Title:
Recommendations for a step-wise comparative approach to the evaluation of new screening tests for colorectal cancer

Running Title: Colon cancer screening test evaluation

Authors
Graeme P. Young, MD, FRACP, FTSE AGAF. Flinders Centre for Innovation in Cancer, Flinders University. Adelaide Australia
Carlo Senore, MD MSc. SCDO Epidemiologia, AOU Città della Salute e della Scienza – CPO Piemonte, Torino, Italy
Jack S. Mandel, PhD MPH. Emeritus Professor, University of Minnesota, Minneapolis, Minnesota, USA.
James E. Allison, MD. Kaiser Division of Research, University of California, and Kaiser Division of Research, San Francisco, USA.
Wendy S. Atkin, MPH, PhD. Gastrointestinal Epidemiology, Imperial College, London, UK.
Robert Benamouzig, MD PhD. Gastroenterology Department, Avicenne Hospital, Paris 13 University, Paris, France.
Patrick Bossuyt, PhD. Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands.
Mahinda De Silva, MB BS. Department of Gastroenterology, Repatriation General Hospital, Adelaide, Australia.
Lydia Guittet, MD PhD. INSERM U1086 Cancers & Preventions Centre, Caen University Hospital, Caen, France.
Stephen P. Halloran, MBE FRCPath. University of Surrey, Guildford UK.
Ulrike Haug, PhD. Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology - BIPS Achterstr. 30, 28359 Bremen, Germany.
Geir Hoff, MB ChB PhD. Telemark Hospital, Skein; Cancer Registry of Norway, University of Oslo, Oslo, Norway.
Steven H. Itzkowitz, MD, FACP, FACG, AGAF. GI Division, Icahn School of Medicine at Mount Sinai, New York, USA.
Marcis Leja, MD, MBA, PhD, AGAF. Digestive Diseases Centre, University of Latvia, GASTRO Riga, Latvia.
Bernard Levin, MB Bch FACP. Division of Cancer Prevention, University of Texas M.D. Anderson Cancer Center, Houston, USA.

Gerrit A. Meijer, MD PhD. Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

Colm A. O’Morain, MD. Faculty of Health Science, Trinity College Dublin, Dublin, Ireland.

Susan Parry, MbCHB. FRACP. Ministry of Health Bowel Cancer Programme, Auckland Hospital, Auckland, New Zealand.

Linda Rabeneck, MD MPH. Prevention and Cancer Control, Cancer Care Ontario, and University of Toronto, Toronto, Canada.

Paul Rozen, MD. Department of Gastroenterology, Sestopali Fund for Gastrointestinal Cancer Prevention, Tel Aviv, Israel (posthumous)

Hiroshi Saito, MD PhD. Research Center for Cancer Prevention and Screening, National Cancer Center, Tokyo, Japan.

Robert E. Schoen, MD PhD. Department of Medicine and Epidemiology University of Pittsburgh, Pittsburgh, USA.

Helen E. Seaman, BSc PhD. NHS Bowel Cancer Screening Southern Programme Hub, Royal Surrey County Hospital, Guildford, UK.

Robert J.C. Steele, MD FRCS. Department of Surgery, University of Dundee, Dundee, Scotland.

Joseph J.Y. Sung, MD PhD. Vice Chancellor and President, The Chinese University of Hong Kong, Shatin, China.

Sidney J. Winawer, MD. Gastroenterology and Nutrition Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York USA.

Corresponding author:

Graeme P Young, MD FRACP,
Flinders Centre for Innovation in Cancer,
Flinders University,
Adelaide, Australia

Postal Address: FCIC, Level 3, Flinders Medical Centre, Bedford Park, Adelaide, SA 5042, Australia

graeme.young@flinders.edu.au

Phone +61 3 9850 8630
Fax: +61 8 8204 3943
Numbers: 1) text pages – 16 excluding Tables etc.; 2) Tables - 3 and Boxes - 3; 3) Figures – 1 and 4) On-line supplementary tables – 2.

Funding:
This work was supported by a grant from the World Gastroenterology Organization.

Conflict of interest Disclosures:
The following authors declare the following potential competing interests: Consultancy (GY, BL, SI, SW, GM), Grants or support for research (GY, WA, GM), Patents - holders and applicants (GY, GM)

Précis:
To provide practical guidance on how to compare new with proven screening tests for colorectal cancer, a panel of experts has reviewed the literature and propose a four-phase evaluation process that includes comparison to existing proven tests but which does not necessarily require RCTs with mortality as the endpoint. New screening tests can be evaluated efficiently by this stepwise comparative approach.

Key words: colorectal cancer, screening test, fecal occult blood test, molecular diagnostics, colonoscopy.

Abbreviations: FOBT, fecal occult blood test; gFOBT, guaiac-based fecal occult blood test; FIT, fecal immunochemical test for hemoglobin; CRC, colorectal cancer; RCT, randomized controlled trial; QALY, quality adjusted life years; ROC, receiver operating characteristic; TP, true-positive hence TPR (TP rate); FP, false-positive hence FPR (FP rate); TN, true-negative; FN, false-negative.
Abstract (245 words)

New screening tests for colorectal cancer continue to emerge but the evidence needed to justify their adoption in screening programs remains uncertain. A review of the literature and a consensus approach by experts was undertaken to provide practical guidance on how to compare new with proven screening tests.

Findings and recommendations: Adoption of a new screening test requires evidence of effectiveness relative to a proven comparator test. Clinical accuracy supported by programmatic population evaluation in the screening context on an intention-to-screen basis, including acceptability, are essential. Cancer-specific mortality is not essential as an endpoint provided that the mortality benefit of the comparator has been demonstrated and that the biological basis of detection is similar. Effectiveness of the guaiac-based fecal occult blood test provides the minimum standard to be achieved by a new test. A four-phase evaluation is recommended. An initial retrospective evaluation in cancer cases and controls (Phase 1), followed by a prospective evaluation of performance across the continuum of neoplastic lesions (Phase 2). Phase 3 follows demonstration of adequate accuracy in these two prescreening phases and addresses programmatic outcomes at one screening round on an intention-to-screen basis. Phase 4 involves more comprehensive evaluation of on-going screening over multiple rounds. Key information is provided from the following parameters: test positivity rate in a screening population, true- and false-positive rates, and number needed to colonoscope to detect a target lesion.

Conclusions: New screening tests can be evaluated efficiently by this stepwise comparative approach.
INTRODUCTION

New tests to screen for colorectal cancer (CRC) continue to emerge and are based on new biomarkers, new imaging modalities, or variations to existing methods. Efficient evaluation of these options presents a challenge. It has been observed that new *diagnostic* tests frequently enter practice without evidence of improved outcomes. For *screening* tests, the requirement for evidence is more demanding since more than clinical test accuracy (i.e. sensitivity and specificity) is required to justify adoption. Safety, public acceptability and cost-effectiveness need to be assessed even more carefully for tests that are to be applied to ostensibly healthy people.

The intention of a cancer screening program, or secondary prevention, is to significantly reduce cancer-site-specific mortality and burden of that disease in the target population through programmatic use of a test that detects neoplasia at a stage early enough for treatment to be successful and/or incidence to be reduced.

Certain screening tests have been shown to reduce cancer-site-specific mortality and/or incidence by randomized population-based evaluation on an intention-to-screen basis, thereby limiting biases such as lead-time, length, and self-selection that are often present in simpler studies that employ surrogate measures of mortality or intermediate endpoints. Evaluation of every new CRC screening test to the endpoint of mortality would be a huge and expensive undertaking and would markedly slow - if not prohibit - implementation of promising new technologies. Fortunately, simpler studies employing surrogate measures or intermediate endpoints can be used to evaluate new tests, provided that a carefully validated reference standard is used and biases are minimized.

To define what is justifiably required to support the use of a new test for CRC screening, we propose an efficient and rigorous method for how to compare the alternative/new (hereafter “new”) with the proven/established screening tests.

METHOD

To establish the guiding principles for comparative evaluation, including the informative endpoints and the appropriate study design, we established a consensus based on the Glaser and Delphi approaches. The membership was chosen from experts because of their knowledge or experience in practice or research relevant to screening for CRC. The problem was defined by using the consensus process to agree on the goal. To support the consensus process, systematic literature searches were undertaken using Medline and other relevant databases. One search string was optimized for diagnosis and screening with inclusion of measures such as sensitivity, another was optimized for cancer and a third attempted to identify papers focusing on comparison of tests. We also searched for review papers addressing the evidence supporting screening for CRC cancer.
A series of specific questions that focused on the definition of appropriate study designs and outcomes for the comparison of different screening tests were established by agreement. The answers to these questions were reached by consensus (requiring 75% agreement) based on dissemination of summaries of the literature searches, detailed examination of methodological papers, a series of semi-structured discussions with dissemination of decisions following each critique, followed by consultation with external advisors. Based on these processes, progressive drafts of the recommendations were then prepared, circulated and critiqued.

This paper presents:

1. the underlying guiding principles which emerged from the consensus,
2. an expert opinion on the methods appropriate to evaluate a new test in comparison to a proven comparator test (what is needed),
3. practical guidance on how to apply these methods in a four-step phased evaluation (how to do it), and
4. examples of published research that exemplify these phases (how it has been done).

As such, it will guide researchers and enable practitioners to decide if a new test is suitable for the context in which they practice.

GUIDING PRINCIPLES

The guiding principles that emerged from the consensus approach and the literature review are outlined in Box 1, together with their key consequences for test comparison. A presentation of the reasoning underlying these principles is presented in Supplementary Table 1 (on-line).

With regard to Principle 3, which states that “Population randomized controlled trials (RCTs) set the standard for evaluation of new tests”, Table 1 outlines the characteristics of major screening tests known to reduce CRC mortality and/or incidence together with the type of evidence supporting their value. Such tests are ideal as a reference point against which to compare a new test. Table 1 also describes the test target (which serves as an informative outcome for comparison), as discussed in Principle 5.
Box 1: The guiding principles that underpin a strategy for comparing screening tests which emerged from the consensus approach and the literature review. (See Supplementary Table 1 (on-line) for a discussion of each.)

1. Screening aims to reduce the burden of disease in the targeted population, without adversely affecting the health status of those who participate in screening, through early detection and treatment of cancer and/or through detection of pre-cancer lesions, which reduces incidence.

2. The screening test is just one event in a process that includes engagement of the public, testing, validation, communication and treatment.

3. Population randomized controlled trials (RCTs) with mortality as primary outcome set the standard for evaluation of new tests.

4. New tests can be assessed in parallel to an existing test all the way through the screening process from population engagement to population outcomes/measures.

5. New screening tests might detect a different neoplasia-dependent biology, and as a consequence, the value of treatment and benefit to mortality reduction might not be the same.

6. In two-step screening, the screening test selects participants who proceed to diagnostic verification by colonoscopy, as a positive test increases the likelihood of neoplasia being present.

7. It is not ethically justifiable to proceed to study a test in the screening environment including acceptability to invitees or other screening program outcomes, without studies that indicate that the new test is of acceptable accuracy when compared to a proven comparator test.

8. New tests must be clearly defined with provision of adequate technical details, quality assurance procedures and performance standards.
Table 1: Characteristics of established screening tests known to reduce CRC mortality and the type of evidence supporting their value. This information is derived from several publications 5,6,15-20.

<table>
<thead>
<tr>
<th>Detection Goal</th>
<th>Technology</th>
<th>Strongest evidence for benefit</th>
<th>Test objective</th>
<th>Sensitivity determinant</th>
<th>Specificity determinants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fecal blood</td>
<td>Guaiac-based FOBT (gFOBT)</td>
<td>Population RCTs – reduced incidence and mortality</td>
<td>Heme component of hemoglobin</td>
<td>Amount of fecal heme exceeds that needed to generate a positive result (fixed by manufacturer)</td>
<td>Dietary peroxidases; agents interfering with peroxidase reaction; bleeding nonneoplastic lesions; amount of stool in sample.</td>
</tr>
<tr>
<td>Fecal immunochemical test for hemoglobin (FIT)</td>
<td>Case-control &amp; cohort studies – reduced incidence and mortality. Comparative screening cohorts (randomized) – higher detection rates and participation compared to gFOBT.</td>
<td>Globin component of human hemoglobin</td>
<td>Amount of fecal hemoglobin exceeding selected cut-off concentration (may be fixed by manufacturer or selected by end-user)</td>
<td></td>
<td>Bleeding nonneoplastic colonic lesions; amount of stool in sample.</td>
</tr>
<tr>
<td>Endoscopic visualization of lesion</td>
<td>Colonoscopy</td>
<td>Case-control &amp; cohort studies – reduced incidence and mortality</td>
<td>Visually apparent lesions (ulcerative, polypoid or flat/depressed) suspicious of neoplasia</td>
<td>Quality of procedure; Ability to negotiate the colonic lumen with adequate views; Nature of the lesion</td>
<td>Histopathological clarification</td>
</tr>
<tr>
<td></td>
<td>Sigmoidoscopy (flexible)</td>
<td>Population RCTs – reduced incidence and mortality</td>
<td>Visually apparent lesions within reach</td>
<td>Quality of procedure; depth of insertion; Ability to negotiate the colonic lumen with adequate views; Nature of the lesion</td>
<td>Histopathological clarification</td>
</tr>
</tbody>
</table>

Abbreviations: FOBT, fecal occult blood test; gFOBT, guaiac-based fecal occult blood test; FIT, fecal immunochemical test for hemoglobin; CRC, colorectal cancer; RCT, randomized controlled trial.
A FRAMEWORK FOR EVALUATING A NEW SCREENING TEST

With these principles in mind, a practical framework for evaluating a “new” against a proven test can be built. The test of effectiveness for the proven test demands proof at the population level—hence the context for evaluation must eventually include population outcomes and not just the testing of capacity to detect lesions.

When a RCT establishes that a test is effective in reducing mortality, then a new test does not need to be evaluated with such rigor provided it is compared to the proven test. This is true provided that Principle 5 (Box 1) applies, namely that the value of treatment and benefit in mortality is not compromised due to potential differences in the biology of detected lesions.

In applying this view, there are three types of readily-determined outcome other than effects on CRC mortality and stage, that inform the value of a new test: accuracy, acceptability and impact on other screening program outcomes when applied in a screening context (see “Phased evaluation”). Such intermediate/surrogate outcomes facilitate prediction of benefit provided that the new test is directly compared to one that has been proved to be effective on an intention-to-treat basis, i.e. based on an approach that amongst others takes into account imperfect adherence and overcomes other sources of bias.

STUDY DESIGN FOR COMPARING TESTS

Accuracy can be assessed through case-control and cohort studies using the framework shown in Figure 1. This framework can be adapted to any phase of evaluation from prescreening assessment to mass population application.

Choice of comparator test

The first and well-characterized non-invasive test (in terms of effectiveness) is the gFOBT Hemoccult (and variants particularly Hemoccult II). The screening outcomes achieved with this gFOBT represent the minimum that needs to be achieved since the effect of gFOBT on mortality is modest. The more advanced technology provided by fecal immunochemical tests for hemoglobin (FIT) provides better accuracy including improved sensitivity for adenomas as well as CRCs and better acceptability when evaluated on an intention-to-screen basis. Population-based and case-control studies support the value of this technology. Further studies from the Netherlands confirm the value of FIT in a population RCT when analyzed on an intention-to-screen basis relative to the gFOBT Hemoccult II. This evidence has led to recommendations that FIT replace gFOBT. As such, a well-studied FIT sets a new standard against which new tests can be judged. FIT technology tends to have a better capacity to detect adenomas than do gFOBT and repeated testing improves detection.
As population screening trials with FS have now been reported, this screening test will serve as a useful comparator for detection of pre-invasive lesions.

The experts concluded that colonoscopy serves to estimate accuracy of a new test, however, without RCT intention-to-screen evidence of effectiveness, the effectiveness of a new non-invasive test cannot be deduced if assessed relative to colonoscopy only. However, as results emerge from the currently-underway population screening trials evaluating colonoscopy, we will be able to use colonoscopy as a comparator knowing its benefit to mortality in an un-biased setting.

**EVALUATION OF ACCURACY**

Clinical accuracy (sensitivity, specificity and predictive values) is crucial to whether a new test is fully evaluated in screening. It is not appropriate to study acceptability, or other screening program outcomes, without having first measured accuracy. Consequently, comprehensive test evaluation must be phased – see Principle 7.

The two key measures of accuracy – sensitivity and specificity – are often difficult to ascertain, especially for screen-relevant lesions (i.e. the earlier-stage cancers and adenomas that one would encounter in a largely asymptomatic typical screening population). A valid estimate of these accuracy measures would require costly and time-consuming testing of an unselected screening population containing sufficient cases with such lesions where all test confounders were likely to be encountered and where every subject, both test-positive and test-negative, underwent diagnostic verification.

Fortunately, when a comparator test is available, a paired study design (which improves statistical power) facilitates evaluation of effectiveness of the new test and estimation of the relative impact on screening outcomes. We conclude, in line with others, that existing tests, namely gFOBT/FIT and FS, have been shown to be effective and can be used to facilitate assessment of relative benefit.

Another simplification is based on the proposition that the two key questions concerning clinical accuracy are: 1) detection – a test that is more sensitive in practical terms returns more true-positives, and 2) the burden associated with detection – a test that is more specific in practical terms returns fewer false-positives. The assessment of these two parameters is achieved by a thorough diagnostic verification of every test-positive case (both comparator and new test positives) to determine if it is a true-positive or a false-positive.

As shown in Table 2, the simple dichotomous measures of true-positive rate (TPR) and false-positive rate (FPR) are direct and practical measures of accuracy, sometimes referred to as test “operating characteristics”. They are used when undertaking ROC (receiver operating characteristic)
analysis. TPR reflects detection (sensitivity) and FPR the burden associated with detection (1-specificity). Consequently, relative sensitivity and specificity are determined by comparing TPR and FPR, respectively, between tests.
Table 2: The relationship between direct, practical measures (operating characteristics) of a screening test result, how each informs assessment of test accuracy and what the consequence of the result is for a screening program.

<table>
<thead>
<tr>
<th>Test result</th>
<th>Diagnostic verification; operating characteristic</th>
<th>Corresponding accuracy characteristic</th>
<th>Issue addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True (i.e. target condition* present); true-positive rate (TPR)</td>
<td>Sensitivity (positivity rate in those with the target condition).</td>
<td>Detection.</td>
</tr>
<tr>
<td></td>
<td>False (i.e. target condition* not present); false-positive rate (FPR)</td>
<td>Specificity (1 – FPR)</td>
<td>Burden associated with detection</td>
</tr>
<tr>
<td>Negative</td>
<td>True; true-negative rate (TNR)</td>
<td>Negative predictive value (TNR/TNR + FNR)</td>
<td>Elimination/exclusion of targeted clinical lesion (stage specified)</td>
</tr>
<tr>
<td></td>
<td>False; false-negative rate (FNR)</td>
<td>Missed lesion</td>
<td>Burden of failed detection</td>
</tr>
</tbody>
</table>

* Targeted clinical lesion is either cancer and/or advanced adenoma depending on the question being asked of the test as tests might detect these to differing degrees.

**Abbreviations:** TP, true-positive hence TPR (TP rate); FP, false-positive hence FPR (FP rate); TN, true-negative hence TNR; FN, false-negative hence FNR.
Comparing test accuracy – the scenarios

The approach based on verification of positive tests, classifying them as true- or false-positives, provides a straightforward but powerful strategy for comparing the accuracy (operating characteristics) of two screening tests. The concepts presented apply regardless of whether the target lesion is cancer and/or adenoma.

In comparing accuracy, one needs to clearly define the targeted clinical lesion (hereafter referred to as targeted lesion), which can be cancer and/or adenoma or combinations thereof. Performance characteristics related to sensitivity and specificity need to be compared for the same clinical endpoint. Depending on the phase of evaluation and the question being addressed, the target lesion might be early-stage cancer or advanced adenoma or “advanced neoplasia” a term referring to cancer plus advanced adenoma (see Phase 2 below for definition). Tests might differ in their capacity to detect lesions at specific stages and this needs to be explored. It should be noted that clinical accuracy depends on the presence of the biomarker that forms the basis of the test objective (see Table 1) and this in turn might be important to treatment response (Principle 5 and online supplementary Table1).

Two simple questions, modified from Lord et al, guide assessment in a practical manner.

1) Is the new test better at detecting target lesions?

This is true if the TPR (which reflects sensitivity) for the target lesions, is improved using the new test. It is likely that improved outcomes (reduced mortality and/or incidence) will follow from use of the new test especially if the TPR is greater for early stage cancers.

Complexity arises if the new test is better at detection (higher sensitivity) but it returns more false-positives (lower specificity) than the old test, raising concerns about cost and potential harms. Hemoccult Sensa compared to Hemoccult II is an example. Note, however, that a test with more true positives and higher initial colonoscopy rate (whether due to true- and/or false-positives) will make the program more expensive initially but might create longer-term savings as a result of better detection. This will become clearer in formal cost-effectiveness analyses that measure cost per QALY saved.

There are several ways to address such complex scenarios. The operating characteristics of the two tests can be plotted as an ROC curve (TPR versus FPR) as a way to judge which test has the best balance of true and false positives; overall, the test with the greatest area under the curve has the best discriminatory power. This is particularly applicable to pre-screening phases in the evaluation process that focuses on accuracy (see below).
Another objective approach is to calculate the number needed to colonoscope for detecting one target lesion using each test (the reciprocal of the positive predictive value). Calculating the number needed to colonoscope also facilitates comparison of two tests when each is applied to a different cohort, although comparability of populations needs careful consideration. However, the number needed to colonoscope should be determined only in Phase 3 studies conducted in settings that represent the natural prevalence of neoplasia and not in studies in which prevalence is biased due to recruitment processes.

2) If not better at detecting target lesions, does it have other advantages?

A new test might have other benefits. For instance, a significantly better specificity without improved sensitivity. Comparison is made simply in this circumstance by calculating the number needed to colonoscope to detect one target lesion for each test. The new test might also have programmatic benefits (see Phase 3 evaluation) such as greater acceptance by the screening population or improved technical reliability. In similar fashion, the number needed to invite to detect one target lesion will offer additional comparative information by capturing the product of participation and accuracy, although this approach is susceptible to the method of invitation and how the invitation is framed. It should be noted that many consider the sensitivity of gFOBT, which has shown a statistically significant but only relatively small impact on CRC mortality, to be inadequate. As a consequence, they would argue that there is only a place for a new test that returns a better sensitivity than gFOBT.

Study Populations

The population selected for study will depend on the question being asked and the phase of the evaluation. The testing path may involve paired testing in a single group (that comprises cases and controls) or parallel testing of randomized cohorts (see Figure 1). Which is chosen depends on the stage of evaluation (see Box 2). The subsequent discussion on phased evaluation provides more detail.
Box 2: Study populations and testing path.

- **Initial testing of accuracy (Phases 1 and 2):** Ideally a single clinical group of patients undertaking *paired* testing (i.e. each person does both the new and old test) as shown in Figure 1. This is an efficient design. Initially diagnostic verification of all cases by colonoscopy is carried out regardless of test result. Pairing reduces cohort size because of improved statistical power for assessing incremental benefit. It ensures subjects are comparable and avoids imbalances in variables that affect test results and in other biases between the tests. If the new test shows promise, then larger numbers of subjects undertaking paired testing can be further studied with colonoscopic follow-up in test-positive cases only.

- **Subsequent testing in the screening context (Phases 3 and 4):** Individuals may be randomly assigned to do either the proven or the new test, in the context of the screening pathway on an intention-to-screen basis, when it has been shown first that the accuracy of the new test is not worse than that of a suitable proven comparator test. When assessing test accuracy in parallel groups, the inclusion criteria for the study group must be carefully characterized and the detected lesions fully described. Without this, transferability from one setting to another is not possible. 

---

This article is protected by copyright. All rights reserved.
COMPARING TESTS IN THE SCREENING PATHWAY

In addition to accuracy, it is essential that the effect of a new test on other variables in the screening pathway is determined, e.g. safety, cost, feasibility, ease of use for a screening participant and acceptability. New tests must undergo evaluation in unselected typical screening populations and intention-to-screen evaluation is necessary to justify large-scale adoption.

In mass population screening, detection of target lesions is the product of participation and sensitivity because without participation (sometimes referred to as compliance or uptake) there can be no detection. Consequently, measuring participation with one test relative to another in separate cohorts randomly selected from the same population can document test acceptability, provided that framing of information is carefully balanced.

PHASED EVALUATION

Phased (i.e. sequential) evaluation in a step-wise increasingly complex manner is most appropriate. Initial evaluation (Phases 1 and 2) starts with a simple prescreening evaluation that addresses accuracy of the new test and proceeds, if judged appropriate, to more thorough evaluation addressing outcomes in the population screening context (Phases 3 and 4) as shown in Box 3. Phased evaluation takes into account the issues shown in Box 3.

The primary and secondary aims and general characteristics of these Phases are provided in Table 3.

The phased approach is indeed undertaken in practice. Supplementary Table 2 (on-line) provides selected examples of studies that have been conducted that demonstrate elements of each Phase, together with their main characteristics. If one tracks through these phases for tests such as different FIT products and designs, and the fecal DNA tests, it can be seen that early simple studies are followed by more complex and informative studies. There are many other possible examples - those provided serve to demonstrate the increasing complexity of each Phase, what the design options are within a Phase and what information might be gleaned from such studies.
Box 3: The four Phases of test evaluation and associated issues

Phases:
1. Retrospective estimation of ability to discriminate between cancer cases and normal
2. Detection of pre-symptomatic stages along the neoplastic continuum, prospective clinical studies
3. Initial screening evaluation – participation and prevalence studies
4. Screening program evaluation

Issues to be noted:
- In two-step screening, screening tests select subjects who then undergo the reference diagnostic test,
- Pathway parameters in screening, such as participation rates, are as crucial to population benefit as is technical performance,
- Relative test accuracy is simply addressed in a paired design,
- The value of the new test should be compared with the old test in the context of how the new test is to be implemented in the existing screening pathway,
- The specific phases of screening are a guide to evaluation reflecting a continuum from simple to increasingly complex evaluation where each step may be adjusted for complexity according to outcomes in the prior phases.
- Costing each phase is subject to local considerations but if one puts aside the costs of diagnostic verification, Phase 1 studies might cost several hundred thousand dollars while Phase 3 and 4 will cost several to many millions of dollars.
Table 3: Phased evaluation for comparison of screening tests for colorectal cancer. Discussion of group sizes and approximate costs for each phase are included in the text.

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Nature</th>
<th>Primary aim</th>
<th>Secondary aims</th>
</tr>
</thead>
</table>
| Phase 1    | Prescreening: Retrospective estimation of ability to discriminate between cancer cases and controls without neoplasia. | **Test detects established cancer:**  
1.1 To estimate TPR* and FPR* (test operating characteristics) as the primary measures of accuracy relative to an established test.  
1.2 Establish the test sampling process.  
1.3 Optimize processes for quality assurance.  
1.4 Fine tune test endpoint. |  
1.2 Establish the test sampling process.  
1.3 Optimize processes for quality assurance.  
1.4 Fine tune test endpoint.  
Cases known to have cancer, ideally with a majority with potentially curable stages and including some who are asymptomatic.  
Controls to be free of neoplasia.  
Concordance between tests should be reported.  
Ideally paired testing, all results verified at diagnostic procedure. |
| Phase 2    | Detection of lesions along the neoplastic continuum; prospective clinical studies. | **Test detects early neoplasia before it becomes apparent:**  
2.1 To estimate test operating characteristics for detection of neoplasia at stages along the oncogenesis continuum, especially pre-clinical disease including advanced adenomas.  
2.2 To determine the final format of the test (sample and end point).  
*Minimum requirement for test registration.*  
2.3 More reliably estimate operating characteristics.  
2.4 Information on co-variates affecting test performance.  
2.5 Ascertain number of samples and threshold (fine-tune the endpoint).  
2.6 Test to be registerable with authorities.  
2.7 Clarify if there are subgroups where the test might fail to detect lesions. |  
2.3 More reliably estimate operating characteristics.  
2.4 Information on co-variates affecting test performance.  
2.5 Ascertain number of samples and threshold (fine-tune the endpoint).  
2.6 Test to be registerable with authorities.  
2.7 Clarify if there are subgroups where the test might fail to detect lesions.  
Cases covering all stages of colorectal neoplasia especially early-stage cancer and/or advanced adenomas with knowledge of whether cases are symptomatic.  
Asymptomatic where possible.  
Controls to be free of neoplasia.  
Results in subjects with common benign diseases and how they affect test result need ascertainment.  
Testing undertaken prior to scheduled diagnostic procedure.  
Ideally paired testing.  
Concordance between tests should be reported. |
Table 3: (continued).

<table>
<thead>
<tr>
<th>Phase 3</th>
<th>Initial screening evaluation; single round of screening</th>
<th>Characteristics of neoplasia detected when screening; false-referral rate; acceptability:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3.1 In a screening population, to determine the operating characteristics of the test, what is detected and the workload associated with detection including the false referral rate.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.2 Determine test acceptability. Minimum requirement for use in organized screening.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3 Describe the characteristics and frequency of neoplasia detected when screening.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.4 Determine feasibility.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.5 Preliminary assessment of costs including diagnostic workload.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing in a typical screening environment using a single prevalent screen.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Separate cohorts perform the new test or comparator (potentially in the form of “usual care”) and outcomes are followed from invitation to outcome of interest. Only test-positive cases need colonoscopy (unless direct comparison to screening colonoscopy is required). Start with initial small studies addressing simpler pathway outcomes and progress to larger programs addressing detection rates. Analyze by intention-to-screen.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 4</th>
<th>Screening program evaluation over multiple rounds</th>
<th>Impact of screening on reducing burden of neoplasia, adverse events:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>4.1 To estimate or model reductions in cancer mortality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.2 Broader benefits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.3 Accurate costs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.4 Participation with re-screening.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5 Compliance with diagnostic follow up.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.6 Treatability of lesions detected.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.7 Screening intervals.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.8 Missed cancer rate.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.9 Program detection rates with repeated screening.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.10 Diagnostic follow-up rate across all rounds.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.11 Number needed to</td>
</tr>
</tbody>
</table>
4.12 Unexpected adverse events.

<table>
<thead>
<tr>
<th>TPR – true positive rate; FPR – false positive rate; RCT – randomized controlled trial</th>
</tr>
</thead>
</table>

screen to detect a lesion.
Phase 1 – Retrospective estimation of ability to discriminate between cancer cases and normals:

The ability to distinguish between cancer and non-cancer states is essential for a test to be useful and can initially be evaluated in cases with established cancer compared to those free of neoplasia (controls). While initially guiding evaluation, the accuracy measures obtained in this way may be biased and the cases used are not necessarily representative of pre-clinical cancer, the critical target of any screening program.

Cases and Controls:

An initial indication can be obtained comparing cases with established cancer to those free of neoplasia (controls). For cases, it is helpful to have a range of different histological features and stages, meaning that all must have had diagnostic colonoscopy.

Intervention:

Design should follow that charted in Figure 1, with cases and controls performing both the new and the comparator tests, i.e. “paired-testing”. Persons developing the test sample should be blinded with respect to the subject’s status. If the tests require collection of biological samples, it needs to be ensured that the sampling process and pre-analytical conditions should be exactly the same for cases and controls (such as time interval from the colonoscopy, setting of the examination, conditions of sample storage and so on).

Outcomes and Sample Size:

A sample size of 60 pairs has approximately 80% power to detect a difference in true positivity rate of 20% when the proportion of discordant pairs is expected to be 30% in cases affected by the cancer; such conditions may be encountered. “Discordant pairs” refers to those cases positive by one or the other but not both tests. The minimum standard approach and its analysis is described in detail by Pepe et al. Basic considerations in measuring power when true- and false-positive rates are the main outcomes and the design is not paired, have been provided.

For studies on marker combinations that require training before validation, if the training and validation cases are drawn from the same population, the sample size requirements should be fulfilled by the validation set independent of the training set.

The proportions of subjects with lesions where both new and comparator tests are positive and where only one or other test is positive, should be reported. This clarifies concordance between the tests and addresses Principle 5.

To compare tests in a paired design, calculation is simply performed by determining the...
confidence interval of the difference in test positivity or by McNemar’s test. Fine-tuning the test endpoint, i.e. the threshold set for positivity (the “criterion” value), is crucial for those tests with a quantitative or semi-quantitative endpoint. An ROC curve should be constructed and analyzed. For each cut-off selected for positivity in the ROC curve, the confidence interval of the difference in positivity rates between new and comparator test can be calculated.

If the new test is at least comparable to the comparator then it is justified to proceed to Phase 2 evaluation. In exceptional circumstances, skipping phases prior to Phase 3 might be justifiable especially if screen-detected cases were included.

**Phase 2 – Detection of neoplasia across the oncogenic continuum, prospective clinical studies:**

Paired testing is undertaken prospectively in subjects before undergoing the diagnostic procedure, i.e. before they are identified as cases or controls. Test operating characteristics need to be understood across the spectrum of stages of oncogenesis, with the particular interest being performance in the earlier stages when treatment is more likely to be successful. This is especially important if the new test has a different objective (i.e. it detects a different biology) from the proven comparator. The risk in practice is that seeking a higher detection rate for early stages or pre-invasive neoplasia (adenomas) raises the possibility of a higher false positive rate and over-diagnosis (detection of inconsequential colorectal neoplasia).

There are two clinical targets of particular interest. One is a shift to earlier stage of cancer as CRC screening RCTs show reduced mortality to be linked to earlier detection. This can only be examined in very large screening studies but a surrogate measure is provided by estimating sensitivity for earlier stage cancer. The second target is that of pre-invasive neoplasia, particularly “advanced” adenomas (size >9mm, Villous component >25%, high grade dysplasia or more than 2 of any characteristic), as detection of adenomas by screening flexible sigmoidoscopy is beneficial and advanced adenomas are more likely to progress to cancer.

An important purpose of Phase 2 can be to determine the final test format, i.e. criterion endpoint fine-tuning, prior to population evaluation in Phase 3. The operational nature of the test, e.g. in the case of a laboratory test the assay details and analyte, should be carefully defined (see Principle 8) and a provisional threshold set for positivity, i.e. the characteristic that would direct that person to undergo diagnostic evaluation. For tests requiring a biological sample, the sampling process must be clear; information on stability of the analyte and robustness of the sampling method regarding pre-analytical variations should be published. If any of these matters remain uncertain, then simple pilot studies in typical screening populations should be undertaken. While a new test might detect lesions at an earlier stage it might also fail at certain stages, or it might detect
a different type of neoplastic lesion. Ideally, Phase two studies would indicate if these outcomes were likely.

**Cases and Controls:**

Patients scheduled for colonoscopy for any reason, are informative, but more so if asymptomatic.

**Intervention:**

Evaluation parallels that for Phase 1 with subjects performing paired testing prior to colonoscopy. Subjects should be classified according to stage of oncogenesis and presence or absence of neoplasia, specifically: cancer stage, advanced adenoma, non-advanced adenoma, benign pathology or normal organ.

Generalized linear modeling can be used to examine the relationship between covariates and test result\(^{37,43}\). This will highlight the factors other than pathology in the organ that must be considered in Phase 3 as potential covariates.

**Outcomes and Sample Size**

The low prevalence of cancer, even in subjects scheduled for colonoscopy, requires recruitment of many subjects. As discussed for Phase 1, a meaningful comparison may be achieved if approximately 60 of the desired target lesions are included in the study population, given paired-testing. To calculate the total population size required to provide sufficient power, one then needs to know the likely prevalence of the target lesion in the population. As a general rule, 1,000-5,000 people should be recruited depending on whether attempts to enrich the population with cancer cases are successful. Advanced adenomas are likely to be ascertained at a rate approximately three to ten times that of cancer, when evaluating screening tests for CRC.

The data provided from Phase 2 evaluation may be sufficient to have a test registered with appropriate authorities for medical use. If performance has been demonstrated to be at least equivalent to the comparator, then it is justifiable to proceed to population screening studies.

**Phase 3 – Initial screening evaluation – participation and prevalence studies:**

Phase 3 evaluation seeks to confirm that the new test improves outcomes when the test is applied in the screening context as a one-time event, i.e. a prevalent screen. Usually, separate cohorts are randomized to each test so as to provide intention-to-screen outcomes. An organized screening program starts with an offer of the test, the test sample is obtained by the person (ideally under optimal conditions) but entirely at their own discretion, the sample is submitted for analysis.
and each positive test result must be verified by a diagnostic test. \(53\). This is the minimum level of evidence required to justify use in large-scale organized screening.

The Population:
Study groups should be derived randomly from a population that would be targeted in a screening program. Unbiased selection of invitees is highly desirable.

Intervention
In randomized screening trials, participants usually perform one test only, as though this were a typical screening program. If they do both, then intention-to-screen outcomes cannot be determined. Prospective testing with either the new test or the comparator test requires that collection of the samples is undertaken prior to the ascertainment of the diagnosis. Events should be tracked from offer of screening to completion of diagnostic verification (see Principle 4), except in small studies that seek to gather information on participation as the only outcome.

Outcomes and Sample size.
Intention-to-screen analysis of results as well as per-protocol analysis should be undertaken. For per-protocol (i.e. participant) analyses, in addition to the above outcomes, the overall test positivity rate which defines the total diagnostic work load, i.e. colonoscopy, is informative. For intention-to-screen analyses, test participation rates and tracking return of tests over time is also informative.

Adjusted logistic regression analyses can be undertaken to adjust for covariates. Given that separate groups are studied in this type of design, covariates may not be equal between the groups and especially might not be equal between those undertaking testing or returning positive test results.

Sample size depends on the degree of incremental improvement being sought, the target lesion of interest, whether one is focusing on an intention-to-screen or participatory (per protocol) outcome, and the outcome being addressed. For instance, test positivity or participation rates are often the initial outcomes of interest in Phase 3 studies and are easily estimated. With study group sizes of \(n=376\), a 2-group \(\chi^2\) test with a 0.05 2-sided significance level gives 80% power to detect a 10% change in participation where participation in the reference group is 30% \(43\). Where the ultimate consideration is difference in detection rates of cancer, if one expects a difference in detection rates of cancer of 3 per thousand invitees \(17\) then sample size should be at least 6083 if one expects a gFOBT comparator to detect 2 per 1000.
Therefore, it is sensible within Phase 3 studies to progressively stage evaluation, starting with smaller study groups of say 400 to 500 to measure overall test positivity rate (which estimates the number of colonoscopies needing to be done) and participation rates and to gain further estimates of true- and false-positive rates and associated covariates. This informs sample sizes for larger studies that then address detection rates. Modeling cost-effectiveness is an important element of Phase 3 as it provides real-world estimates of test positivity rates and participation, variables that are important to accurate cost-modeling. Indeed, as outcomes are accumulated, extensive modeling can be undertaken using models such as MISCAN\textsuperscript{54}, to predict impact and so enable adjustment of programs to maximize the likely benefit.

**Phase 4 -- Screening Program Evaluation:**

Screening aims to reduce the burden of disease by reducing CRC mortality at the population level. It is important that it does not adversely affect the health status of those who choose to participate. A new test might be associated with some unexpected adverse events that would counterbalance mortality benefits predicted by better detection and/or participation; Phase 4 studies conducted over multiple rounds should identify these.

Comparing new CRC screening tests using CRC mortality as the endpoint will probably never be feasible on the grounds of size, time and cost. Phase 4 evaluation is not so much about comparison of tests but about monitoring how the new test performs when applied to a large unselected population, ideally over repeated rounds of screening. Measures such as shift to an earlier stage and interval (missed) cancers are ascertainable, as well as unexpected adverse events. Knowledge of these will improve cost-effectiveness determinations. As a consequence, Phase 4 evaluation would normally proceed as a process of careful evaluation of an organized screening program applied to a large population and monitored over a considerable time, often involving multiple rounds of screening.

**Outcome measures that demonstrate benefit**

In considering what to measure to assess health benefit in screening programs, intermediate measures associated with demonstrated RCT effectiveness can be informative\textsuperscript{55}.

The gFOBT RCTs show that a shift to an earlier stage of cancer in a program involving repeated screening offers is associated with reduced mortality\textsuperscript{7,8,10,55}. Thus earlier detection by a new test to at least a comparable degree is highly desirable. For instance, screening with FIT has now been shown to lead to earlier detection\textsuperscript{50}.

The association of adenoma detection and removal in screening with the reduction of CRC
incidence and mortality is now proven by the RCTs of flexible sigmoidoscopy (FS) screening\textsuperscript{5}. FS is thus an expeditious comparator for evaluating new tests targeting pre-invasive lesions, a potential surrogate measure for predicting reduction in incidence being detection (true-positive rates) of those lesions considered to be at high risk of progressing to CRC.

Interval cancers, i.e. missed or new cancers, occur in programs and monitoring these for each test would be valuable although to obtain valid and accurate comparative data an adequate follow-up time and a very large sample size are required. Nonetheless, interval cancer rates need to be determined especially where the earlier Phases of evaluation have focused primarily on assessment of test-positive cases (i.e. an endoscopic method is not routinely undertaken in test-negative participants).

Comparing tests over multiple rounds is also an important goal of Phase 4 testing and will require prolonged follow-up. Cumulative detection rates should be considered when the stipulated screening interval of the tests being compared is different. Also, methods for reporting participation over multiple rounds of screening have not been well applied to CRC screening\textsuperscript{56}, but, as long as repeated participation is required to achieve the expected screening benefit, this represents a relevant indicator to be assessed. Participation in screening, a central performance indicator for population screening, can vary across the population and it is important not only to monitor the effect of a new test on overall uptake, but also its acceptability to all socio-economic and ethnic groups in order to avoid widening the inequalities gap.

**Phase 4 study design:**

Studies should follow the design outlined for Phase 3 evaluation but include multiple rounds of screening (several at least with the interval matched to the perceived duration of effect of each test), with plans to ascertain the outcomes relating to those measures deemed to be important, namely participation, detection, cost, adverse effects, earlier detection and interval (new or missed) lesions.

Such studies will be extremely costly and normally only feasible in the context of public health screening strategies that are already in place, where methods to collect outcome measures are already designed and operational. In other words, Phase 3 evaluation is sufficient to lead to incorporation of a new test into a pilot within a formal organized population program and Phase 4 evaluation serves to confirm the expected promise by evaluation of screening programs. Given good information on costs, comparative cost-effectiveness of different tests can be determined as described\textsuperscript{57}. 

This article is protected by copyright. All rights reserved.
NEW BIOMARKERS:

The discovery of new biomarkers, such as fecal or blood tests for DNA, RNA or protein, adds complexity. Initial research usually precedes Phase 1\(^3\) as we describe it but also requires fine-tuning the test endpoints in Phases 1 and 2. This is especially true if a panel of markers is being used.

The process of discovery starting with tissue banks has been discussed in detail elsewhere\(^3,5,8\). Sophisticated retrospective molecular analyses of material in bio-specimen banks can serve to identify candidate biomarkers that might become the objective of the screening test.

If such laboratory research identifies a promising biomarker, it can be initially evaluated as for Phase 1 and 2 by a simple study in cases and-controls. Doing this, however, may assume that the retrospective bio-specimen banks are adequate to identify the best candidate. Usually this is not the case as discovery is often undertaken on limited numbers of samples obtained from strictly categorized materials that are often not typical of screen-detected lesions. A further technological challenge arises if resected tissue specimens are used to identify the biomarker but use of the biomarker in screening involves measurement in a biological sample such as blood or feces. Many factors might influence the appearance of the biomarker in the biological sample and there is a chance that it might not be of the same molecular structure in blood or feces as in tissue, as degradation or other processing might occur.

This makes it likely that the best discovery process first develops a putative panel of markers and then uses clinical studies set up in such a way that the panel can be explored in clinical specimens as part of Phase 1 or 2, or perhaps even Phase 3 studies. Indeed, access to the appropriately characterized population with biological samples that themselves serve as a source of materials for discovery of potential biomarkers, may be very useful. The usefulness of panels of multiple markers can then be explored, i.e. “validated”, in Phase 1-3 studies\(^5,8\).

DISCUSSION

This phased approach provides an efficient method for evaluating a new screening test that increases in cost and complexity only if key attributes are worthwhile. It assesses both accuracy and acceptance because screening of a general population requires good participation as well as good detection and the same principles can be applied to adenoma detection.

Study costs increase considerably with each Phase. The high cost of undertaking Phase 3 studies might be reduced by obtaining government regulatory approval for the use of a test on the basis of Phase 2 studies. Some authors suggest this can wait until Phase 4 studies have been undertaken\(^3\) but this seems impractical as no commercial entity would proceed with test development under such circumstances. Utilizing logistics and infrastructure of existing screening
programs can also help reduce costs of such studies.

Expensive studies have included evaluation of new non-invasive tests in people participating in colonoscopic screening\textsuperscript{59}. While useful, this fails to provide comparison with a test known to reduce mortality on an intention-to-screen basis.

The final issue is what justifies progression from one phase to the next. While our proposal sets the principles for the phased evaluation of new tests, researchers, in collaboration with health service providers, should agree on hurdle values before embarking upon a study. Importantly, criteria for equivalence or superiority should be agreed at commencement. Phase 1 studies can be considered as exploratory and of value in helping to determine necessary power and likely outcomes in Phases 2 and 3. What constitutes an acceptable hurdle value will vary with the test and how the test will be used within the health care system.

We consider that this process of comparative, phased evaluation provides a rational, efficient and useful process for evaluating new tests and for progressing a test to a stage where the considerable degree of evidence needed for it to be included in population screening is obtained. Health providers will be able to adopt a test that is soundly based on scientific objectivity and the fundamental principles of screening.

**Supplementary Material**

Note: To access the supplementary material accompanying this article, visit ??? at http://????.
FIGURE LEGENDS

Figure 1: Conceptualization of design for testing a new test relative to an existing (comparator) test. Solid lines represent essential paths in the process. Dotted lines represent discretionary paths not essential in some phases of evaluation.
ACKNOWLEDGEMENTS

We are grateful to the secretariat of the World Endoscopy Organization and to Alexandra Whibley, Erin Symonds and Emma-May Palmer for their assistance in the consensus processes and manuscript preparation.

Members of the original working party responsible for the consensus process underlying these recommendations were convened by the World Endoscopy Organization Colorectal Cancer Screening Committee and were: Young GP, Mandel J, Sung JJY, Allison JE, Atkin W, Benamouzig R, Hoff G, Itzkowitz SH, Levin TR, McFarlane EG, O’Morain C, Parry S, Rabeneck L, Rozen P, Saito H, Schoen RE, Senore C, Steele RJC, Winawer SJ, Wong BCY

The listed authors were members of the WEO/WGO New Screening Tests Expert Working Party, and met the criteria for authorship of this manuscript.

This paper is dedicated to the memory of Professor Paul Rozen, a true pioneer in CRC screening, who passed away in early 2013.
REFERENCES


Figure 1: Conceptualization of design for testing a new test relative to an existing (comparator) test. Solid lines represent essential paths in the process. Dotted lines represent discretionary paths not essential in some phases of evaluation.

Further endpoints in screening pathway (see “Phased evaluation”)
Supplementary Table 1: The guiding principles that underpin a strategy for comparing screening tests which emerged from the consensus approach and the literature review, the key points giving rise to each one, together with their key consequences for comparing tests.

1. Screening aims to reduce the burden of disease in the targeted population, without adversely affecting the health status of those who participate in screening, through early detection and treatment of cancer and/or through detection of pre-cancer lesions, which reduces incidence. Criteria that justify screening were defined more than four decades ago. Balancing benefit with harm in otherwise healthy individuals is crucial, which means that considering screening only as an exercise in test-based detection may fail to provide information as to the overall benefit. Published randomized controlled trials (RCTs) in CRC screening show benefit to mortality through early detection of cancer and to incidence through detection of adenomas.

2. The screening test is just one event in a process that includes engagement of the public, testing, validation, communication and treatment. The International Agency for Research on Cancer (IARC) has emphasized that organized screening programs, as opposed to opportunistic ad hoc screening, provide greater protection against many of the harms of screening, including over-screening, poor quality and complications of screening, and poor follow-up of those who test positive. Consequently, the screening test is just one event in a process that includes engagement, testing, validation, communication and treatment. The individual as an entity and how he or she benefits, rather than what the test detects, is the focus of screening. The nature of a test has an impact upon participation, communication and cost amongst other program elements so screening is much more than simply an exercise in lesion detection.

3. Population randomized controlled trials (RCTs) with mortality as primary outcome set the standard for evaluation of new tests. Two-step screening using a guaiac-based fecal occult blood test (gFOBT) as the initial test has been shown to be effective in reducing CRC-specific mortality in multiple controlled trials analyzed on an intention-to-screen basis. The extent of the evidence has been reviewed; see also Table 1 in the main manuscript. These trials have pointed to those intermediate endpoints that predict potential for reducing mortality since those endpoints can be directly related to the main outcome of interest. The relatively insensitive gFOBT Hemoccult offered biennially reduced CRC-mortality by 15-20%.

This improves to a 33% reduction with rehydration of Hemoccult offered annually, a
process that increases sensitivity, but results in considerable deterioration in specificity. The increased sensitivity achieved with rehydrated Hemoccult is also associated with a 20% reduction in CRC incidence when followed up for 18 years, presumably resulting from increased detection and removal of adenomas. Together with this benefit on CRC mortality, the associated parameters regarding screening participation, test accuracy and cancer stage are well known.

More recently, the results of four sigmoidoscopy screening RCTs consistently showed that endoscopic excision of colorectal adenomas is associated with a substantial reduction in CRC incidence (18%-23%) and mortality (26%-31%) on an intention-to-screen basis. Considering subjects who were actually screened, the reduction in CRC incidence ranged between 31% and 33% and CRC-specific mortality was reduced by 38% - 43%. The observed protective effect refers to a follow-up of 11 years and was mainly limited to the distal colon. The reduction in CRC incidence in the proximal colon was small and not statistically significant either in the UK (3%) or in the Italian (15%) trials. A statistically significant 14% reduction in CRC incidence in the proximal colon was documented only in the PLCO trial, but this was not associated with a mortality reduction.

Based on the effect observed for flexible sigmoidoscopy it can be concluded that structural detection (i.e. visualization at endoscopy) of lesions brings significant benefit in terms of reduced incidence, morbidity and CRC mortality. But it should be noted that the majority of adenomas would not progress to cancer during a person’s life time if left in situ.

Whether the benefit of polypectomy extends to the proximal colon is not yet certain. This uncertainty is underscored in observational studies that showed that the use of colonoscopy was not associated with a reduction in the risk of dying from right-sided CRC. One case-control study has shown a reduction in proximal CRC incidence associated with self-reported use of colonoscopy in the preceding 1 to 10 years, and only in subjects older than 60 years. These findings underscore how crucial the quality of diagnostic examinations is to maximizing effectiveness in screening and to optimize the balance between potential harms and benefit. They also suggest that effectiveness of one-step colonoscopic screening in practice might not be as great as is often assumed.

Therefore, even though colonoscopy improves detection of both invasive lesions and pre-invasive lesions (adenomas), adding the potential to prevent cancer, the benefit of colonoscopic screening, either in terms of CRC mortality, or incidence reduction, has not been assessed by mass population RCTs in the setting of mass population screening. Such studies are, however, underway.
4. New tests can be assessed in parallel to an existing test all the way through the screening process from population engagement to population outcomes/Measures.

The impact of the new test on an important screening outcome, such as CRC mortality, can be predicted from a direct comparison of the new with a proven test using intermediate/surrogate endpoints such as test accuracy, true-positive rate and change in stage distribution, as long as the association of these outcomes to the expected health impact of screening has been previously documented. Also, measuring participation rates achieved by the different tests, together with their detection rates for the target lesion, allows estimation of the impact of the new test on the disease burden at the population level on an intention-to-screen basis. As mortality effects are dependent on multiple elements within a screening program, tests need to be compared at all of those points where the test itself might have an effect.

5. New screening tests might detect a different neoplasia-dependent biology, and as a consequence, the value of treatment and benefit to mortality reduction might not be the same.

It is beyond the scope of this paper to comprehensively describe all the existing and potential tests available at their varying stages of development and evaluation. Table 1 in the main manuscript provides an overview for the main categories of screening tests for CRC regarding the type of evidence supporting their value and the target of detection (the test objective).

If the objective of a new test reflects a biology that is not necessarily shared by existing tests, it is possible that what is detected might have a different probability of progressing to (advanced) cancer and/or might respond differently to treatment – sometimes referred to as a shift in treatability of a disease. An example is comparison of a fecal DNA test with an FOBT. Here, test objectives are different and the biological behavior associated with bleeding might be different from that associated with DNA characteristics of cells shed into the lumen. It has been suggested that in this situation, it might be necessary to characterize the lesions detected and to make evidence-based judgments about the generalizability between the cases detected by the new test versus those detected by the proven test. However, given that structural detection of cancers and premalignant adenomas by flexible sigmoidoscopy leads to reduction in CRC mortality, it is unlikely that a shift in treatability will prove to be a real issue of concern, at least when considering distal lesions. Furthermore, concordance between the new test with its different objective and the comparator test will identify whether the lesions being detected show overlap between the different methods.
Measuring test concordance and examining for differences between discordant cases will be important in comparative studies.25

6. In two-step screening, the screening test selects participants who proceed to diagnostic verification by colonoscopy as a positive test increases the likelihood of neoplasia being present.

“Two-step screening” is where an initial test (the “screening test”) is used to identify a subpopulation more likely to have neoplasia as indicated by a positive test.13,26 If we only consider a test’s ability to detect neoplasia at a given point in time, no simple test for use in two-step screening will be as accurate as one-step screening with colonoscopy. Colonoscopic validation of test-positive individuals is essential as part of the process of evaluating test accuracy, although colonoscopy is not an infallible standard. Colonoscopy serves dual purposes: as the best means of diagnostic verification of a positive test in two-step screening for CRC and as a therapeutic procedure for certain lesions.

The change in likelihood for a disease when a screening test is positive27,28 can be simply calculated as sensitivity divided by 1-specificity. In the gFOBT RCTs, the likelihood of finding a cancer given a positive test was 8- to 25-fold times greater13.

7. It is not ethically justifiable to proceed to study a test in the screening environment including acceptability to invitees or other screening program outcomes, without studies that indicate that the new test is of acceptable accuracy when compared to a proven comparator test.

Before exposing a population to a new test, evidence must be provided that the test will perform as required and expected, both from the clinical and laboratory perspectives. The consensus view of the authors was that the accuracy of a new test should be at least equivalent to the minimum standard set by a proven test such as a gFOBT.

Given that the cost and resources required for a definitive evaluation of a new test at the population level are substantial, preliminary studies are essential. Reasoned consideration of what is required gives rise to the concept of a phased or sequential evaluation starting with relative accuracy in moderate-scale studies followed by screening pathway evaluation requiring larger populations. Each phase is more stringent than the preceding one but the process is not conceived as being so rigid that phases cannot be modified, or even omitted, as circumstances require or allow.
8. New tests must be clearly defined with provision of adequate technical details, quality assurance procedures and performance standards.

It is not the purpose of this paper to discuss quality matters but it must be emphasized that tests should be performed according to professional standards or manufacturer’s instructions, following appropriate training and in accordance with professionally-defined quality control and assessment procedures and with good laboratory and clinical practice. Test procedures should be reported in a manner that enables comparison of the results with those from other studies. Clinical evaluation has little value if the test is not reliable or robust and cannot perform consistently well in a typical screening setting.

These critical issues can be easily illustrated with fecal immunochemical tests (FIT) for hemoglobin\(^30\). Before an FIT can be recommended for evaluation in population screening or even regulatory approval, it needs calibration using internationally recognized standards, well-characterized reproducibility data for within and between laboratory analysis, and linearity of response across the clinically relevant concentration range (if a quantitative test). Even the design of a FOBT, its packaging, instructions for use and its mode of distribution need to be considered since they are critical to reaching the target population and achieving acceptable test performance. While some of these issues might be refined with experience, the fundamental analytical issues must be identified and addressed at the outset.
References


**Supplementary Table 2:** The four main phases of screening test evaluation with selected studies that demonstrate elements of a Phase, together with their main characteristics. There are many other possible examples and these are provided to demonstrate the increasing complexity of each Phase, what the design options are within a Phase and what information might be gleaned from such studies.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Study</th>
<th>Test objective</th>
<th>Population</th>
<th>Comparator test</th>
<th>Intention-to-screen?</th>
<th>Colonoscopy</th>
<th>Outcomes ascertained</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Prescreening: Retrospective estimation of ability to discriminate between cancer cases and controls without neoplasia.</td>
<td>Ahlquist D et al 2000¹</td>
<td>Multiple DNA markers; feces</td>
<td>Cancer cases and normal controls etc., n=61</td>
<td>No marker for comparison</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false)</td>
</tr>
<tr>
<td></td>
<td>Young et al 2003²</td>
<td>Globin; feces</td>
<td>Cases with known cancer or adenomas, and controls, n=300.</td>
<td>Single group, paired (2 FIT)</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false); relativity between FIT; Test preference</td>
</tr>
<tr>
<td></td>
<td>Lofton-Day et al 2008³</td>
<td>Methylated SEPT9; plasma</td>
<td>Cases with known cancer and controls, n=312.</td>
<td>No marker for comparison</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false)</td>
</tr>
<tr>
<td></td>
<td>Melotte et al, 2009⁴</td>
<td>Methylated NDRG4; feces</td>
<td>Cases with known cancer or advanced adenomas, and controls, n=150</td>
<td>No marker for comparison</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false)</td>
</tr>
<tr>
<td></td>
<td>Ruffin et al 2010⁵</td>
<td>D-galactose-β-[1,3]-N-acetyl-D-galactosamine; feces</td>
<td>Cancer cases and normal controls etc., n=299</td>
<td>Single group, paired (FOBT)</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false); relativity to FOBT</td>
</tr>
<tr>
<td></td>
<td>Bosch et al 2012⁶</td>
<td>Methylated PHACTR3; feces</td>
<td>Cases with known cancer or advanced adenomas, and controls, n=92</td>
<td>Single group, paired (FIT)</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false); combination of marker with FIT</td>
</tr>
<tr>
<td></td>
<td>Pedersen et al 2014⁷</td>
<td>Methylated CAHM, plasma</td>
<td>Cancer cases and normal controls, n=220</td>
<td>No marker for comparison</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false)</td>
</tr>
</tbody>
</table>
### Supplementary Table 2: (Continued)

<table>
<thead>
<tr>
<th>Study</th>
<th>Methodology</th>
<th>Patients</th>
<th>Controls</th>
<th>Design</th>
<th>Follow-up</th>
<th>Positivity Rate</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levin, Hess et al. 1997&lt;sup&gt;8&lt;/sup&gt;</td>
<td>Heme; feces</td>
<td>Prospective cohort of screening subjects (n=8293)</td>
<td>Single group, paired (Hemoccult in 3 versions)</td>
<td>No</td>
<td>Only if positive</td>
<td>Population positivity rate and PPV</td>
<td></td>
</tr>
<tr>
<td>Imperiale T et al 2004&lt;sup&gt;9&lt;/sup&gt;</td>
<td>Multiple DNA markers; feces</td>
<td>Cases with broad range of neoplastic lesions and non-neoplastic controls; n=5486</td>
<td>Single group, paired (DNA markers and gFOBT)</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false); relativity between tests</td>
<td></td>
</tr>
<tr>
<td>Smith et al 2006&lt;sup&gt;10&lt;/sup&gt;</td>
<td>Globin; feces</td>
<td>Cases with broad range of neoplastic lesions and non-neoplastic controls; n=2351</td>
<td>Single group, paired (FIT and gFOBT)</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false); relativity between tests</td>
<td></td>
</tr>
<tr>
<td>Park, Ryu et al. 2010&lt;sup&gt;11&lt;/sup&gt;</td>
<td>Heme, globin; feces</td>
<td>Prospective cohort of screening subjects (N=770)</td>
<td>Single group, paired (gFOBT or quantitative FIT at different cut-offs)</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false); relativity between tests and between cut-offs</td>
<td></td>
</tr>
<tr>
<td>Johnson et al. 2014&lt;sup&gt;12&lt;/sup&gt;</td>
<td>Methylated SEPT9; plasma</td>
<td>Cases with CRC (N=102) and prospective cohort of screening subjects (n=199)</td>
<td>Single group, paired (FIT and SEPT9)</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false); relativity between tests</td>
<td></td>
</tr>
<tr>
<td>Pedersen et al 2013&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Methylated IKZF1 and BCAT1</td>
<td>Prospective - patients having colonoscopy: CRC (N=129) and neoplasia-free patients (n=838)</td>
<td>No marker for comparison</td>
<td>No</td>
<td>All</td>
<td>Positivity rate in cases (true) and controls (false)</td>
<td></td>
</tr>
</tbody>
</table>
### Supplementary Table 2: (Continued).

<table>
<thead>
<tr>
<th>3. Initial screening evaluation; single round of screening</th>
<th>Allison, Tekawa et al. 1996</th>
<th>Heme, globin; feces</th>
<th>Screening population, N=8065</th>
<th>Single group, paired two gFOBT vs a FIT</th>
<th>No</th>
<th>Only positives; but prolonged follow-up</th>
<th>True and false-positive rates with imputed sensitivity and specificity.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cole et al 2003</td>
<td>Heme, globin; feces</td>
<td>Screening population, N=1800</td>
<td>Parallel cohorts; randomized (gFOBT or one of two FIT)</td>
<td>Yes</td>
<td>Only positive</td>
<td>Participation rate.</td>
<td></td>
</tr>
<tr>
<td>Segnan N et al. 2005</td>
<td>Structural lesion, Globin in feces</td>
<td>Screening population, N=26682</td>
<td>Parallel cohorts; randomized (FIT or flexible sigmoidoscopy)</td>
<td>Yes</td>
<td>Only positive</td>
<td>Participation and population detection rate; Relativity between tests</td>
<td></td>
</tr>
<tr>
<td>Allison et al 2007</td>
<td>Heme, globin; feces</td>
<td>Screening population, N=5841</td>
<td>Single group, paired a gFOBT vs a FIT</td>
<td>No</td>
<td>Only positives; but prolonged follow-up</td>
<td>True and false-positive rates with imputed sensitivity and specificity.</td>
<td></td>
</tr>
<tr>
<td>van Rossum, van Rijn et al. 2008</td>
<td>Heme, globin; feces</td>
<td>Screening population, n&gt;20,000</td>
<td>Parallel cohorts; randomized (FIT or gFOBT)</td>
<td>Yes</td>
<td>Only positive</td>
<td>Participation and population detection rate;</td>
<td></td>
</tr>
<tr>
<td>van Roon, Wilschut et al. 2011</td>
<td>Globin; feces</td>
<td>Screening population, N&gt;8100</td>
<td>Parallel cohorts; randomized(one or two sample FIT)</td>
<td>Yes</td>
<td>Only positive</td>
<td>Participation and population detection rate; Relativity between tests</td>
<td></td>
</tr>
<tr>
<td>Regge et al 2014</td>
<td>Structural lesion</td>
<td>Screening population, N=7000</td>
<td>Parallel cohorts; randomized (flexible sigmoidoscopy or CT)</td>
<td>Yes</td>
<td>Only positive</td>
<td>Participation and population detection rate; Relativity between tests</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Test Details</td>
<td>Study Details</td>
<td>Comparison</td>
<td>Results</td>
<td>Conclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
<td>---------------</td>
<td>------------</td>
<td>---------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imperiale T et al, 2014</td>
<td>Multitarget DNA and globin; feces</td>
<td>Screening cases; n=9989</td>
<td>Single group, paired (multitarget test vs FIT)</td>
<td>No</td>
<td>Sensitivity and specificity etc; Relativity to FIT.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Supplementary Table 2**: (Continued).

<table>
<thead>
<tr>
<th>4. Screening program evaluation, generally over multiple rounds</th>
<th>Malila et al 2011(^{22})</th>
<th>Heme; feces</th>
<th>Screening population (N=74592)</th>
<th>No comparator test.</th>
<th>Yes</th>
<th>Only positive cases</th>
<th>Participation rate by round; Positivity rate by round, PPV by round.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quintero et al 2012(^{23})</td>
<td>Globin; feces</td>
<td>Screening population (N=&gt;53000)</td>
<td>Parallel cohorts; randomized. Single round colonoscopy or 6 rounds of FIT</td>
<td>Yes</td>
<td>Only positive cases (in FIT arm).</td>
<td>Participation and population detection rate. Relative workloads.</td>
<td></td>
</tr>
<tr>
<td>Sali et al 2013(^{24})</td>
<td>globin in feces, structural lesion</td>
<td>Screening population (N=14000)</td>
<td>Parallel cohorts; randomized. Single round colonoscopy or CT colonography or 3 rounds of FIT</td>
<td>Yes</td>
<td>Only positive cases (in FIT or colonography arms).</td>
<td>Participation and population detection rate. Relativity between tests</td>
<td></td>
</tr>
<tr>
<td>Cole et al 2013(^{25})</td>
<td>Globin; feces</td>
<td>Cancer cases in general population (N=3481)</td>
<td>Invitees compared to those not-invited; national screening program.</td>
<td>Yes</td>
<td>If positive or considered clinically indicated.</td>
<td>Cancer stage distribution</td>
<td></td>
</tr>
<tr>
<td>Van Roon AH et al.(^{26})</td>
<td>Globin; feces</td>
<td>Screening population (N=7501)</td>
<td>No comparator test.</td>
<td>Yes</td>
<td>Only positive cases</td>
<td>Participation rate by round; Positivity rate by round, PPV by round.</td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations**: FOBT, fecal occult blood test; gFOBT, guaiac-based fecal occult blood test; FIT, fecal immunochemical test for hemoglobin; CRC, colorectal cancer; RCT, randomized controlled trial.
References


