Comparison of different FIT test kits in the Dutch CRC screening program

Clasine de Klerk*, Els Wieten*, Annemieke van der Steen, Christian Ramakers, Ernst Kuipers, Bettina Hansen, Iris Lansdorp-Vogelaar, Patrick Bossuyt, Evelien Dekker, Manon Spaander
Introduction

- Fecal immunochemical tests (FIT) are recommended as fecal test of choice for colorectal cancer (CRC) screening\(^1,2\)

- Various FITs are available for screening

- Comparative data on FIT accuracy to detect advanced neoplasia (AN\(^*\)) Needed:
  - Head-to-head comparisons (paired)
  - Enough power to determine (small) relevant differences
  - Within organized CRC screening setting

- Strong evidence is needed to allow informed decisions on FIT-brand implementation

\(^1\) Halloran et al. Endoscopy, 2012

* CRC and/or advanced adenoma: adenoma ≥ 10 mm, with ≥ 25% villous component and/or high-grade dysplasia
Dutch nationwide CRC screening

- All 55-75 year old individuals (biennially) invited to sample **FOB-Gold** (Sentinel, Italy)
- Previous pilot screening programs in NL: **OC-Sensor** (Eiken, Japan)

- FOB-Gold higher positivity rates and more non-analyzable tests?¹
- 6 months after implementation: higher positivity rates and lower PPV
- Minister of Health: Compare these FITs in the national screening program

¹ Zubero et al. Frontiers in Pharmacology, 2014
Methods

- Large paired prospective cohort study within the Dutch CRC screening program
- May 2016 to March 2017

- Screening naïve individuals aged 59, 61, 63, 71, or 75 years

- Two FITs (one FOB-Gold and one OC-Sensor) by postal mail
- Instruction: sample both in the same bowel movement

- Positivity cut-off ≥15 µg Hb/g feces (≥88 ng Hb/ml FOB-Gold and ≥75 ng Hb/ml OC-Sensor)\(^1\)
- Referral for colonoscopy intake if ≥ 1 positive FIT

\(^{1}\) Fraser et al. Journal of the National Cancer Institute, 2012
Outcomes

- **Primary outcome:**
  Difference in diagnostic yield AN between FOB-Gold and OC-Sensor
  Diagnostic yield = AN/all invitees
  *Hypothesis: equivalence (<0.15%)*\(^1,2\)

- **Secondary outcomes:**
  - Diagnostic yield of CRC\(^1,2\)
  - Paired accuracy (two completed FITs)\(^2,3\)
    - Relative true positive rate (rTPR or relative sensitivity)
    - Relative false positive rate (rFPR)
  - Non-analyzable tests
    Fecal overload, buffer loss, missing barcode, other technical problem

---

Power

Sample size calculations\(^1\):

- 1.5% diagnostic yield of AN per FIT
- 50% study participation
- Absolute difference in diagnostic yield of 0.15% or more
- 40,000 invitees → 90% power to demonstrate equivalence
- \(\alpha=0.05\)

\(^1\) Liu et al. Stat Med, 2002
Results

Invited
n=42,179

Participants
n=22,064 (52%)

≥1 FIT positive
n=2,112 (9.7%)

Colonoscopy
n=1,778 (82%)

Total AN n=716 (40%)
Total CRC n=82 (4.6%)

‘Non-participants’
- No informed consent
- No FIT returned
Diagnostic yield: equal between FITs

Detection of **AN** in all invitees (n=42,179):
- FOB-Gold 1.45% (n=610)
- OC-Sensor 1.44% (n=606)
→ absolute difference of 0.01% (CI -0.06% to 0.08%): < 0.15%

Detection of **CRC** in all invitees:
- FOB-Gold 0.18% (n=74)
- OC-Sensor 0.18% (n=78)
→ absolute difference of -0.009% (CI -0.027% to 0.008%)
Paired accuracy

Participants that completed two FITs n=21,078

<table>
<thead>
<tr>
<th>OC-Sensor</th>
<th>FOB-Gold</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Positive*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive*</td>
<td>AN</td>
<td>1,163</td>
<td>464</td>
<td>1,627 (7.7%)</td>
</tr>
<tr>
<td></td>
<td>CRC</td>
<td>500</td>
<td>104</td>
<td>604 (2.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
<td>7</td>
<td>77 (0.4%)</td>
</tr>
<tr>
<td>Negative</td>
<td>AN</td>
<td>419</td>
<td>19,032</td>
<td>19,451 (92.3%)</td>
</tr>
<tr>
<td></td>
<td>CRC</td>
<td>83</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>AN</td>
<td>1,582 (7.5%)</td>
<td>19,496 (92.5%)</td>
<td>21,078</td>
</tr>
<tr>
<td></td>
<td>CRC</td>
<td>583 (2.8%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>73 (0.4%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Overall colonoscopy attendance rate of participants with ≥positive FIT(s) was 84%
Relative TPR & FPR: FOB vs OC

rTPR of AN: 0.97 (95%CI: 0.92 to 1.01)

rFPR of AN: 0.99 (95%CI: 0.93 to 1.05)

rTPR of CRC: 0.95 (95%CI: 0.87 to 1.03)
## Reasons for non-analysable FIT tubes

<table>
<thead>
<tr>
<th>Total tests returned</th>
<th>FOB-Gold n=22,057</th>
<th>OC-Sensor n=21,369</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-analysable tests, n (%)</td>
<td>49 (0.22)</td>
<td>14 (0.07)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Technically impossible</td>
<td>10 (20)</td>
<td>4 (29)</td>
<td>0.52</td>
</tr>
<tr>
<td>Barcode unreadable</td>
<td>5 (10)</td>
<td>4 (29)</td>
<td>0.09</td>
</tr>
<tr>
<td>Broken tube</td>
<td>2 (4.1)</td>
<td>0</td>
<td>0.45</td>
</tr>
<tr>
<td>No buffer</td>
<td>23 (47)</td>
<td>1 (7.1)</td>
<td>0.005</td>
</tr>
<tr>
<td>Sample too large</td>
<td>6 (12)</td>
<td>1 (7.1)</td>
<td>0.60</td>
</tr>
<tr>
<td>Sample too small</td>
<td>3 (6.1)</td>
<td>3 (21)</td>
<td>0.09</td>
</tr>
<tr>
<td>No sample taken</td>
<td>0</td>
<td>1 (7.1)</td>
<td>0.06</td>
</tr>
</tbody>
</table>
Conclusions and implications

- **FOB-Gold and OC-Sensor** have similar accuracy in detecting **AN** and **CRC** in an organized CRC screening program.

- Other features can guide informed decision making when selecting one of the two brands for CRC screening:
  - Participation, ease of use, ease of laboratory analysis, cost-effectiveness

- Apart from its results, our study may serve as an example how to assess and possibly improve screening effectiveness within an ongoing program.

  .. and may, hopefully, inspire future initiatives.
Acknowledgements

bevolking onderzoek zuid-west
Ministerie van Volksgezondheid, Welzijn en Sport
ZonMw

Star Medisch Diagnostisch Centrum
DaklaPack® europe
EIKEN CHEMICAL CO., LTD.

Erasmus MC Universitair Medisch Centrum Rotterdam
reumatologie
post.nl

NKI-AVL The Netherlands Cancer Institute Antoni van Leeuwenhoek Hospital

bevolking onderzoek landelijk beheer
Positive predictive value

Positivity
FOB-Gold 7.5%
OC-Sensor 7.7% (p=0.140)

PPV Advanced neoplasia
43.8% for FOB-Gold
44.3% for OC-sensor
(absolute difference 0.5%; 95%CI: -3.3% to 4.2%, p <0.05).

PPV CRC
5.5% for FOB-Gold
5.7% for OC-sensor
(absolute difference 0.2%; 95%CI: -1.5% to 2.0%).
Statistical analyses and power

- **Statistical analysis:**
  - Diagnostic yield: equivalence test for paired binary proportions
    - Method Lui et al\(^1\)
  - 95% confidence intervals for paired proportions
    - Wald interval with Bonett-Price adjustment\(^2\)
  - Relative TPR and relative FPR
    - Method Alonzo et al\(^3\)

- **Sample size calculations\(^1\):**
  - 1.5% diagnostic yield of AN per FIT
  - Absolute difference in diagnostic yield of 0.15% or more
  - 40,000 invitees → 90% power to demonstrate equivalence
  - \(\alpha=0.05\)

---

\(^1\) Liu et al. Stat Med, 2002  
\(^2\) Fagerland et al. Stat Med, 2014  
\(^3\) Alonzo TA et al. Stat Med, 2004
Tests for equivalence for paired binary data:

Commonly used and accepted approach to assessing equivalence compares the asymptotic confidence interval on the difference of two response rates with some clinical meaningful equivalence limits.

Article compares: two asymptotic test statistics, a Wald-type (sample-based) test statistic and a restricted maximum likelihood estimation (RMLE-based) test statistic

A bioequivalence study is tested by comparing the 100(1–2) per cent confidence limits of the difference with the equivalence limits (−; ) Equivalence is inferred if the entire confidence interval falls within the equivalence limits.

The RMLE-based test without the continuity correction appears to be an acceptable approach for establishing the equivalence or non-inferiority between two paired binary endpoints.

- Evaluated 24 methods for the analysis of paired binomial proportions and 95% CI:
  - Wald with Bonett–Price Laplace adjustment
  - Hoewel bij ons weinig verschillen want grote aantallen
Relative accuracy for binary screening tests (only positives gold standard)

Table I. Data from a screen positive study comparing accuracy of screening tests A and B. Data in brackets are unknown.

<table>
<thead>
<tr>
<th></th>
<th>Diseased</th>
<th></th>
<th>Non-diseased</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B+</td>
<td>B−</td>
<td>B+</td>
<td>B−</td>
</tr>
<tr>
<td>A+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td></td>
<td>b</td>
<td>m₁</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>[d]</td>
<td>m₂</td>
<td>[n_D]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A−</td>
<td>m₂</td>
<td>[n_D]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td></td>
<td>f</td>
<td>m₃</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td></td>
<td>[h]</td>
<td>m₄</td>
<td>[n_D]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m₄</td>
<td></td>
<td>[n_D]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\hat{rTPR} = \frac{a + b}{a + c} \\
\hat{rFPR} = \frac{e + f}{e + g}
\]