIT and the test result was discarded. If a specimen received in the laboratory >7 days from date of sample collection and test result was ≥ 10 µg Hb/g faeces, the participant received a positive test result and a reference for colonoscopy. The number of and reasons for non-analysable FITs were recorded by the laboratory analysts.

The OC-Sensor FITs were analysed on two OC-Sensor µ systems (Eiken, Japan), and the FOB-Gold FITs were analysed on a Sentinel Sentifit 270 system (Sentinel, Italy). All FIT samples were allowed to warm to room temperature before analysis and analysed once. The analytical working range was 1-200 µg Hb/g faeces for the OC-Sensor µ, and 1-170 µg Hb/g faeces for the Sentifit 270. Samples with Hb level above the upper analytical working limits were not diluted or reanalysed. Before the start of the study the OC-Sensor µ system and the Sentifit 270 system were compared. Fifty- five faecal samples were spiked with different concentrations of Hb, and from each spiked sample two OC-Sensor FITs and two FOB-Gold FITs were taken and analysed on OC-Sensor µ or Sentifit 270, respectively. Using paired Student’s t-test no significant difference was found by levels ≤ 65 µg Hb/g faeces (p=0.412). The OC-Sensor µ was calibrated with six calibrators. In the Erasmus MC this was done every week and in advance of every analytical run two quality controls (low and high) were measured. In the Academic Medical Centre Amsterdam (AMC) the calibration took place when necessary, indicated by the results of the controls, and was monitored by two controls at the start and at the end of each run; calibrators and controls were from Eiken, Japan. The Sentifit 270 was calibrated with six calibrators every month or earlier when another latex lot was used for analysis. Three quality controls (low, middle and high) were run every analytical run before and after analysis of the samples. Calibrators and controls were from Sentinel, Italy. Analyses were carried out in a 14-month period by seven technicians and two staff members, with over 8 years’ expertise in FIT analysis. Samples that appeared to be over range were not diluted and reanalysed.

Follow-up evaluation

Participants with a positive FIT result were scheduled for colonoscopy within 4 weeks. In case of an incomplete colonoscopy, a CT colonography was performed. Experienced endoscopists, all board-certified gastroenterologists who had performed at least 1,000 colonoscopies, performed all colonoscopies for the current trial. The maximum reach of the endoscope, quality of bowel preparation, data on location, size, macroscopic aspect, morphology, and endoscopic assessment of completeness of resection was recorded for all lesions detected during colonoscopy. All lesions were collected and evaluated by experienced GI pathologists according to the Vienna criteria and WHO classification [21,22]. Advanced adenomas (AA) were defined as an adenoma with a diameter ≥10 mm, and/or with a ≥25% villous component, and/or high-grade dysplasia. Advanced neoplasia (AN) included AA and CRC. Cancers were staged according to the 7th edition of the American Joint Committee on Cancer classification [23]. Advice regarding surveillance colonoscopy after removal of adenomatous polyps, large (≥10 mm) serrated lesions or cancer was given to the clients according to the Dutch guideline. Participants with a negative colonoscopy were referred back to the screening programme, but were considered not to require FIT screening for 10 years.
Statistical analysis
The analysis was based on the intention-to-screen principle. The primary outcome measure was the diagnostic yield for AN, defined as the proportion of participants being diagnosed with AN relative to the total number of invitees. When more than one lesion was present, the screenee was classified according to the most advanced lesion.

Additional outcome measures were the participation rate, usability, the PR, the positive predictive value (PPV) for CRC and AN, the diagnostic yield of CRC defined as the proportion of participants with CRC relative to the number of invitees, and detection rate for AN and CRC, defined as the proportion of subjects with AN/CRC relative to all participants. The participation rate was calculated as the number of participants relative to all eligible invitees. The PR was defined as the proportion of participants with a positive test result. Usability was defined as the number of non-analysable tests. The PPV refers to the participants in whom AN is detected relative to those undergoing colonoscopy after a positive FIT or, in case the colonoscopy was incomplete, CT colonography. As this is a dynamic cohort, outcomes were also separately analysed for first-time and repeat-round invitees. Adenoma detection rate was defined as the proportion of colonoscopies in which one or more adenomas were found.

Differences in proportions between groups were analysed for statistical significance using the $\chi^2$ test statistic. Differences in means between groups were tested using the Student’s t statistic. Participation rate, PR, detection rate (DR), and PPV were calculated and described as proportions with 95% CI. All p values were two-sided and considered significant if <0.05. Analyses were conducted using SPSS for Windows version 21.0.

Sample size calculation
Sample size was guided by the size of the dynamic cohort, invited for the three previous screening rounds. This cohort consists of approximately 20,000 people. We anticipated that differences in participation would be crucial in driving any differences in diagnostic yield. Diagnostic yield would also be affected by failures, positivity, and PPV. Inviting 20,000 people would then have a power of at least 80% to detect an absolute difference in diagnostic yield of 5 per 1,000 invitees or more, assuming 60% participation with FIT-based screening, a 6% PR, and a 30% PPV, using two-sided testing at a 5% significance level.

Ethical approval
The study was approved by the Dutch National Health Council (Population Screening Act; publication no. 2013/20).

RESULTS
We invited 19,618 persons, of whom 327 had to be excluded because they met one of the exclusion criteria (n=306), had moved (n=19), or died (n=2), leaving 19,291 eligible invitees (Figure 1). Of the 9,669 invitees who received the OC-Sensor, 4,706 (49%) were male. Of the 9,622 invitees who received the FOB-Gold, 4,584 (48%) were male. The median age in both study arms was 61 years
(IQR 57-67). The proportion of first-time invitees was 14.2% in the OC-Sensor group and 14.8% in the FOB-Gold group (p=0.23).

In the OC-Sensor group, 6,040 returned the FIT (63%), versus 6,014 in the FOB-Gold group (63%) (p=0.96). Participation was lower among first-time invitees, yet over all rounds participation was similar for both FIT brands (57% for OC-Sensor and 56% for FOB-Gold; p=0.73). Non-analysable tests were reported in 41 (0.7%) participants in the OC-Sensor arm versus 118 (2.0%) in the FOB-Gold

Figure 1. Study design
CHAPTER 3

arm (p<0.001). The main reason for an unanalysable test was a too large sample of faeces collected in the tube by the participant (Table1).

More participants tested positive at the pre-specified cut-off of 10 µg Hb/g faeces with the OC-Sensor: 479 (7.9%) versus 390 (6.5%) for FOB-Gold (p=0.002). The difference in PR disappeared at higher cut-offs (Figure 2).

PR among first time invitees were higher and similar for both FITs, respectively 9.5% for OC-Sensor and 9.1% for FOB-Gold (p=0.86). Faecal Hb concentrations were distributed differently across both tests, with the OC-Sensor measuring higher values of Hb concentrations, ranging up to 1333 µg Hb/g faeces, versus 179 µg Hb/g faeces for FOB-Gold (p <0.0001, Figure 3).

Adherence to colonoscopy among FIT-positive screeners was 90% for both tests. Overall adenoma detection rate was 58%. AN was detected in 137 participants in the OC-Sensor group (1.4%) and in 114 in the FOB-Gold group (1.2%; p=0.15). In 537 FIT-positives (68%), no adenomas or only non-AA were detected. AN were detected in 224 (28.4%) participants, and CRC was diagnosed in 27 (3.4%) participants, with most CRCs detected at an early stage (56% stage I; Table 2). For the two FITs, no significant differences were observed in colonoscopy findings, including non-AA, AN and CRC, neither between stages of CRC.

The diagnostic yield for CRC was 0.1% in both groups (p=0.84). The detection rate of AN among participants was 2.3% for OC-Sensor and 1.9% for FOB-Gold (p=0.15; Table 3). The detection rate for

Table 1. Reasons for failure to analyse FIT.

<table>
<thead>
<tr>
<th>Reasons</th>
<th>OC-sensor (%) (n=41)</th>
<th>FOB-Gold (%) (n=118)</th>
</tr>
</thead>
<tbody>
<tr>
<td>too large sample</td>
<td>40 (98)</td>
<td>97 (82)</td>
</tr>
<tr>
<td>too little sample</td>
<td>1 (2)</td>
<td>5 (4)</td>
</tr>
<tr>
<td>loss of buffer</td>
<td>0</td>
<td>7 (6)</td>
</tr>
<tr>
<td>analytical technical failure</td>
<td>0</td>
<td>9 (8)</td>
</tr>
</tbody>
</table>

Figure 2. Positivity rates (PR) for the OC-sensor and FOB-Gold at different cut-offs. * dashed lines represent positivity rates of Hb concentrations under the pre-specified cut-off (10 µg Hg/g feces).
CRC was 0.2% for both tests (p=0.84). The PPV for AN was 31% for OC-Sensor and 32% for FOB-Gold at the prespecified cut-off (p=0.80). For CRC the PPV was 3.0% for OC-Sensor versus 4.0% for FOB-Gold (p=0.45).

In addition, as this study concerns a fourth round of screening we have stratified the results per number of participations to (previous) screening rounds for all participants of this fourth round (Table 4). Both the PPV as well as the detection rate for AN were highest in those who participated for the first time. Because equal cut-offs did not result in equal PRs, we next calculated the PPV for different PRs using multiple cut-offs, and this is illustrated in Figure 4. This yielded a similar PPV for both tests when comparing tests at the same PRs, and resulted in similar partial areas under the curve (p=0.48). This figure illustrates that both tests performed equally in terms of diagnostic yield for AN when comparing them at the same PR and thus also at an equal number of colonoscopies required.

Table 2. Colorectal cancer stage for the OC-sensor and FOB-Gold among FIT-positive screenees.

<table>
<thead>
<tr>
<th></th>
<th>OC-sensor (%)</th>
<th>FOB-Gold (%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n=13)</td>
<td>(n=14)</td>
<td></td>
</tr>
<tr>
<td>stage I</td>
<td>8 (61.5)</td>
<td>7 (50.0)</td>
<td></td>
</tr>
<tr>
<td>stage II</td>
<td>1 (7.7)</td>
<td>1 (7.1)</td>
<td></td>
</tr>
<tr>
<td>stage III</td>
<td>3 (23.1)</td>
<td>4 (28.7)</td>
<td>0.63*</td>
</tr>
<tr>
<td>stage IV</td>
<td>1 (7.7)</td>
<td>2 (14.3)</td>
<td></td>
</tr>
</tbody>
</table>

* overall p-value
Table 3. Test characteristics for the OC-sensor and FOB-Gold for all invitees (n=19,291) and for first time invitees (n=2,796).

<table>
<thead>
<tr>
<th></th>
<th>All invitees</th>
<th>First-time invitees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OC-sensor (n=9,669)</td>
<td>FOB-Gold (n=9,622)</td>
</tr>
<tr>
<td>Positive predictive value (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced neoplasia</td>
<td>31.3</td>
<td>32.0</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>3.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Diagnostic yield (invitees, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced neoplasia</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Detection rate (participants, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced neoplasia</td>
<td>2.3</td>
<td>1.9</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Figure 4. Positive predictive value and positivity rate

DISCUSSION

In CRC screening programmes, willingness to undergo a screening modality, easy use of the test and diagnostic accuracy are vital. In this cohort of biennial population-based FIT screening, we observed similar performance of the OC-Sensor and FOB-Gold. Both tests had the same participation rate of 63%. A higher PR was found for OC-Sensor, resulting in more colonoscopies and a slightly higher detection rate of AN. A significantly higher proportion of non-analysable tests was found with FOB-Gold. Both tests performed similar in terms of PPV. However, screening with the OC-Sensor test led to a higher diagnostic yield, mainly due to the previously mentioned higher PR. When comparing both tests at the same PRs by raising the cut-off of the OC-Sensor, they showed similar PPVs.

Our study has several strengths. All invitees were randomly selected from the general population and were at average risk of developing CRC. Invitees from the same household received
Table 4. Positive predictive value and detection rate in participants according to number of participations over all rounds (n=12,054).

<table>
<thead>
<tr>
<th></th>
<th>1st time participants (n=2,582)</th>
<th>2nd time participants (n=2,459)</th>
<th>3rd time participants (n=1,720)</th>
<th>4th time participants (n=5,293)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OC-sensor</td>
<td>FOB-Gold</td>
<td>OC-sensor</td>
<td>FOB-Gold</td>
</tr>
<tr>
<td>Positive predictive value (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced neoplasia</td>
<td>45.1</td>
<td>43.0</td>
<td>30.7</td>
<td>27.3</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>5.7</td>
<td>4.7</td>
<td>0</td>
<td>7.6</td>
</tr>
<tr>
<td>Detection rate (participants, %)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced neoplasia</td>
<td>4.3</td>
<td>3.5</td>
<td>1.9</td>
<td>1.5</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>0.5</td>
<td>0.4</td>
<td>0</td>
<td>0.4</td>
</tr>
</tbody>
</table>
the same test brand, so that confusion about use or shape of the test was avoided. In addition, risk of exchanging the two FITs between participants living in the same household was thereby prevented. Adherence to colonoscopy after a positive FIT was high with 90% of screenees undergoing colonoscopy. A possible limitation for the results of our study is that it was performed in a fourth round of FIT screening, with the consequence that the majority of the population was not screening-naïve. However, as the cohort included previous participants as well as newly invited individuals, it represents a true population-based screening population. Second, it is possible that participants were more familiar with the OC-Sensor, which had been used in the first three rounds, and this may have influenced the usability of the FOB-Gold. However, a higher rate of non-analysable tests was also found among first-time participants (1.0% for the OC-Sensor vs 2.9% for the FOB-Gold, p value <0.001). Finally, a more precise comparison of the diagnostic test accuracy of the two FITs could have been made by applying both tests on the same faecal samples, with all participants undergoing subsequent colonoscopy. This would have allowed to evaluate prime indicators of diagnostic performance, including sensitivity, specificity and the area under the receiver-operating curve. However, such a design might influence the willingness to participate and would not have allowed for a fair comparison of participation rates and usability.

In our study, the main reason for non-analysable tests was a too large sample of faeces collected in the tube by the participant. These findings are in line with previously published results [24,25]. A possible explanation for this sampling error could be the round shape of the opening of the FOB-Gold, making it possible to sample larger volumes of faeces, whereas OC-Sensor has a small oval opening (Figure 5). It should be noted that the FOB-Gold test-tube that was used in our study has been adjusted before the start of the study by the manufacturer to prevent participants from unscrewing the wrong side of the test causing loss of buffer [26].

Concerns have been raised about the so-called prozone effect in FIT screening [27,28]. This effect could lead to an underestimation of high Hb concentrations because the relative large amount of antigen (in this case Hb) is greater than the quantity of antibody present in the test. This could lead to underestimation of true Hb concentration at very high faecal Hb concentrations. Our results indicate that OC-Sensor is less subject to this effect than the FOB-Gold, with a large difference in distribution of the highest concentrations of Hb between both tests. The ability to measure high values is often driven by properties of the analytical equipment. However, it is important to realise that such high concentrations are usually much higher than most cut-offs, and are therefore not likely to influence PRs.

In our cohort OC-Sensor had a higher PR than FOB-Gold at a cut-off of 10 µg Hb/g faeces. This is in contrast with results from a Spanish study, in which the FOB-Gold had a higher PR than the OC-Sensor [24]. This difference can be explained by multiple factors. First, in the Spanish study a higher positivity cut-off (100 ng Hb/mL) was used; we found that for higher cut-offs the difference in PR between the tests was less pronounced. Second, the Spanish study used the same cut-off (expressed in ng Hb/mL buffer) for both tests. However after standardising this to µg Hb/g faeces, this cut-off relates for the OC-Sensor to a higher cut-off than for the FOB- Gold, 20 µg Hb/g and 17 µg Hb/g faeces, respectively [14,29]. Last, the Spanish study was conducted in 2009, after which date both manufacturers have improved their buffer to increase the conservation of Hb in the sample.
Besides participation, test accuracy is crucial for screening effectiveness. In our study, the PPV was comparable for both tests. In literature, the above-mentioned Spanish study is the only other study that compared the same two brands of FIT in a screening population, allowing comparison of participation rates in addition to test accuracy. This study was performed in a first round of screening and demonstrated superiority of the OC-Sensor in participation rates, with similar PPVs for both tests [24]. Other studies that have evaluated the OC-Sensor and/or FOB-Gold test relied on different designs to evaluate test performance. Such study designs however do not allow comparison of participation rates. One of those studies compared the two FITs to gFOBT, and concluded that apart from the superiority of FIT over gFOBT, no differences in performance were found between FITs [9]. These findings are in accordance with our results, confirming that similar test performance of both FITs can also be expected in a population-based screening setting. Another study sent both FITs to screenees, and in case of one or two positive test results participants were referred for colonoscopy. In this setting a relative sensitivity could be calculated, showing a lower sensitivity for the FOB-Gold, and comparable specificity [30]. As all participants in our study only received one out of two tests, sensitivity could not be calculated. However, as both groups were randomised, prevalence of AN should be similar between groups and as a result the detection rate is a direct reflection of the relative sensitivity. Also, future follow-up will allow determining the incidence of interval carcinomas to determine programme sensitivity per test.

As it has become clear that different FITs use a variety of sampling techniques and report Hb concentration in different units, more attention has been given to standardising faecal haemoglobin concentrations [14]. Due to these differences a comparison between FITs could prove to be arduous. It has been proposed to standardise the measuring units to µg Hb/g faeces, taking into account the amount of buffer and sampling volume [14]. However, a Taiwanese study concluded that even after standardisation, different brands of quantitative FITs perform differently [13].

Figure 5. The FOB-Gold (left) and OC-Sensor (right).
findings are in line with these results, showing different PRs and PPVs per test at the same cut-off. This is likely explained by the fact that standardising FITs to µg Hb/g faeces is hampered by several factors. First, wet sample FITs do not directly determine Hb concentration in faeces, but determine Hb concentration in the kit’s storage buffer. This depends on both the faecal Hb concentration and the amount of faecal material put into the buffer. Although manufacturers assume that the volume of faecal material sampled is stable per device, sampling volumes can in practice vary substantially. This affects the reported faecal Hb concentrations. Second, different FIT brands make use of antibodies against different epitopes. This could potentially influence test performance and PR. As a result, the same cut-off in µg Hb/g faeces can lead to different PRs depending on the FIT brand. Therefore we chose to compare the PR between both FITs, and this resulted in comparable PPVs. Another advantage of this approach is that the PR directly reflects the required colonoscopy capacity. Evidently, a higher PR leads to more colonoscopies and consequently results into a higher detection rate of AN in case of similar PPVs. As colonoscopy capacity is for most countries the main determinant in a CRC screening programme, information on required capacity using test PR is crucial when implementing a population-based screening programme or when contemplating on changing to a different FIT brand within an already existing program [12]. This necessitates a comparative analysis of test performances not solely based on similar cut-offs but first based on similar PRs. Moreover, if programmes intent to use different cut-off concentrations for different populations (eg, based on age, gender or screening history) PRs for the different cut-offs per subgroup should be addressed. Our comparison shows that both tests perform equally regarding detection rate and PPV at cut-offs that result in equal PRs requiring the same number of colonoscopies. Notably, miss rates of early-stage cancers between tests is another important outcome, and the fact that both tests led to very similar numbers of cancers detected strongly suggests that they will also perform similar in terms of interval cancers.

This trial shows that the OC-Sensor and FOB-Gold can be expected to perform similar in population-based screening, with no major differences in diagnostic yield. Despite standardisation on Hb concentration, differences in PR and diagnostic yield can be expected, but adjusting for PR will result in an equal number of colonoscopies, and a similar diagnostic yield. When comparing different FIT brands, our results indicate a need for standardising on PR, rather than on faecal Hb concentration.

ACKNOWLEDGEMENTS

This work is part of a research programme that is (partly) financed by the Netherlands Organisation for Health Research and Development (ZonMw 120720012). The authors would like acknowledge with gratitude the essential contribution of all involved endoscopists, A.N. Reijm, A.C. de Groot, J. Schiewold, K. Izelaar and H. ’t Mannetje to this study.
REFERENCES


