Accepted Manuscript

Direct Comparison of Diagnostic Performance of 9 Quantitative Fecal Immunochemical Tests for Colorectal Cancer Screening

Anton Gies, PhD student, Katarina Cuk, PhD, Petra Schrotz-King, PhD, Hermann Brenner, MD, MPH

PII: S0016-5085(17)36177-2
DOI: 10.1053/j.gastro.2017.09.018
Reference: YGAST 61440

To appear in: *Gastroenterology*
Accepted Date: 17 September 2017

Please cite this article as: Gies A, Cuk K, Schrotz-King P, Brenner H, Direct Comparison of Diagnostic Performance of 9 Quantitative Fecal Immunochemical Tests for Colorectal Cancer Screening, *Gastroenterology* (2017), doi: 10.1053/j.gastro.2017.09.018.

Direct Comparison of Diagnostic Performance of 9 Quantitative Fecal Immunochemical Tests for Colorectal Cancer Screening

Short Title: Stool tests for colorectal cancer screening.

Anton Gies, PhD student;[1] **Katarina Cuk**, PhD;[2] **Petra Schrotz-King**, PhD;[1] and **Hermann Brenner**, MD, MPH [1, 2, 3]

Author affiliations:
[1]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany
[2]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany
[3]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

Grant support: There was no external funding for this study.

Abbreviations: AA, advanced adenoma; AN, advanced neoplasm; AUC, area under the curve; CI, confidence interval; CRC, colorectal cancer; FIT, fecal immunochemical test; FSD, fecal sampling device; Hb, hemoglobin; ROC, receiver operating characteristic.

Correspondence: Hermann Brenner
Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, D-69120 Heidelberg, Germany
Phone ++49-6221-421300
Fax ++49-6221-421302
E-mail h.brenner@dkfz.de

Disclosures: The authors disclose no conflicts of interest. All test kits were provided free of charge by the manufacturers. The manufacturers had no role in the study design, in the collection, analysis, and interpretation of data, or approval of submission for presentation/publication.

Writing Assistance: No assistance in writing was provided.

Author Contributions: AG contributed to the fieldwork and laboratory analyses, carried out the statistical analyses and drafted the manuscript. KC led and contributed to the laboratory analyses and contributed to revision of the manuscript. PSK contributed to revision of the manuscript. HB designed, led and supervised the study and contributed to revision of the manuscript. All authors approved the final, submitted version of the manuscript.

Acknowledgments: The authors acknowledge Sabine Eichenherr, Romana Kimmel and Ulrike Schlesselmann for their excellent work in laboratory preparation of stool samples. They also acknowledge Volker Herrmann for his help in preparing the study.

ABSTRACT

Background & Aims: A variety of fecal immunochemical tests (FITs) for hemoglobin (Hb) are used in colorectal cancer (CRC) screening. It is unclear to what extent differences in reported sensitivities and specificities reflect true heterogeneity in test performance or differences in study populations or varying pre-analytical conditions. We directly compared the sensitivity and specificity values with which 9 quantitative (laboratory-based and point of care) FITs detected advanced neoplasms (AN) in a single CRC screening study.

Methods: Pre-colonoscopy stool samples were obtained from participants of screening colonoscopy in Germany from 2005 through 2010 and frozen at –80°C until analysis. The stool samples were thawed, homogenized, and used for 9 different quantitative FITs in parallel. Colonoscopy and histology reports were collected from all participants and evaluated by 2 independent, trained research assistants who were blinded to the test results. Comparative evaluations of diagnostic performance for AN were made at preset manufacturers´ thresholds (range: 2.0–17.0 μg Hb/g feces), at a uniform threshold (15 μg Hb/g feces), and at adjusted thresholds yielding defined levels of specificity (99%, 97%, and 93%).

Results: Of the 1667 participants who fulfilled the inclusion criteria, all cases with AN (n=216) and 300 randomly selected individuals without AN were included in the analysis. Sensitivities and specificities for AN varied widely when we used the preset thresholds (21.8%–46.3% and 85.7%–97.7%, respectively) or the uniform threshold (16.2%–34.3% and 94.0%–98.0%, respectively). Adjusting thresholds to yield a specificity of 99%, 97%, or 93% resulted in almost equal sensitivities for detection of AN (14.4%–18.5%, 21.3%–23.6%, and 30.1%–35.2%, respectively) and almost equal positivity rates (2.8%–3.4%, 5.8%–6.1% and 10.1%–10.9%, respectively).

Conclusions: Apparent heterogeneity in diagnostic performance of quantitative FITs can be overcome to a large extent by adjusting thresholds to yield defined levels of specificity or positivity rates. Rather than simply using thresholds recommended by the manufacturer, screening programs should choose thresholds based on intended levels of specificity and manageable positivity rates.

KEY WORDS: fecal occult blood test; colon cancer; advanced adenoma; early detection

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer globally, accounting for approximately 1.4 million new cases and 700,000 deaths per year.[1] Randomized controlled trials have shown that annual or biannual screening with traditional, guaiac-based fecal occult blood tests could reduce CRC mortality by up to 30%.[2-4] Even stronger mortality reduction should be possible with newer fecal immunochemical tests (FITs) for hemoglobin (Hb), which have been shown to have substantially higher sensitivity, not only to detect CRC, but also its most important precursor, advanced adenoma (AA).[5-7] Therefore, FITs are meanwhile widely recommended as primary CRC screening tests[8, 9] and used as such in an increasing number of countries.[10] With the growing market for FIT-based screening, a large number of FITs from diverse manufacturers are meanwhile being offered. Although diagnostic performance of specific FIT brands has been evaluated in previous studies,[11, 12] the heterogeneity of study designs, study populations, pre-analytical sample handling, and positivity thresholds makes a comparative evaluation of the diagnostic performance of different FIT brands difficult if not impossible. Only very few studies have evaluated more than one FIT based on the same stool samples. In a previous study, we evaluated diagnostic performance of six different qualitative point of care FITs in a cohort of participants of screening colonoscopy in Germany.[13] A large diversity of sensitivities and specificities was observed, with the most sensitive FITs showing the lowest specificity and vice versa. This diversity probably mostly reflects problems with the fixed thresholds in qualitative FITs. More flexible analyses are possible with quantitative FITs which allow flexible adjustment of thresholds based on quantitative measurements of fecal Hb concentrations.

The aim of this study was to evaluate and directly compare diagnostic performance of nine different quantitative commercially available and clinically used FITs, including both laboratory-based FITs as well as point of care FITs, based on the same stool samples collected from our large cohort of participants of screening colonoscopy.

MATERIALS AND METHODS

This article is following the STARD (Standards for Reporting of Diagnostic Accuracy) statement[14] and the FITTER (Fecal Immunochemical TesTs for Hemoglobin Evaluation Reporting) checklist.[15]

Study design and study population

This project is based on the BliTz (Begleitende Evaluierung Innovativer Testverfahren zur Darmkrebsfrüherkennung) study, an ongoing prospective study among participants of screening colonoscopy. The BliTz study is conducted in cooperation with 20 gastroenterology practices in Southern Germany, with the aim to collect blood and stool samples for the evaluation of novel CRC screening tests. Participants of the German screening colonoscopy program are informed and recruited at a preparatory visit in the practice, typically one week before colonoscopy. Because of the low number of CRC cases in a true screening setting, an additional separate group of CRC cases was included for ancillary analyses who were recruited in the DACHSplus satellite sub-study of the DACHS (DArmkrebs: CHancen der Verhütung durch Screening) study, a case-control study with a focus on the role of colonoscopy in CRC prevention. In the DACHSplus sub-study cancer patients were referred by general practitioners or gastroenterologists for surgery to one of four collaborating hospitals, where the patients were informed about the study and recruited prior to initiation of any therapy.

Further information on both, BliTz and DACHSplus has been provided elsewhere.[7, 13, 16, 17] Both studies were approved by the Ethics committee of the University of Heidelberg and by the State Chambers of Phyiscians of Baden-Wuerttemberg, Rhineland-Palatinate and Hesse.

Between 2005 and 2010, participants from BliTz and DACHSplus received a study kit which included a stool collection container (60ml). These individuals were considered for this project.

Figure 1 shows the exclusion criteria and flow diagrams of the study participants. Briefly, 566 samples were analyzed in total. From the main study, conducted in the screening setting (BliTz study), all eligible advanced neoplasm (AN)-cases (n=216) were included (i.e., cases with CRC or AA, defined as adenoma with at least one of the following features: ≥1cm in size, tubulovillous or villous components, or high-grade dysplasia). The 300 participants without AN (including participants with non-advanced adenomas, hyperplastic polyps and no neoplasms) were randomly selected from all eligible participants (n=1437) in this group. Due to the low number of CRC cases which is typical of true screening settings, 50 CRC cases from the DACHSplus study (clinical setting) were additionally included for ancillary analyses.

Sample and data collection

After giving written informed consent, participants were asked to collect one stool sample from a single bowel movement, without any specific recommendations for dietary or medicinal restrictions, before bowel preparation for colonoscopy (screening setting) or surgery (clinical setting). Participants were furthermore asked to keep the stool-filled container in a freezer or, if not possible, in a refrigerator at home until their colonoscopy appointment (screening setting) or hospital admission (clinical setting). Upon receipt the stool-filled containers were immediately frozen at –20℃ in the practice (screening setting) or in the hospital (clinical setting), then shipped on dry ice to a central laboratory and finally stored at –80℃ at the German Cancer Research Cente r (DKFZ) study center.

In addition, participants were asked to fill out a questionnaire focusing on CRC risk factors. Colonoscopy and histology reports were collected from all participants of the screening colonoscopy. Colonoscopists were blinded for test results. After surgery, medical reports on the clinical patients were collected from the hospital. Relevant information was extracted by two independent trained research assistants who were blinded to the test results.

Specimen collection and handling

For the purpose of this evaluation, which was conducted in fall 2016, the stool samples were thawed overnight in a refrigerator at the study center and homogenized with a sterile plastic stick. A defined stool amount was extracted in a randomized order using each company's brand-specific fecal sampling device (FSD). Each FSD was a small vial, containing a defined volume of Hb-stabilizing buffer, with a lid that was attached to a serrated plastic stick for stool collection. After stabbing the collection stick into three different parts of the stool sample, we checked if all serrations on the stick were filled completely. Then we inserted the stick with the collected stool back into the vial. The vials have a tight membrane at their entrance which removes most of the stool, leaving only a specified quantitative amount of stool in the serrations even though this may not be consistently successful in practice. The only exception was the ImmoCare-C vial, where a supplied custom-fitted scraper was used to remove excess stool material from the collection stick. All FSDs were subsequently mixed on a vortexer, so that the stool could move out of the serrations into the buffer. Stool-filled FSDs were stored over night at a median temperature of 21.5℃ (range: 20.0℃–24.0℃).

On the following day, all laboratory analyses were conducted in parallel by laboratory-experienced staff, which was blinded to the colonoscopy results. Test calibrators and test controls were performed on a regular basis according to the manufacturers´ instructions. Due to limited laboratory space and resources, five quantitative FITs had to be evaluated externally. After vortexing, the stool-filled FSDs were immediately packed and directly shipped, without any cooling, to the cooperating companies providing the respective tests (CARE diagnostica [CAREprime Hb and ImmoCARE-C], Immundiagnostik [Hb ELISA and QuantOn Hem] and R-Biopharm [RIDASCREEN Hb]) for evaluation. The mean outdoor temperature in the study center area, extracted from the German Meteorological Service, was 7.7℃ (range: –6.7℃–18.7℃). [18] Detailed information about all nine quantitative FITs is shown in Table 1. Our analysis included five laboratory-based FITs and four point of care FITs; one of the latter (QuantOn Hem) would not even require a local analytical instrument, but could be run with remote testing using a smartphone with an App for optical analysis of the test cassette.

Statistical analyses

Sensitivities were calculated for CRC, AA, and their combination, AN, with their corresponding 95% confidence interval (CI) at preset manufacturers´ thresholds and at adjusted thresholds, using colonoscopy results as the reference standard. Specificities were calculated for the absence of any AN. The Clopper-Pearson method was used to calculate 95% CIs. The expected positivity rate of the tests in a true screening setting was calculated by applying the observed sensitivity and specificity to all eligible participants with and without ANs from the screening setting (n=230 and n=1437, respectively), using the following formula:

$$\text{Expected positivity rate} = (\text{Sensitivity}*230 + (1-\text{Specificity})*1437)/(230+1437).$$

In order to evaluate the diagnostic performance of the tests across different thresholds, receiver operating characteristic (ROC) curves were constructed and the areas under the curves (AUCs) were determined. Because of the low number of CRC cases in the screening setting and the similarity of sensitivities for CRC cases in the screening and the clinical setting, the ROC plot for CRC was constructed combining both groups of cases. The ROC analysis for AN on the other hand is purely based on the screening setting.

In addition to analyses for both sexes combined, sex specific analyses were performed. For CRC cases, stage specific sensitivities were evaluated in addition to overall sensitivities. Stages were categorized according to the Union for International Cancer Control (UICC) classification (7th edition), and sensitivity was determined for early (0/I/II) stages versus late (III/IV) stages. Due to the small number of CRC cases in the screening setting and very similar sensitivity results for CRC patients recruited in the screening setting and the clinical setting, stage specific analyses were performed for both groups combined.

All statistical analyses were conducted using SAS Enterprise Guide, version 6.1 (SAS Institute, Cary, North Carolina, USA).

RESULTS

Study population

A total of 2042 participants of the BliTz study, who were recruited between 2005 and 2010, provided stool samples in 60ml collection containers (screening setting) (**Figure 1, A**). 375 participants were excluded, because they were not between 50 and 79 years old (n=52), had inflammatory bowel disease (n=10), had a personal history of CRC, adenoma or polyps (n=39), had a previous colonoscopy in the last 5 years (n=114), provided their stool sample not before colonoscopy (n=75), had an incomplete colonoscopy (n=8) or performed an inadequate bowel preparation (n=77). Out of the 1667 eligible individuals 230 were AN-cases, of whom 14 had to be excluded due to an insufficient stool amount, leaving 216 AN-cases (16 cases with CRC, 200 cases with AA) for the analyses. With 300 randomly selected individuals without AN, a total of 516 participants from the screening setting were included in the main study.

From the DACHSplus study (clinical setting) (**Figure 1, B**), a total of 184 CRC cases provided stool-filled containers. After exclusion of participants with neoadjuvant therapy before stool sampling (n=51), age <50 or ≥80 years (n=30), inflammatory bowel disease (n=3) or personal history of CRC (n=6), 94 individuals with CRC were eligible for this study. All patients diagnosed through a screening colonoscopy and supplying a sufficient stool amount (n=27) were included. From the 65 patients whose CRC was detected otherwise (i.e. not by screening colonoscopy) 23 individuals were randomly selected for this study. Finally, 50 clinical CRC cases were included for the ancillary analyses.

An overview on basic characteristics of the study participants is provided in **Table 2**. A slight majority of participants in both the main study (screening setting) and the ancillary study (clinical setting) were males, mean ages were 63.2 and 65.8 years, respectively. The majority of CRC cases from both the screening setting (9/16) and the clinical setting (30/50) were diagnosed at an early stage (0/I/II).

Comparison of test characteristics

Table 3 and **Table 4** display sensitivities and specificities of the nine quantitative FITs at preset manufacturers´ thresholds, at a uniform threshold and at adjusted thresholds. The results of both tables are sorted by the sensitivities for AN.

At preset thresholds the sensitivities (95% CI) for AN ranged from 21.8% (16%–28%) to 46.3% (40%–53%), with corresponding specificities (95% CI) between 97.7% (95%–99%) and 85.7% (81%–89%) (upper part of **Table 3**). This apparent strong variation in sensitivity and specificity seemed to be determined to a large extent by the variation of preset thresholds, with sensitivity strongly decreasing and specificity increasing with increasing thresholds. The sensitivities for AN were mostly determined by the sensitivities for AA, which make up the vast majority of AN in screening settings. Sensitivities for AA ranged from 18.0% to 43.5%, whereas much higher sensitivities, ranging from 62.5% to 81.3%, were observed for CRC. Using the thresholds preset by the manufacturers also yielded strongly varying expected positivity rates, ranging from 5.7% to 18.7%, when applying the tests in a true screening setting.

By adjusting the thresholds to yield the same levels of specificity for all tests the heterogeneity in the sensitivities and the expected positivity rates were substantially reduced or disappeared almost entirely (**Table 4**). With thresholds yielding a specificity of 99.0%, sensitivities (95% CI) for AN ranged from 14.4% (10%–20%) to 18.5% (14%–24%), and expected positivity rates ranged from 2.8% to 3.4%. For one test (RIDASCREEN Hb), the threshold could not be increased above the upper analytical range (50 μg Hb/g feces) to yield a specificity of 99%. With thresholds yielding a specificity of 96.7%, sensitivities (95% CI) for AN ranged from 21.3% (16%–27%) to 23.6% (18%–30%), and expected positivity rates ranged from 5.8% to 6.1%. With thresholds yielding a specificity of 93.0%, the sensitivities (95% CI) for AN ranged from 30.1% (24%–36%) to 35.2% (29%–42%), and the expected positivity rates ranged from 10.1% to 10.9%. For one test (QuikRead go iFOBT), the threshold could not be lowered to yield a specificity of 93%, because of the limited analytical

range (lower limit 15 μg Hb/g feces). For one other test (SENTiFIT-FOB Gold) the specificity of 93.3% was achieved at the lower end of its analytical range.

However, identical levels of specificities and very similar levels of sensitivities and expected positivity rates were achieved at apparently very different thresholds. Thresholds [μg Hb/g feces] that yielded specificities of 99.0%, 96.7% and 93.0% ranged from 18.20 to 53.38, from 6.11 to 29.54 and from 1.70 to 12.27, respectively. Vice versa, using a uniform threshold (here: 15 μg Hb/g feces, the lower end of the analytical range of one of the tests) resulted in strongly varying sensitivities (range of sensitivities for AN: 16.2% to 34.3%) and expected positivity rates (3.4% to 9.9%) (lower part of **Table 3**).

Overall, the sensitivities for CRC cases recruited in the screening setting and for CRC cases recruited in the clinical setting were very similar, but CIs were much narrower for the latter due to the substantially larger case number. When investigating the sensitivities according to CRC stage, sensitivities were higher for late (III/IV) stages versus early (0/I/II) stages for eight of the nine tests, with a median difference of 9 percent units (**Table 5**).

Figure 2 shows ROC curves and AUCs for the detection of CRC **(Figure 2, A**; cases from screening and clinical setting combined) and AN (**Figure 2, B**; screening setting only). The AUCs (95% CI) for CRC ranged from 79% (73%–85%) to 89% (84%–94%). For the detection of AN, the AUCs (95% CI) ranged from 59% (57%–62%) to 72% (68%–77%). Most of the apparent differences in ROC curves and AUCs resulted from the varying limits of the tests' analytical range, with ROC curves going either straight to the upper-right or to the lower-left corner for thresholds below or above the analytical range, respectively. Therefore no statistical tests for differences between the AUCs were performed. Segments of the ROC curves not affected by the limits of the analytical range were generally very close.

In sex specific analyses sensitivity was consistently higher and specificity was consistently lower among men than among women at the same thresholds, but ROC curves and AUCs were essentially identical.

DISCUSSION

To our knowledge, this is the first comprehensive comparative evaluation of diagnostic performance of a large number of quantitative FITs in a screening setting. Apparent large differences in diagnostic performance parameters were seen when using either preset thresholds recommended by the manufacturers or a uniform threshold. However, these apparent large differences almost entirely disappeared when thresholds were adjusted in such a way that all tests achieved defined levels of specificity (here: 99.0%, 96.7% and 93.0%), at which sensitivities were also all very close. Along the same lines, ROC curves and AUCs were all very similar except for some variation due to differences in the lower or upper end of the analytical range.

In a previous study from our group, similarly large apparent differences in diagnostic performance had been found for six different qualitative FITs.[13] Like in the present study, qualitative FITs with higher sensitivities had shown lower specificities and vice versa, pointing to differences in the threshold definition. However, due to the qualitative nature of the tests, no further exploration of the impact of shifting thresholds had been possible. Quantitative tests offer the advantage of flexible definition of thresholds. Such flexibility can be very useful to enable the best balance between sensitivity and specificity or to adapt positivity rates (which are close to 1 minus specificity in screening settings in which prevalence of AN is low) to colonoscopy capacities available for the screening population. Further advantages include the possibility of automated, objective measurements under quality controlled laboratory conditions.

Although a large number of studies have meanwhile evaluated the diagnostic performance of single quantitative FITs,[11, 12, 16, 19-22] and results even have been combined in meta-analyses,[11] only very few studies have evaluated more than one quantitative FIT in the same study population. It was therefore essentially unknown to what extent the reported partly very large differences in sensitivity and specificity might have resulted from true differences in diagnostic performance of the tests, or from differences in the populations studied or other

specific study characteristics, such as collection and pre-analytical handling of fecal samples. To our knowledge, only two studies from our group directly compared the diagnostic performance of two quantitative FITs (OC Sensor and RIDASCREEN Hb) among participants of screening colonoscopy, evaluating identical stool samples.[7, 17] Similar diagnostic performance of the two quantitative FITs was observed when the thresholds were adjusted to yield the same overall positivity rate (5%)[7] or the same specificity levels (90% and 95%).[17]

In a study from Taiwan, Chiang et al[23] compared the CRC detection rate and the positive predictive value of two different quantitative FITs (OC Sensor and HM-Jack) at the same threshold (20μg Hb/g feces). In agreement with our findings, Chiang et al[23] found major differences between tests despite identical thresholds. However, because colonoscopy was done in FIT positive participants only, direct estimates of sensitivity and specificity were not available. The same also applies to a randomized trial from the Netherlands, which found different positivity rates between OC Sensor and FOB Gold (two of the laboratory-based quantitative FITs included in our comparative analysis), despite the use of an identical threshold (10μg Hb/g feces).[24]

The design of our study essentially precluded any differences in study populations or sample handling as a cause of differences in observed diagnostic performance: All tests were evaluated in exactly the same study participants who were recruited in a true screening setting among participants of screening colonoscopy. Stool samples were collected in exactly the same manner, and additional homogenization of stool samples after thawing and before stool extraction for the single tests should further have eliminated the variation of Hb concentrations within a single bowel movement. Under these precautions, all nine tests included were shown to perform essentially equally well overall, with the remaining, apparent heterogeneity being almost exclusively threshold-related. However, in agreement with the findings from Chiang et al[23], our results illustrate that the threshold-related heterogeneity cannot simply be overcome by using the same threshold across different quantitative FIT brands. The most plausible reason for that seems to be variation in the “translation” of tests

results into Hb concentrations given by the manufacturers for the various tests. While the reasons for such variation cannot be disclosed by our study, our results underline the need of enhanced efforts for standardization and quality control. Interestingly, setting the thresholds to ensure defined levels of specificity (which is independent of such "translation") ensured levels of sensitivity to be quite similar as well.

In practice, determining test specificity or defining a threshold according to specificity in the context of an established FIT-based screening program is often difficult, as typically only FIT positive participants would undergo colonoscopy. However, choosing a threshold to ensure a defined positivity rate is straightforward. With AN as the major outcome, which typically has a prevalence of less than 10% in screening populations, the positivity rate is closely related to specificity (it is typically a few percentage points higher than one minus specificity). For example, thresholds yielding specificities of 99.0%, 96.7% and 93.0% in our study resulted in very narrow ranges of positivity rates from 2.8% to 3.4%, from 5.8% to 6.1% and from 10.1% to 10.9%, respectively. Vice versa, adjusting the threshold to defined levels of the positivity rate would have resulted in very narrow ranges of specificities across tests (data not shown). An additional advantage of choosing thresholds according to a defined positivity rate would be that the latter directly reflects the colonoscopy workload associated with the FIT-based screening program, which is a limiting factor in many countries.

Given that diagnostic performance of the various tests evaluated in our study was very similar after threshold adjustments, additional factors might determine advantages and disadvantages of the different tests. One obvious factor directly evident from our analyses is the width of the analytical range which delineates possibilities of threshold adjustment. Other factors to be considered which are beyond the scope of our study, might be, for example, costs of tests, convenience of sample collection, sample stability under routine environmental conditions, laboratory requirements, and ease of laboratory analysis. Interestingly, apart from the high lower end of the analytical range of one of the point of care tests, no consistent differences in diagnostic performance were seen between laboratory-based and point of care

tests, and equivalent diagnostic performance was even achieved with a smartphone-based test that could be conducted by the participants at their home without the need of any sample shipment, suggesting interesting perspectives for novel telemedicine applications. Nevertheless, the possibility should be kept in mind that diagnostic performance of point of care tests might be somewhat lower when these tests are applied in routine medical practice.

Specific strengths of our study include the first time parallel evaluation of a large number of quantitative FITs in a screening setting, with screening colonoscopy results as reference in all participants. However, our study also has a number of limitations that require careful discussion. First, stool samples were originally collected in small containers rather than FSDs provided by the manufacturers and stored frozen at –80°C over several years prior to analysis. This was though probably the only way to realize a comparative study like this, as it is difficult to imagine that study participants would be willing to collect nine fecal samples with nine different FSDs, each with different sample collection instructions. Nevertheless, the original FSDs provided by the manufacturers were used when extracting the fecal samples from the thawed stool, and prior homogenization of the thawed stool ruled out variation of Hb concentration within the same bowel movement as an additional source of variation of results between tests (even though this might lead to somewhat better diagnostic performance compared to routine practice where such homogenization is not performed). In a previous examination based on one of the tests included in the current study (SENTiFIT-FOB Gold), we furthermore found only small differences in comparative analyses of the diagnostic performance based on frozen fecal samples or fecal samples collected according to the manufacturer's instructions.[16] Similarly, two of the tests (OC Sensor and RIDASCREEN Hb) which had been evaluated in an overlapping selection of the same fecal samples (with one less freeze-thaw cycle, and without prior homogenization) several years earlier,[7, 17] showed very similar results in the overlapping segments of the study populations (data not shown). Second, despite the overall large size of the study, with targeted selection of samples (including those from all CRC cases) from more than 1600 participants of screening colonoscopy, the number of CRC cases from the screening setting was still rather low

(n=16), leading to broad CIs for the sensitivity estimates for CRC. More precise estimates were possible, however, by additionally considering CRC cases from our ancillary study from the clinical setting (of whom approximately half also had screen-detected CRC). Given the similarity of sensitivity estimates for CRC cases recruited in the screening setting and in the clinical setting for all of the nine tests evaluated, combining the analyses for both groups of CRC patients seems justified.

Despite its limitations, our study provides important information regarding the diagnostic performance and its comparability for a large number of quantitative FITs including quantitative FITs that are now widely used in screening practice, such as SENTiFIT-FOB Gold, the test used in the nationwide screening program in the Netherlands. With appropriate threshold adjustments all of the tests included in our evaluation seemed to perform almost equally well. Therefore additional criteria, such as costs, convenience of sample collection and analysis, or stability of results over prolonged sample storing or shipping times, to be evaluated in further, similarly highly standardized comparative investigations, as well as the analytical range may be relevant when selecting one or more quantitative FIT brands for specific screening programs. Furthermore, rather than simply using thresholds recommended by the manufacturer, screening programs should choose thresholds based on intended levels of specificity and manageable positivity rates.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359-86.
2. Hewitson P, Glasziou P, Watson E, et al. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. Am J Gastroenterol 2008;103:1541-9.
3. Scholefield JH, Moss SM, Mangham CM, et al. Nottingham trial of faecal occult blood testing for colorectal cancer: a 20-year follow-up. Gut 2012;61:1036-40.
4. Shaukat A, Mongin SJ, Geisser MS, et al. Long-term mortality after screening for colorectal cancer. N Engl J Med 2013;369:1106-14.
5. Zhu MM, Xu XT, Nie F, et al. Comparison of immunochemical and guaiac-based fecal occult blood test in screening and surveillance for advanced colorectal neoplasms: a meta-analysis. J Dig Dis 2010;11:148-60.
6. Park DI, Ryu S, Kim YH, et al. Comparison of guaiac-based and quantitative immunochemical fecal occult blood testing in a population at average risk undergoing colorectal cancer screening. Am J Gastroenterol 2010;105:2017-25.
7. Brenner H, Tao S. Superior diagnostic performance of faecal immunochemical tests for haemoglobin in a head-to-head comparison with guaiac based faecal occult blood test among 2235 participants of screening colonoscopy. Eur J Cancer 2013;49:3049-54.
8. Halloran SP, Launoy G, Zappa M, et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First Edition--Faecal occult blood testing. Endoscopy 2012;44 Suppl 3:SE65-87.
9. US Preventive Services Task Force. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA 2016;315:2564-2575.
10. Schreuders EH, Ruco A, Rabeneck L, et al. Colorectal cancer screening: a global overview of existing programmes. Gut 2015;64:1637-49.
11. Lee JK, Liles EG, Bent S, et al. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. Ann Intern Med 2014;160:171.
12. Robertson DJ, Lee JK, Boland CR, et al. Recommendations on Fecal Immunochemical Testing to Screen for Colorectal Neoplasia: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. Am J Gastroenterol 2017;112:37-53.
13. Hundt S, Haug U, Brenner H. Comparative evaluation of immunochemical fecal occult blood tests for colorectal adenoma detection. Ann Intern Med 2009;150:162-9.
14. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.
15. Fraser CG, Allison JE, Young GP, et al. Improving the reporting of evaluations of faecal immunochemical tests for haemoglobin: the FITTER standard and checklist. Eur J Cancer Prev 2015;24:24-6.
16. Chen H, Werner S, Brenner H. Fresh vs Frozen Samples and Ambient Temperature Have Little Effect on Detection of Colorectal Cancer or Adenomas by a Fecal Immunochemical Test in a Colorectal Cancer Screening Cohort in Germany. Clin Gastroenterol Hepatol 2016.
17. Tao S, Seiler CM, Ronellenfitsch U, et al. Comparative evaluation of nine faecal immunochemical tests for the detection of colorectal cancer. Acta Oncol 2013;52:1667-75.
18. Deutscher Wetterdienst - The German Meteorological Service. Klimadaten Deutschland
19. Huang Y, Li Q, Ge W, et al. Optimizing sampling device for the fecal immunochemical test increases colonoscopy yields in colorectal cancer screening. Eur J Cancer Prev 2016;25:115-22.
20. Chang LC, Shun CT, Hsu WF, et al. Fecal Immunochemical Test Detects Sessile Serrated Adenomas and Polyps With a Low Level of Sensitivity. Clin Gastroenterol Hepatol 2016.
21. Kim NH, Park JH, Park DI, et al. The fecal immunochemical test has high accuracy for detecting advanced colorectal neoplasia before age 50. Dig Liver Dis 2016.

22. Kim NH, Yang HJ, Park SK, et al. Does Low Threshold Value Use Improve Proximal Neoplasia Detection by Fecal Immunochemical Test? Dig Dis Sci 2016;61:2685-93.
23. Chiang TH, Chuang SL, Chen SL, et al. Difference in performance of fecal immunochemical tests with the same hemoglobin cutoff concentration in a nationwide colorectal cancer screening program. Gastroenterology 2014;147:1317-26.
24. Grobbee EJ, van der Vlugt M, van Vuuren AJ, et al. A randomised comparison of two faecal immunochemical tests in population-based colorectal cancer screening. Gut 2016.

Table 1 Overview of the nine quantitative FITs

Quantitative FIT brand	Manufacturer	Fecal sampling device (fecal mass/buffer volume)	Analytical instrument	Analytical range [μg Hb/g feces]	Preset threshold [μg Hb/g feces]
Laboratory-based					
CAREprime Hb	Alfresa Pharma, Tokyo, Japan	Specimen Collection Container A (10mg/1.9ml)	CAREprime	0.76-228.0	6.30
Hb ELISA	Immundiagnostik, Bensheim, Germany	IDK Extract (15mg/1.5ml)	Dynex System X	0.086-50.0	2.00
OC Sensor	Eiken Chemical, Tokyo, Japan	OC Auto-Sampling Bottle 3 (10mg/2.0ml)	OC Sensor io	10-200	10
RIDASCREEN Hb	R-Biopharm, Darmstadt, Germany	RIDA TUBE Hb (10mg/2.5ml)	Dynex System X	0.65-50.0	8.00
SENTiFIT-FOB Gold	Sentinel Diagnostics, Milan, Italy	SENTiFIT pierceTube (10mg/1.7ml)	SENTiFIT 270 analyzer	1.70-129.88	17.0
Point of care					
Eurolyser FOB test	Eurolyser Diagnostica, Salzburg, Austria	Eurolyser FOB sample Collector (19.9mg/1.6ml)	Eurolyser CUBE	2.01-80.4	8.04
ImmoCARE-C	CARE diagnostica, Voerde, Germany	Sample Collection Tube (20mg/2.5ml)	CAREcube	3.75-250.0	6.25
QuantOn Hem	Immundiagnostik, Bensheim, Germany	QuantOn Hem TUBE (15mg/1.5ml)	Smartphone* with App/iOS	0.30-100.0	3.70
QuikRead go iFOBT	Orion Diagnostica, Espoo, Finland	QuikRead FOB Sampling Set (10mg/2.0ml)	QuikRead go	15-200	15

FIT=Fecal immunochemical test; Hb=Hemoglobin; App=mobile application software; iOS=iPhone operating system; *iPhone 6s was used for this study.

Table 2 Study population according to screening and clinical setting

Characteristic	Participants of screening colonoscopy (main study)	Colorectal cancer patients recruited in clinical setting (ancillary study)
Total [N]	516	50
Sex		
Men, [N (%)]	287 (55.6)	30 (60.0)
Age [years]		
Range	50-79	51-78
Mean (standard deviation)	63.2 (6.4)	65.8 (7.9)
Most advanced findings [N]		
Advanced neoplasm	216	50
- Colorectal cancer	16	50
- Advanced adenoma	200	0
No advanced neoplasm	300	0
- Non-advanced adenoma	63	0
- Hyperplastic polyp	33	0
- None of above	204	0
Colorectal cancer stage* [N]		
0/I	8	15
II	1	15
III	7	15
IV	0	4
Missing	0	1

*According to the Union for International Cancer Control (UICC) classification (7th edition).

Table 3 Comparison of sensitivity and specificity of quantitative FITs at preset thresholds and at a uniform threshold

Quantitative FIT brand	Threshold [µg Hb/g feces]	Participants of screening colonoscopy (main study)					Clinical setting (ancillary study)
		Sensitivity [%] (95% CI)			Specificity [%] (95% CI)	Expected positivity rate [%]	Sensitivity [%] (95% CI)
		CRC (n=16)	AA (n=200)	AN (n=216)	No AN (n=300)		CRC (n=50)
Thresholds preset by the manufacturers							
Hb ELISA	2.00	**81.3** (54-96)	**43.5** (37-51)	**46.3** (40-53)	**85.7** (81-89)	**18.7**	**84.0** (71-83)
QuantOn Hem	3.70	**81.3** (54-96)	**41.5** (35-49)	**44.4** (38-51)	**85.7** (81-89)	**18.5**	**84.0** (71-83)
ImmoCARE-C*	6.25	**81.3** (54-96)	**35.2** (29-42)	**38.6** (32-45)	**90.0** (86-93)	**13.9**	**76.0** (62-87)
CAREprime Hb	6.30	**81.3** (54-96)	**31.0** (25-38)	**34.7** (28-41)	**91.3** (88-94)	**12.3**	**74.0** (60-85)
RIDASCREEN Hb	8.00	**81.3** (54-96)	**36.0** (29-43)	**33.3** (27-40)	**90.7** (87-94)	**13.5**	**74.0** (60-85)
Eurolyser FOB test	8.04	**62.5** (35-85)	**19.5** (14-26)	**22.7** (17-29)	**97.0** (94-97)	**5.7**	**66.0** (51-79)
OC Sensor	10.00	**68.8** (41-89)	**18.0** (13-24)	**21.8** (16-28)	**97.7** (95-99)	**6.8**	**68.0** (53-80)
QuikRead go iFOBT	15.00	**62.5** (35-85)	**18.5** (13-25)	**21.8** (16-28)	**96.7** (94-98)	**5.9**	**64.0** (49-77)
SENTiFIT-FOB Gold	17.00	**68.8** (41-89)	**18.0** (13-24)	**21.8** (16-28)	**96.3** (94-98)	**6.2**	**70.0** (55-82)
Thresholds adjusted to 15µg Hb/g feces							
RIDASCREEN Hb	15	**81.3** (54-96)	**30.5** (24-37)	**34.3** (28-41)	**94.0** (91-96)	**9.9**	**72.0** (58-84)
ImmoCARE-C*	15	**75.0** (48-93)	**23.1** (17-30)	**27.0** (21-33)	**96.0** (93-98)	**7.2**	**70.0** (55-82)
QuantOn Hem	15	**75.0** (48-93)	**22.5** (17-29)	**26.4** (21-33)	**95.0** (92-97)	**8.0**	**76.0** (62-87)
SENTiFIT-FOB Gold	15	**68.8** (41-89)	**19.0** (14-25)	**22.7** (17-29)	**96.0** (93-98)	**6.6**	**70.0** (55-82)
CAREprime Hb	15	**68.8** (41-89)	**18.0** (13-24)	**21.8** (16-28)	**97.0** (94-99)	**5.6**	**68.0** (53-80)
QuikRead go iFOBT	15	**62.5** (35-85)	**18.5** (13-25)	**21.8** (16-28)	**96.7** (94-98)	**5.9**	**64.0** (49-77)
Hb ELISA	15	**68.8** (41-89)	**17.5** (13-23)	**21.3** (16-27)	**96.3** (94-98)	**8.0**	**72.0** (58-84)
Eurolyser FOB test	15	**56.3** (30-80)	**13.5** (9-19)	**16.7** (12-22)	**98.0** (96-99)	**4.0**	**56.0** (41-70)
OC Sensor	15	**56.3** (30-80)	**13.0** (9-18)	**16.2** (12-22)	**97.0** (94-99)	**3.4**	**68.0** (53-80)

FIT=Fecal immunochemical test; Hb=Hemoglobin; CI=Confidence interval; CRC=Colorectal cancer; AA=Advanced adenoma; AN=Advanced neoplasm; *Calculation is based on 199 AA and 215 AN.

Table 4 Comparison of sensitivity and specificity of quantitative FITs at adjusted thresholds yielding defined levels of specificity

Quantitative FIT brand	Threshold [µg Hb/g feces]	Participants of screening colonoscopy (main study)					Clinical setting (ancillary study)
		Sensitivity [%] (95% CI)			Specificity [%] (95% CI)	Expected positivity rate [%]	Sensitivity [%] (95% CI)
		CRC (n=16)	AA (n=200)	AN (n=216)	No AN (n=300)		CRC (n=50)
Thresholds adjusted to 99.0% specificity							
QuikRead go iFOBT	23.00	**56.3** (30-80)	**15.5** (11-21)	**18.5** (14-24)	**99.0** (97-100)	**3.4**	**60.0** (45-74)
ImmoCARE-C*	36.80	**56.3** (30-80)	**13.1** (9-19)	**16.3** (12-22)	**99.0** (97-100)	**3.1**	**62.0** (47-75)
OC Sensor	18.20	**56.3** (30-80)	**13.0** (9-18)	**16.2** (12-22)	**99.0** (97-100)	**3.1**	**66.0** (51-79)
CAREprime Hb	26.22	**56.3** (30-80)	**13.0** (9-18)	**16.2** (12-22)	**99.0** (97-100)	**3.1**	**62.0** (47-75)
Hb ELISA	29.16	**62.5** (35-85)	**12.0** (8-17)	**15.7** (11-21)	**99.0** (97-100)	**3.0**	**62.0** (47-75)
QuantOn Hem	29.81	**62.5** (35-85)	**11.0** (7-16)	**14.8** (10-20)	**99.0** (97-100)	**2.9**	**62.0** (47-75)
SENTiFIT-FOB Gold	53.38	**56.3** (30-80)	**11.0** (7-16)	**14.4** (10-20)	**99.0** (97-100)	**2.8**	**56.0** (41-70)
Eurolyser FOB test	21.15	**56.3** (30-80)	**11.0** (7-16)	**14.4** (10-20)	**99.0** (97-100)	**2.8**	**50.0** (36-64)
RIDASCREEN Hb	Not possible to adjust the threshold above 50µg Hb/g feces						
Thresholds adjusted to 96.7% specificity							
Eurolyser FOB test	6.11	**68.8** (41-89)	**20.0** (15-26)	**23.6** (18-30)	**96.7** (94-98)	**6.1**	**70.0** (55-82)
OC Sensor	6.60	**68.8** (41-89)	**20.0** (15-26)	**23.6** (18-30)	**96.7** (94-98)	**6.1**	**68.0** (53-80)
CAREprime Hb	12.35	**68.8** (41-89)	**20.0** (15-26)	**23.6** (18-30)	**96.7** (94-98)	**6.1**	**68.0** (53-80)
ImmoCARE-C*	17.30	**62.5** (35-85)	**20.1** (15-26)	**23.3** (18-29)	**96.7** (94-98)	**6.1**	**68.0** (53-80)
QuantOn Hem	17.73	**75.0** (48-93)	**18.5** (13-25)	**22.7** (17-29)	**96.7** (94-98)	**6.0**	**74.0** (60-85)
RIDASCREEN Hb	29.54	**62.5** (35-85)	**19.0** (14-25)	**22.2** (17-28)	**96.7** (94-98)	**5.9**	**66.0** (51-79)
QuikRead go iFOBT	15.00	**62.5** (35-85)	**18.5** (13-25)	**21.8** (16-28)	**96.7** (94-98)	**5.9**	**64.0** (49-77)
SENTiFIT-FOB Gold	17.68	**68.8** (41-89)	**18.0** (13-24)	**21.8** (16-28)	**96.7** (94-98)	**5.9**	**70.0** (55-82)
Hb ELISA	15.32	**68.8** (41-89)	**17.5** (13-23)	**21.3** (16-27)	**96.7** (94-98)	**5.8**	**70.0** (55-82)
Thresholds adjusted to 93.0% specificity							
Hb ELISA	4.80	**81.3** (54-96)	**31.5** (25-38)	**35.2** (29-42)	**93.0** (90-96)	**10.9**	**76.0** (62-87)
RIDASCREEN Hb	12.27	**81.3** (54-96)	**31.0** (25-38)	**34.7** (28-41)	**93.0** (90-96)	**10.8**	**72.0** (58-84)
Eurolyser FOB test	2.01	**75.0** (48-93)	**31.0** (25-38)	**34.3** (28-41)	**93.0** (90-96)	**10.8**	**74.0** (60-85)
ImmoCARE-C*	9.20	**81.3** (54-96)	**29.7** (23-37)	**33.5** (27-40)	**93.0** (90-96)	**10.6**	**72.0** (58-84)
CAREprime Hb	6.65	**81.3** (54-96)	**29.5** (23-36)	**33.3** (27-40)	**93.0** (90-96)	**10.6**	**74.0** (60-85)
SENTiFIT-FOB Gold	1.70	**68.8** (41-89)	**28.5** (22-35)	**31.5** (25-38)	**93.3** (90-96)[#]	**10.1**	**74.0** (60-85)
QuantOn Hem	9.59	**75.0** (48-93)	**28.0** (22-35)	**31.5** (25-38)	**93.0** (90-96)	**10.4**	**80.0** (66-90)
OC Sensor	3.60	**75.0** (48-93)	**26.5** (21-33)	**30.1** (24-36)	**93.0** (90-96)	**10.2**	**72.0** (58-84)
QuikRead go iFOBT	Not possible to adjust the threshold below 15µg Hb/g feces						

FIT=Fecal immunochemical test; Hb=Hemoglobin; CI=Confidence interval; CRC=Colorectal cancer; AA=Advanced adenoma; AN=Advanced neoplasm; *Calculation is based on 199 AA and 215 AN; [#]Not possible to adjust the threshold below 1.70µg Hb/g feces.

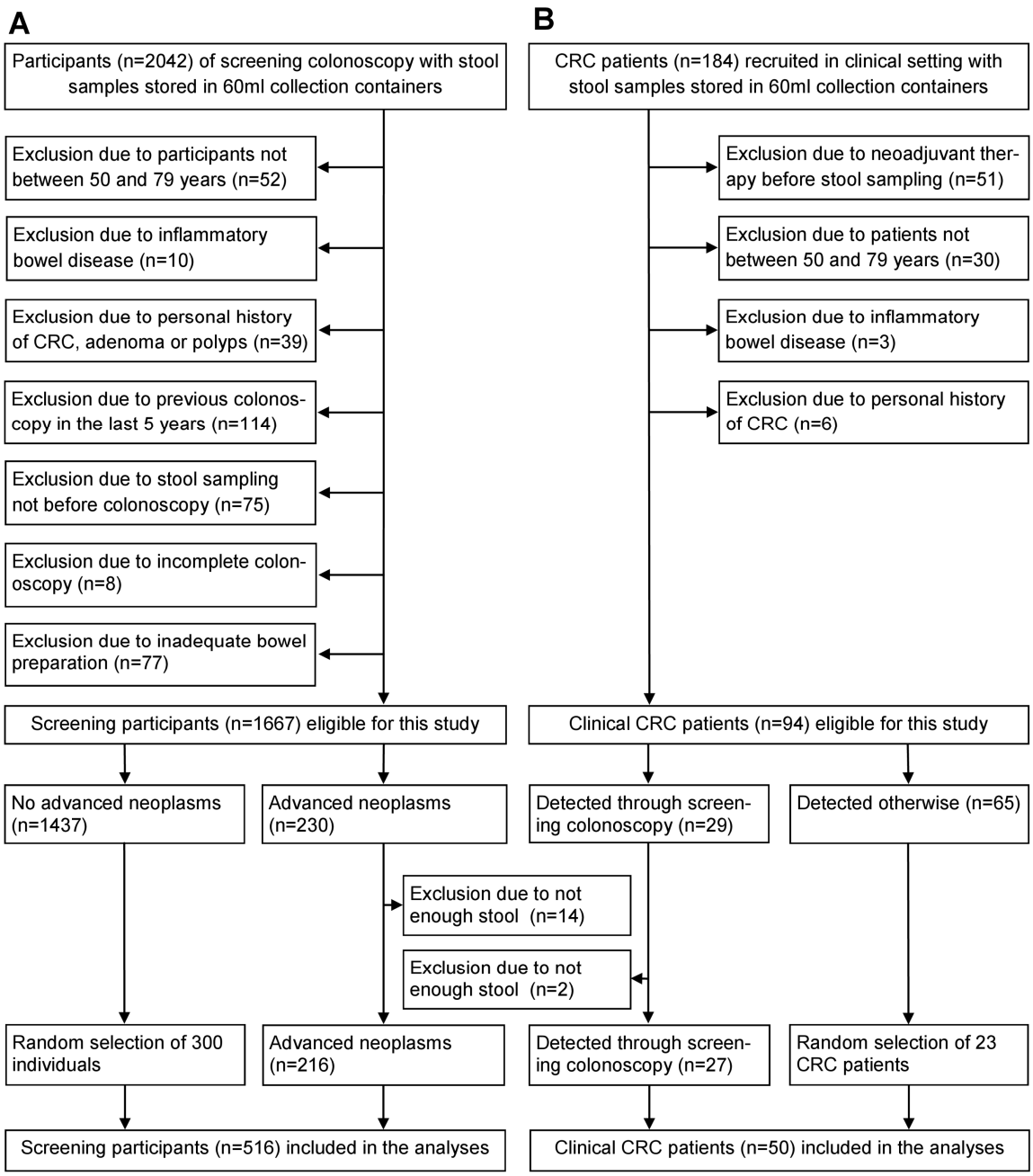
Table 5 Sensitivities according to early and late CRC stages

Quantitative FIT brand	Threshold [µg Hb/ g feces]	Sensitivity [%] (95% CI)		
		Screening and clinical CRC cases (n=65)*		
		Early stages (0/I/II) (n=39)	Late stages (III/IV) (n=26)	Difference [% units]
Thresholds preset by the manufacturers				
Hb ELISA	2.00	**76.9** (61-89)	**92.3** (75-99)	15.4
QuantOn Hem	3.70	**76.9** (61-89)	**92.3** (75-99)	15.4
RIDASCREEN Hb	8.00	**69.2** (52-83)	**84.6** (65-96)	15.4
CAREprime Hb	6.30	**71.8** (55-85)	**80.8** (61-93)	9.0
OC Sensor	10.00	**64.1** (47-79)	**73.1** (52-88)	9.0[M]
Eurolyser FOB test	8.04	**61.5** (45-77)	**69.2** (48-86)	7.7
ImmoCARE-C*	6.25	**74.4** (58-87)	**80.8** (61-93)	6.4
SENTiFIT-FOB Gold	17.00	**66.7** (50-81)	**73.1** (52-88)	6.4
QuikRead go iFOBT	15.00	**64.1** (47-79)	**61.5** (41-80)	-2.6
Thresholds adjusted to 96.7% specificity				
Hb ELISA	15.32	**64.1** (47-79)	**76.9** (56-91)	12.8
QuantOn Hem	17.73	**69.2** (52-83)	**80.8** (61-93)	11.6
ImmoCARE-C*	17.30	**61.5** (45-77)	**73.1** (52-88)	11.6
OC Sensor	6.60	**64.1** (47-79)	**73.1** (52-88)	9.0
CAREprime Hb	12.35	**64.1** (47-79)	**73.1** (52-88)	9.0[M]
RIDASCREEN Hb	29.54	**61.5** (45-77)	**69.2** (48-86)	7.7
Eurolyser FOB test	6.11	**66.7** (50-81)	**73.1** (52-88)	6.4
SENTiFIT-FOB Gold	17.68	**66.7** (50-81)	**73.1** (52-88)	6.4
QuikRead go iFOBT	15.00	**64.1** (47-79)	**61.5** (41-80)	-2.6

Hb=Hemoglobin; FIT=Fecal immunochemical test; CRC=Colorectal cancer; *One CRC patient with a missing stage classification was excluded; M=Median difference.

Figure 1 Flow diagram for selection of study population: (A) Main study (screening setting); (B) Ancillary study (clinical setting).

Figure 2 Comparison of ROC curves and AUCs among test brands: (A) for screening and clinical colorectal cancer cases and (B) for screening advanced neoplasm cases (results are based on screening setting cases only). *ImmoCare-C results are based on one advanced adenoma case less.


A
Participants (n=2042) of screening colonoscopy with stool samples stored in 60ml collection containers
Exclusion due to participants not between 50 and 79 years (n=52)
Exclusion due to inflammatory bowel disease (n=10)
Exclusion due to personal history of CRC, adenoma or polyps (n=39)
Exclusion due to previous colonoscopy in the last 5 years (n=114)
Exclusion due to stool sampling not before colonoscopy (n=75)
Exclusion due to incomplete colonoscopy (n=8)
Exclusion due to inadequate bowel preparation (n=77)
Screening participants (n=1667) eligible for this study
No advanced neoplasms (n=1437)
Advanced neoplasms (n=230)
Exclusion due to not enough stool (n=14)
Random selection of 300 individuals
Advanced neoplasms (n=216)
Screening participants (n=516) included in the analyses
B
CRC patients (n=184) recruited in clinical setting with stool samples stored in 60ml collection containers
Exclusion due to neoadjuvant therapy before stool sampling (n=51)
Exclusion due to patients not between 50 and 79 years (n=30)
Exclusion due to inflammatory bowel disease (n=3)
Exclusion due to personal history of CRC (n=6)
Clinical CRC patients (n=94) eligible for this study
Detected through screening colonoscopy (n=29)
Detected otherwise (n=65)
Exclusion due to not enough stool (n=2)
Detected through screening colonoscopy (n=27)
Random selection of 23 CRC patients
Clinical CRC patients (n=50) included in the analyses