TABLE 2. Relation Between Direct Practical Measures (Operating Characteristics) of a Screening Test Result, How Each Informs Assessment of Test Accuracy, and the Consequences of the Result for a Screening Program

Test Result	Diagnostic Verification; Operating Characteristic	Corresponding Accuracy Characteristic	Issue Addressed
Positive	True (ie, target condition present); true-positive rate (TPR)[a]	Sensitivity (positivity rate in those with the target condition)	Detection
		Positive predictive value (TPR/TPR + FPR)	Efficiency of detection
	False (ie, target condition not present); false-positive rate (FPR)[a]	Specificity (1 − FPR)	Burden associated with detection
Negative	True; true-negative rate (TNR)	Negative predictive value (TNR/TNR + FNR)	Elimination/exclusion of targeted clinical lesion (stage specified)
	False; false-negative rate (FNR)	Missed lesion	Burden of failed detection

[a] A targeted clinical lesion is either cancer and/or advanced adenoma, depending on the question being asked of the test, because tests might detect these to differing degrees.

screening.[1] 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 2 key measures of accuracy—sensitivity and specificity—are often difficult to ascertain, especially for screen-relevant lesions (ie, the earlier stage cancers and adenomas that would be encountered 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 that included a sufficient number of participants with such lesions in which all test confounders were likely to be encountered and in which every participant, 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,[21] that existing tests, namely gFOBT/FIT and FS, have demonstrated effectiveness and can be used to facilitate assessment of relative benefit.

Another simplification is based on the proposition that the 2 key questions concerning clinical accuracy[3,34,35] 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 2 parameters is achieved by a thorough diagnostic verification of every test-positive case (both comparator and new test-positives) to determine whether it is a true-positive or a false-positive.*[3,36]

The simple dichotomous measures of the true-positive rate (TPR) and the false-positive rate (FPR) are direct and practical measures of accuracy, sometimes referred to as test "operating characteristics," as indicated in Table 2. They are used when undertaking receiver operating characteristic (ROC) analysis. The TPR reflects detection (sensitivity), and the FPR reflects the burden associated with detection (1-specificity). Consequently, relative sensitivity and specificity are determined by comparing the TPR and the FPR, respectively, between tests.

Comparing Test Accuracy: The Scenarios

The approach based on verification of positive tests, classifying them as true-positive or false-positive, provides a straightforward but powerful strategy for comparing the accuracy (operating characteristics) of 2 screening tests. The concepts presented apply regardless of whether the target lesion is cancer and/or adenoma.

In comparing accuracy, the targeted clinical lesion (hereafter referred to as *targeted lesion*), which can be cancer, and/or adenoma, or combinations thereof, needs to be clearly defined. 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) (Supporting Table 1; see online supporting information).

Two simple questions, modified from Lord et al,[1] guide assessment in a practical manner:

Is the new test better at detecting target lesions?

This is true if the TPR (which reflects sensitivity) for the target lesion 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 returns more false-positives (lower specificity) than the old test, raising concerns about cost and potential harms. Hemoccult Sensa, compared with Hemoccult II, is an example.[37,38] Note, however, that a test with more true-positives and a higher initial colonoscopy rate (whether because of true-positives 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 the cost per quality-adjusted life year saved.

There are several ways to address such complex scenarios. The operating characteristics of the 2 tests can be plotted as an ROC curve (TPR vs FPR) as a way to judge which test has the best balance of true-positives and false-positives; overall, the test with the greatest area under the curve has the best discriminatory power.[39] This is particularly applicable to prescreening phases in the evaluation process that focusses on accuracy (see below).

Another objective approach is to calculate the number needed to screen (colonoscope) to detect 1 target lesion using each test (the reciprocal of the positive predictive value). Calculating the number needed to colonoscope also facilitates comparison of 2 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 because of recruitment processes.

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

A new test might have other benefits, for instance, significantly better specificity without improved sensitivity. Comparison is made simple in this circumstance by calculating the number needed to colonoscope to detect 1 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 1 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 demonstrated a statistically significant but only relatively small impact on CRC mortality, to be inadequate. Consequently, 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 Fig. 1). Which is chosen depends on the stage of evaluation (see Box 2). The subsequent discussion on phased evaluation provides more detail.

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, eg, safety, cost, feasibility, ease of use for a screening participant, and acceptability. New tests must undergo evaluation in unselected, typical screening populations, and an 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.[41] Consequently, measuring participation with 1 test relative to another in separate cohorts randomly selected from the same population can document test acceptability,[42] provided that framing of information is carefully balanced.

PHASED EVALUATION

Phased (ie, sequential) evaluation in a step-wise, increasingly complex manner is most appropriate.[3,20,43-46] 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

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 (ie, each does both the new test and the 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 results. Pairing reduces cohort size because of improved statistical power for assessing incremental benefit. It ensures that individuals are comparable and avoids imbalances in variables that affect test results and in other biases between the tests. If the new test demonstrates promise, then larger numbers of individuals undertaking paired testing can be further studied with colonoscopic follow-up in test-positive individuals 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 demonstrated 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 1 setting to another is not possible.[40]

BOX 3: The 4 Phases of Test Evaluation and Associated Issues

Phase 1. Retrospective estimation of ability to discriminate between cancer cases and normal;
Phase 2. Detection of presymptomatic stages along the neoplastic continuum, prospective clinical studies;
Phase 3. Initial screening evaluation—participation and prevalence studies; and
Phase 4. Screening program evaluation.
Issues to be noted:

- In 2-step screening, screening tests select participants[21] who then undergo the reference diagnostic test.
- Pathway parameters in screening, such as participation rates, are as crucial to population benefit as accuracy.[41]
- Relative test accuracy is simply addressed in a paired design.[3]
- 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.[21]
- The specific phases of screening are a guide to evaluation reflecting a continuum from simple to increasingly complex evaluations in which each step may be adjusted for complexity according to outcomes in the previous phases.[21]
- The cost for each phase is subject to local considerations; however, if the costs of diagnostic verification are put aside, then Phase 1 studies might cost several hundred thousand dollars, whereas Phases 3 and 4 will cost several to many millions of dollars.

thorough evaluation addressing outcomes in the population screening context (Phases 3 and 4), as indicated in Box 3. Phased evaluation takes into account the issues described in Box 3. The primary and secondary objectives and general characteristics of these phases are provided in Table 3.

The phased approach is indeed undertaken in practice. Supporting Table 2 (see online supporting information) provides selected examples of studies that have demonstrated the elements of each phase, together with their main characteristics. When tracking through these phases for tests, such as different FIT products and designs or the fecal DNA tests, it is observed 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, design options within a phase, and information that may be gleaned from such studies.

Phase 1: Retrospective Estimation of Ability to Discriminate Between Cancer Cases and Normal Controls

The ability to distinguish between cancer and noncancer states is essential for a test to be useful and can initially be evaluated in individuals who have established cancer

TABLE 3. Phased Evaluation for Comparison of Screening Tests for Colorectal Cancer[a]

Evaluation	Nature	Primary Aim	Secondary Aims	Population
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	Individuals known to have cancer, ideally with a majority in potentially curable disease stages and including some who are asymptomatic; controls to be free of neoplasia; concordance between tests should be reported; ideally, paired testing, with 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 preclinical disease, including advanced adenomas 2.2 To determine the final format of the test (sample and endpoint) Minimum requirement for test registration	2.3 More reliably estimate operating characteristics 2.4 Information on covariates affecting test performance 2.5 Ascertain the number of samples and threshold (fine-tune the endpoint) 2.6 Test to be registerable with authorities 2.7 Clarify whether there are subgroups in which 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 individuals with common benign diseases and how they affect test result need ascertainment; testing undertaken before scheduled diagnostic procedure; ideally, paired testing; concordance between tests should be reported
Phase 3	Initial screening evaluation; single round of screening	• Characteristics of neoplasia detected when screening; false-referral rate; acceptability 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. 3.2 Determine test acceptability Minimum requirement for use in organized screening	3.3 Describe the characteristics and frequency of neoplasia detected when screening 3.4 Determine feasibility 3.5 Preliminary assessment of costs including diagnostic workload	Testing in a typical screening environment using a single prevalent screen; 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 those who test positive need colonoscopy (unless direct comparison with 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
Phase 4	Screening program evaluation over multiple rounds	• Impact of screening on reducing burden of neoplasia, adverse events 4.1 To estimate or model reductions in cancer mortality	4.2 Broader benefits 4.3 Accurate costs 4.4 Participation with rescreening 4.5 Compliance with diagnostic follow-up 4.6 Treatability of lesions detected 4.7 Screening intervals 4.8 Missed cancer rate 4.9 Program detection rates with repeated screening 4.10 Diagnostic follow-up rate across all rounds 4.11 Number needed to screen to detect a lesion 4.12 Unexpected adverse events	Randomly selected from populations in which screening program is likely to be implemented; design may use historic controls or else a parallel-arm RCT with screening participants and alternatively screened population; intention-to-screen analysis required

Abbreviations: FPR, false-positive rate; RCT, randomized controlled trial; TPR, true-positive rate.
[a] Discussions of group sizes and approximate costs for each phase are included in the text.

(cases) compared with those who are free of neoplasia (controls). Although they initially guide evaluation, the accuracy measures obtained in this way may be biased, and the cases used are not necessarily representative of preclinical cancer, the critical target of any screening program.

Cases and controls

An initial indication can be obtained comparing individuals who have established cancer (cases) with those who are free of neoplasia (controls). For cases, it is helpful to have a range of different histologic 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 tests and the comparator tests: ie, "paired-testing." The individuals who are developing the test sample should be blinded with respect to participants' status. If the test requires collection of biologic samples, then it needs to be ensured that the sampling process and preanalytic conditions are 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 the TPR of 20% when the proportion of discordant pairs is expected to be 30% in cases affected by the cancer; such conditions may be encountered.[47] "Discordant pairs" refers to those cases who are positive on 1 or the other test but not on both tests. The minimum standard approach and its analysis is described in detail by Pepe et al.[3] Basic considerations in measuring power when the TPR and the FPR are the main outcomes and when the design is not paired have been provided.[3]

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

The proportions of individuals with lesions in which both the new test and the comparator test are positive and in which only 1 or the 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[8] or by using the McNemar test. Fine-tuning the test endpoint, ie, the threshold set for positivity (the *criterion* value), is crucial for those tests that have a quantitative or semiquantitative endpoint. An ROC curve should be constructed and analyzed.[3,39] For each cutoff selected for positivity in the ROC curve, the confidence interval of the difference in positivity rates between the new test and the comparator test can be calculated.[47]

If the new test is at least comparable to the comparator test, then it is justified to proceed to a Phase 2 evaluation. In exceptional circumstances, skipping phases before 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 participants before they undergo the diagnostic procedure: ie, 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 (ie, it detects a different biology) than that of the proven comparator. The risk in practice is that seeking a higher detection rate for early stages or preinvasive neoplasia (adenomas) raises the possibility of a higher FPR and overdiagnosis (detection of inconsequential colorectal neoplasia).[48]

There are 2 clinical targets of particular interest. One is a shift to earlier stage cancer, because CRC screening RCTs demonstrate that reduced mortality is linked to earlier detection. This can only be examined in very large screening studies,[49] but a surrogate measure is provided by estimating sensitivity for earlier stage cancer. The second target is that of preinvasive neoplasia, particularly *advanced* adenomas (size >9 mm, villous component >25%, high-grade dysplasia, or >2 of any characteristic), because the detection of adenomas by screening FS is beneficial,[5,50,51] and advanced adenomas are more likely to progress to cancer.

An important purpose of Phase 2 can be to determine the final test format (ie, criterion endpoint fine-tuning), before the population evaluation in Phase 3. The operational nature of the test (eg, in the case of a laboratory test, the assay details and analyte) should be carefully defined (see Principle 8), and a provisional threshold should be set for positivity: ie, the characteristic that would direct that individual to undergo diagnostic evaluation. For tests requiring a biologic sample, the sampling process must be clear; information on stability of the

analyte and robustness of the sampling method regarding preanalytic variations should be published. If any of these matters remain uncertain, then simple pilot studies in typical screening populations should be undertaken. Although a new test might detect lesions at an earlier stage, it also might fail at certain stages, or it might detect a different type of neoplastic lesion. Ideally, Phase 2 studies would indicate whether these outcomes are likely.

Cases and controls

Individuals who are scheduled for colonoscopy for any reason are informative, but they are more so if asymptomatic.

Intervention

Evaluation parallels that for Phase 1, with individuals undergoing paired testing before colonoscopy. Participants should be classified according to stage of oncogenesis and presence or absence of neoplasia, specifically: cancer stage, advanced adenoma, nonadvanced adenoma, benign pathology, or normal organ.

Generalized linear modeling can be used to examine the relation between covariates and test results.[36,42] 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 individuals who are scheduled for colonoscopy, requires the recruitment of many participants. A meaningful comparison may be achieved if approximately 60 of the desired target lesions are included in the study population given paired-testing, as discussed for Phase 1. To calculate the total population size required to provide sufficient power, the likely prevalence of the target lesion in the population must be known. From 1000 to 5000 individuals should be recruited as a general rule, depending on whether attempts to enrich the population with cancer cases are successful. Advanced adenomas are likely to be ascertained at a rate approximately 3 to 10 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 that of 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 1-time event: ie, a prevalent screen. Usually, separate cohorts are randomized to each test to provide intention-to-screen outcomes. An organized screening program starts with an offer of the test, the test sample is obtained by the participant (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.[52] *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 1 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 sample collection is undertaken before ascertainment of the diagnosis. Events should be tracked from the offer of screening to the 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

Both an intention-to-screen analysis of results and a per-protocol analysis should be undertaken. For per-protocol (ie, participant) analyses, in addition to the outcomes discussed above, the overall test positivity rate, which defines the total diagnostic workload (ie, colonoscopy), is informative. For intention-to-screen analyses, test participation rates and tracking the return of tests over time are also informative.

Adjusted logistic regression analyses can be undertaken to adjust for covariates.[36,42] Because separate groups are studied in this type of design, covariates may not be equal between the groups, and they 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 the focus is 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-square test with a .05 two-sided significance level has 80% power to detect a 10% change in participation, where participation in the reference group is 30%.[42] When the ultimate consideration is the difference in detection rates of cancer, if a difference in detection rates of cancer of 3 per 1000 invitees is expected,[7] then the sample size should be at least 6083 if a gFOBT comparator is expected 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 the overall test positivity rate (which estimates the number of colonoscopies required to be) and participation rates and to gain further estimates of the TPR and FPR 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, because 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 like MISCAN (Microsimulation Screening Analysis)[53] to predict impact and thus enable the adjustment of programs to maximize the likely benefit.

Phase 4: Screening Program Evaluation

The objective of screening is 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 events.

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 the 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 like a shift to an earlier disease stage and interval (missed) cancers are ascertainable, as well as unexpected adverse events. Knowledge of these will improve cost-effectiveness determinations. Consequently, 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 benefits in screening programs, intermediate measures associated with demonstrated RCT effectiveness can be informative.[14] The gFOBT RCTs demonstrate that a shift to an earlier stage of cancer in a program that involves repeated screening offers is associated with reduced mortality.[7,8,10,54] Thus earlier detection by a new test to at least a comparable degree is highly desirable; for instance, it has now been demonstrated that screening with FIT leads to earlier detection.[49]

The association of adenoma detection and removal in screening with the reduction of CRC incidence and mortality is now proven by the RCTs of FS screening.[5] Thus FS is an expeditious comparator for evaluating new tests that target preinvasive lesions, because a potential surrogate measure for predicting a reduction in incidence is the detection (the TPR) of those lesions considered to be at high risk of progressing to CRC.

Interval cancers, ie, 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 when the earlier phases of evaluation have focused primarily on assessment of test-positive cases (ie, 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[55]; however, 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 to monitor not only the effect of a new test on overall uptake but also its acceptability to all socioeconomic and ethnic groups to avoid widening the inequalities gap.

Phase 4 study design

Studies should follow the design outlined for Phase 3 evaluation but should also include multiple rounds of

screening (at least several with the interval matched to the perceived duration of effect of each test), with plans to ascertain the outcomes relating to those measures deemed important; namely, participation, detection, cost, adverse effects, earlier detection, and interval (new or missed) lesions.

Such studies will be extremely costly and normally would be feasible only in the context of public health screening strategies that are already in place, in which methods to collect outcome measures are already designed and operational. In other words, Phase 3 evaluation is sufficient to lead to the 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 an evaluation of screening programs. Given good information on costs, the comparative cost effectiveness of different tests can be determined as described.[56]

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.[3,57] Sophisticated, retrospective molecular analyses of material in biospecimen banks can serve to identify candidate biomarkers that might become the objective of the screening test.

If such laboratory research identifies a promising biomarker, then it can be initially evaluated as for Phases 1 and 2 by a simple study in cases and controls. Doing this, however, may assume that the retrospective biospecimen banks are adequate to identify the best candidate. Usually, this is not the case, because discovery is often undertaken on limited numbers of samples obtained from strictly categorized materials that often are not typical of screen-detected lesions. A further technological challenge arises if resected tissue specimens are used to identify the biomarker; however, use of the biomarker in screening involves measurement in a biologic sample, such as blood or feces. Many factors may influence the appearance of the biomarker in the biologic sample, and there is a chance that it might not be of the same molecular structure in blood or feces as in tissue, because 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 studies, or perhaps even Phase 3 studies. Indeed, access to the appropriately characterized population with biologic samples, which 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, ie, "validated," in Phase 1, 2, and 3 studies.[57]

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 that this can wait until Phase 4 studies have been undertaken,[3] although that seems impractical, because no commercial entity would proceed with test development under such circumstances. Using the logistics and infrastructure of existing screening programs can also help reduce costs of such studies. Expensive studies have included the evaluation of new, noninvasive tests in colonoscopic screening participants.[58] Although 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 1 phase to the next. Although 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 on a study. It is noteworthy that 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 at which the considerable degree of evidence needed for its inclusion 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.

FUNDING SUPPORT

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

CONFLICT OF INTEREST DISCLOSURES

Graeme P. Young reports grants and nonfinancial support from Eiken Chemical Company and grants and personal fees from Clinical Genomics Pty Ltd outside the submitted work; he has a patent "Down Markers: A Method of Diagnosing Neoplasms (II PCT/AU2008/001565) licensed to Clinical Genomics. Carlo Senore reports nonfinancial support from Covidien/Given Imaging, Medical System, and Given Imaging outside the submitted work. Wendy S. Atkin reports nonfinancial support from Eiken Co. Ltd (MAST is UK distributor) outside the submitted work. Stephen H. Itzkowitz reports personal fees from Exact Sciences Corporation outside the submitted work. Bernard Levin reports personal fees from Exact Sciences ad Medial Research during the conduct of the study. Gerrit A. Meijer reports nonfinancial support from Exact Sciences and Sysmex outside the submitted work; he has a patent pending on biomarkers for the early detection of colorectal cancer.

REFERENCES

1. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med.* 2006;144:850-855.
2. Wilson JMG, Junger G, World Health Organization (WHO). Principles and Practice of Screening for Disease. Public Health Papers, no. 34. Geneva, Switzerland, WHO; 1968.
3. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst.* 2001;93:1054-1061.
4. Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet.* 2010;375:1624-1633.
5. Elmunzer BJ, Hayward RA, Schoenfeld PS, Saini SD, Deshpande A, Waljee AK. Effect of flexible sigmoidoscopy-based screening on incidence and mortality of colorectal cancer: a systematic review and meta-analysis of randomized controlled trials [serial online]. *PLoS Med.* 2012;9:e1001352.
6. Holme O, Loberg M, Kalager M, et al. Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: a randomized clinical trial. *JAMA.* 2014;312:606-615.
7. Hardcastle JD, Chamberlain JO, Robinson MH, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet.* 1996;348:1472-1477.
8. Mandel JS, Bond JH, Church TR, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N Engl J Med.* 1993;328:1365-1371.
9. Mandel JS, Church TR, Bond JH, et al. The effect of fecal occult-blood screening on the incidence of colorectal cancer. *N Engl J Med.* 2000;343:1603-1607.
10. Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet.* 1996;348:1467-1471.
11. Segnan N, Armaroli P, Bonelli L, et al. Once-only sigmoidoscopy in colorectal cancer screening: follow-up findings of the Italian Randomized Controlled Trial—SCORE. *J Natl Cancer Inst.* 2011;103:1310-1322.
12. Schoen RE, Pinsky PF, Weissfeld JL, et al. Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. *N Engl J Med.* 2012;366:2345-2357.
13. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Public Health.* 1984;74:979-983.
14. Young GP, Allison J. Screening for colorectal cancer. In: Yamada T, Alpers D, Kaplowitz N, Laine L, Owyang C, Powell D, eds. Textbook of Gastroenterology. 5th ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2008:170-182.
15. Young GP, Cole SR. Which fecal occult blood test is best to screen for colorectal cancer? *Nat Clin Pract Gastroenterol Hepatol.* 2009;6:140-141.
16. van Rossum LG, van Rijn AF, Laheij RJ, et al. Random comparison of guaiac and immunochemical fecal occult blood tests for colorectal cancer in a screening population. *Gastroenterology.* 2008;135:82-90.
17. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Ann Intern Med.* 2011;154:22-30.
18. Baxter NN, Warren JL, Barrett MJ, Barrett MJ, Stukel TA, Doria-Rose PV. Association between colonoscopy and colorectal cancer mortality in a US cohort according to site of cancer and colonoscopist specialty. *J Clin Oncol.* 2012;30:2664-2669.
19. Nishihara R, Wu K, Lochhead P, et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. *N Engl J Med.* 2013;369:1095-1105.
20. Bossuyt PM, Lijmer JG L, Mol BW. Randomised comparisons of medical tests: sometimes valid, not always efficient. *Lancet.* 2000;356:1844-1847.
21. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ.* 2006;332:1089-1092.
22. Hiwatashi N, Morimoto T, Fukao A, et al. An evaluation of mass-screening using fecal occult blood-test for colorectal-cancer in Japan—a case control study. *Jpn J Cancer Res* 1993;84:1110-1112.
23. Saito H, Soma Y, Koeda J, et al. Reduction in risk of mortality from colorectal cancer by fecal occult blood screening with immunochemical hemagglutination test. A case-control study. *Int J Cancer.* 1995;61:465-469.
24. Zappa M, Castiglione G, Grazzini G, et al. Effect of faecal occult blood testing on colorectal mortality: results of a population-based case-control study in the district of Florence, Italy. *Int J Cancer.* 1997;73:208-210.
25. Saito H, Soma Y, Nakajima M, et al. A case-control study evaluating occult blood screening for colorectal cancer with hemoccult test and an immunochemical hemagglutination test. *Oncol Rep.* 2000;7:815-819.
26. Nakajima M, Saito H, Soma Y, Sobue T, Tanaka M, Munakata A. Prevention of advanced colorectal cancer by screening using the immunochemical faecal occult blood test: a case-control study. *Brit J Cancer.* 2003;89:23-28.
27. Lee KJ, Inoue M, Otani T, Iwasaki M, Sasazuki S, Tsugane S. Colorectal cancer screening using fecal occult blood test and subsequent risk of colorectal cancer: a prospective cohort study in Japan. *Cancer Detect Prevent.* 2007;31:3-11.
28. Saito H. Screening for colorectal cancer: current status in Japan. *Dis Colon Rectum.* 2000;43(10 suppl):S78-S84.
29. Oort FA, Terhaar Sive Droste JS, Van Der Hulst RW, et al. Colonoscopy-controlled intra-individual comparisons to screen relevant neoplasia: faecal immunochemical test vs guaiac-based faecal occult blood test. *Aliment Pharmacol Ther.* 2010;31:432-439.
30. von Karsa L, Patnick J, Segnan N, et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. *Endoscopy.* 2013;45:51-59.
31. Halloran SP, Launoy G, Zappa M; International Agency for Research on Cancer. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First edition—faecal occult blood testing. *Endoscopy.* 2012;44(suppl 3):SE65-SE87.
32. Lane JM, Chow E, Young GP, et al. Interval fecal immunochemical testing in a colonoscopic surveillance program speeds detection of colorectal neoplasia. *Gastroenterology.* 2010;139:1918-1926.
33. de Wijkerslooth TR, de Haan MC, Stoop EM, et al. Reasons for participation and nonparticipation in colorectal cancer screening: a randomized trial of colonoscopy and CT colonography. *Am J Gastroenterol.* 2012;107:1777-1783.
34. Cheng H, Macaluso M. Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results. *Epidemiology.* 1997;8:104-106.

35. Cheng H, Macaluso M, Waterbor J. Estimation of relative and absolute test accuracy. *Epidemiology.* 1999;10:566-568.
36. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, UK: Oxford University Press; 2003.
37. St John DJB, Young GP, Alexeyeff MA, et al. Evaluation of new occult blood-tests for detection of colorectal neoplasia. *Gastroenterology.* 1993;104:1661-1668.
38. Allison JE, Tekawa IS, Ransom LJ, Adrain AL. A comparison of fecal occult-blood tests for colorectal cancer screening. *N Engl J Med.* 1996;334:155-159.
39. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ.* 2001;323:157-162.
40. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Evidence base of clinical diagnosis—designing studies to ensure that estimates of test accuracy are transferable. *Brit Med J.* 2002;324:669-671.
41. Young GP, Macrae FA, St John DJB. Clinical methods of early detection: basis, use and evaluation. In: Young GP, Rozen P, Levin B, eds. Prevention and Early Detection of Colorectal Cancer. London, UK: Saunders; 1996:241-270.
42. Cole SR, Young GP, Esterman A, Cadd B, Morcom J. A randomised trial of the impact of new faecal haemoglobin test technologies on population participation in screening for colorectal cancer. *J Med Screen.* 2003;10:117-122.
43. Young GP, St John DJB. Selecting an occult blood test for use as a screening tool for large bowel cancer. *Front Gastrointest Res.* 1991; 18:135-156.
44. Young GP, St John DJ, Winawer SJ, Rosen P, (WHO) World Health Organization and OMED (World Organization for Digestive Endoscopy). Choice of fecal occult blood tests for colorectal cancer screening: recommendations based on performance characteristics in population studies: a (WHO) World Health Organization and OMED (World Organization for Digestive Endoscopy) report. *Am J Gastroenterol.* 2002;97:2499-2507.
45. Young GP. Screening for colorectal cancer: alternative faecal occult blood tests. *Eur J Gastroenterol Hepatol.* 1998;10:205-212.
46. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ.* 1986;134:587-594.
47. Smith A, Young GP, Cole SR, Bampton P. Comparison of a brush-sampling fecal immunochemical test for hemoglobin with a sensitive guaiac-based fecal occult blood test in detection of colorectal neoplasia. *Cancer.* 2006;107:2152-2159.
48. Sillars-Hardebol AH, Carvalho B, van Engeland M, Fijneman RJ, Meijer GA. The adenoma hunt in colorectal cancer screening: defining the target. *J Pathol.* 2012;226:1-6.
49. Cole SR, Tucker GR, Osborne JM, et al. Shift to earlier stage at diagnosis as a consequence of the National Bowel Cancer Screening Program. *Med J Aust.* 2013;198:327-330.
50. Levin B, Liberman DA, McFarland, et al; American Cancer Society Colorectal Cancer Advisory Group. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology.* 2008;134:1570-1595.
51. Atkin WS, Morson BC, Cuzick J. Long-term risk of colorectal cancer after excision of rectosigmoid adenomas. *N Engl J Med.* 1992; 326:658-662.
52. Miles A, Cockburn J, Smith RA, Wardle J. A perspective from countries using organized screening programs. *Cancer.* 2004;101: 1201-1213.
53. Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, van Ballegooijen M, Zauber AG, Jemal A. Contribution of screening and survival differences to racial disparities in colorectal cancer rates. *Cancer Epidemiol Biomarkers Prev.* 2012;21:728-736.
54. Faivre J, Dancourt V, Lejeune C, et al. Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study. *Gastroenterology.* 2004;126:1674-1680.
55. Cole SR, Gregory T, Whibley A, et al. Predictors of re-participation in faecal occult blood test-based screening for colorectal cancer. *Asian Pac J Cancer Prev.* 2012;13:5989-5994.
56. Chen LS, Liao CS, Chang SH, Lai HC, Chen TH. Cost-effectiveness analysis for determining optimal cut-off of immunochemical faecal occult blood tests for population-based colorectal cancer screening. *J Med Screen.* 2007;14:191-199.
57. Ransohoff DF. Rules of evidence for cancer molecular marker discovery and validation. *Nat Rev Cancer.* 2004;4:309-311.
58. Imperiale TF Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med.* 2014; 370:1287-1297.

総論

大腸がん検診のあり方——最近のエビデンスを踏まえて

Author 斎藤 博 国立がん研究センターがん予防・検診研究センター検診研究部

KeyWords ◎大腸がん検診 ◎エビデンス ◎便潜血検査 ◎大腸内視鏡検査

Headline

1 便潜血検査による大腸がん検診は，その有効性ががん検診の中でも最も確立している．
2 sigmoidoscopyによる検診の有効性も確立している．
3 健常者では初回の内視鏡検診で異常がなければ，そうでない場合に比べ10年以上の長期にわたって浸潤がんのリスクが低い可能性が示されている．

本稿に与えられたテーマは，消化器がん診療において最近大きな変化が起こっている領域の一つである大腸がん検診について最新知見を提供することである．大腸がん検診に関連するエビデンスとしては，内視鏡，特にsigmoidoscopyによる検診に関する有効性が確立したことが特筆すべき成果であり，colonoscopyの検診についても評価できる質の研究報告が出始めている．またスクリーニング以外にも，検診関連事項としてポリープ切除後のリスクについても以前から若干の知見は積まれている．それらを踏まえると，スクリーニングおよび発見ポリープのマネジメント等について，より具体的かつ合理的な判断が可能である．たとえばこれまで腺腫が見つかった場合には過剰なフォローアップの内視鏡が行われてきたが，一定の間隔を開ける根拠が示されている．

大腸がんの死亡率・罹患率

大腸がんは過去40年以上にわたってその死亡率が増加を続け，わが国の主たるがん死亡の原因となっている．部位別年齢調整死亡率は，女性では2003年から第1位，男性でも2007年から第3位と男女を通じて高く，わが国のがん対策において重要度の高いがんである[1]．1990年以降はその増加が止まり現在ではやや減少に転じているが，今後もわが国の主要ながんであり続けると考えられる．

大腸がん検診およびスクリーニング法の位置づけ

検診を行う第一の条件は，健康対策上の重要な課題であること(表1)，つまりがん検診の場合は死亡率や罹患率が高いことであり，その意味で大腸がんは世界の先進国に共通の検診の対象がんである．また大腸がん検診は便潜血検査(faecal occult blood test：FOBT)による死亡率減少，さらに罹患率の減少効果も実証されており[2,3]がん検診の中でもそれを対策として行う条件に最も合致しているといえ，世界的にその導入が進んでいる．

大腸がん検診については，次項に要約するように内視鏡による検診についても最近エビデンスが確立しつつある．しかし内視鏡検診はそのキャパシティー，偶発症などの不利益が懸念される他，施策としてのがん検診の条件[4]を満足するには至っておらず，内視鏡は施策としての検診においては精検の位置づけである．スクリーニング法は便潜血検査であり，その精密検査として大腸内視鏡検査を行うのが世界のプロ

0370-999X/15/¥100/頁/JCOPY

グラムの概要である.

スクリーニングのエビデンス(科学的根拠)[1]

1．便潜血検査(FOBT)(表2)

a）化学法FOBT(化学法)の死亡率減少効果

大腸がん検診は欧米で1960年代に開発されたグアヤックろ紙法による化学法FOBTについて，複数のランダム化比較試験(randomized controlled trial:RCT)が行われ，一致して検診による死亡率減少効果が示されている．他のがん検診でRCTによる有効性のエビデンスがあるものはマンモグラフィによる乳がん検診があげられるが，RCTの結果は一致していない．有効性の証拠の水準が最上位であるRCTがすべて一致して有効性を示しているFOBTによる大腸がん検診は，がん検診の中で最も明確に科学的根拠が確立していることをまず認識したい(表2)[3].

最初の米国ミネソタ研究では逐年群で33%の死亡率低下[2]，隔年群でも21%の低下が観察された．さらに，逐年，隔年の間隔での検診について罹患率の減少も示され，スクリーニングにより浸潤がんが減少する(逐年で20%，隔年で17%)ことも明らかとなった[5]．最近，同研究の30年の長期観察後の効果が逐年，2年間隔でそれぞれ大腸がん死亡率は32%，22%低下したと報告され，有効性が長期にわたって確認された.

ミネソタ研究以外に欧州で行なわれた英国とデンマークの2研究，さらにフランスにおける地域ブロック割付による1研究で隔年検診により，15～18%の死亡率低下が観察され，複数のメタ解析ではいずれも14～16%の死亡率低下を示した[3].

b）免疫法FOBT(免疫法)に関する死亡率減少効果

わが国で始まった免疫法FOBTは，有効性の科学的根拠としてはRCTはなく，日本からの4件を含めた5件の症例対照研究と1件のコホー

表1　スクリーニングの基準—Wilson & Junger screening criteria

1. その疾患が健康上の重大な問題になっている
2. 患者に対して認められた治療がある
3. 診断と治療を行う機関がある
4. 診断可能な無症状または症状のある早期の段階がある
5. 適切な検査がある
6. 検査法は集団に受け容れられるものである
7. 潜伏期からの進展など疾患の自然史が十分把握されている
8. 誰を患者として治療すべきか合意の得られた方針がある
9. (診断・治療を含め)スクリーニング*のコストが医療費全体とバランスがとれる
10. スクリーニング*は継続的なプロセスで一回こっきりではない

*：Case finding
(WHO 1968)

表2　便潜血検査の有効性に関する主な証拠—化学法

年	報告者	Journal	地域	研究デザイン	スクリーニング法 検診間隔・対象年齢	RR(95% CI) 罹患	死亡
1993, 1999 2000	Mandel	N Engl J Med	US Minnesota	RCT	逐年・隔年 50～80	逐年 0.80(0.70-0.90) 隔年 0.83(0.73-0.94)	逐年 0.67(0.50-0.87) 隔年 0.80(0.70-0.90)
1996, 2002	Hardcastle	Lancet	UK Nottingham	RCT	隔年 45～74	/	0.85(0.74-0.98)
1996, 2002	Kronborg	Lancet	Denmark Funen	RCT	隔年 45～75	/	0.82(0.68-0.99)
2004	Faivre	Gastroenterology	French	RCT	隔年 45～74	1.01(0.91-1.12)	0.84(0.71-0.99)

RCT:randomized controlled trial

表3　便潜血検査の有効性に関する主な証拠―免疫法

年	FOBT	報告者	Journal	地域	研究デザイン	検診間隔・対象年齢	RR(95% CI) 罹患	死亡
1993	免疫法	Hiwatashi	Jpn J Cancer Res	Japan	CCS	逐年 45～69	/	0.24(0.08-0.76)
1995	免疫法単独	Saito	Int J Cancer	Japan	CCS	逐年 40～79	/	0.40(0.17-0.92)
1997	免疫法＋化学法	Zappa	Int J Cancer	Italy Florence	CCS	隔年 41～75	/	0.54(0.3-0.9)
2000	免疫法＋化学法	Saito	Oncol Rep	Japan	CCS	逐年 >40	/	逐年 0.20(0.08-0.49) 隔年 0.17(0.04-0.75)
2003	免疫法単独	Nakajima	Br J Cancer	Japan	CCS	逐年・隔年 40～79	進行がん 0.54(0.29-0.99)	/
2007	免疫法(＋化学法)	Lee	Cancer Causes Control	Japan	Cohort	40～59	進行がん 0.41(0.27-0.63)	0.28(0.13-0.61)

ト研究が報告されている(表3)[3]．死亡率減少効果に関する免疫法単独による研究は1研究のみであるが，FOBT 1日法の逐年検診により60%の死亡率が減少すると示唆されている[16]．ほかの一部化学法を含む検診に関する4研究もあわせ，一致して死亡リスク減少効果を示す結果が示され，リスク低下は46～80%と報告されている．また進行がんのリスク低下も報告されている[3](表3)．

c）FOBTの感度

化学法については研究対象の全例に内視鏡とFOBTを行い，内視鏡を基準としてFOBTで1回スクリーニングする感度(スクリーン感度)は30%と報告されている

免疫法の精度に関しては健常者コホートにおいて全例に内視鏡検査と免疫法FOBTを行い，免疫法FOBTのスクリーン感度は1日法56～67%，2日法77～83%，3日法89%[2]，特異度は97～98%と報告されている．化学法と免疫法を同時に行って比較した研究では免疫法の感度が高いと報告されている[5]．また一般の人口集団のランダム割付により免疫法と化学法を行う群に分けて，受容度(受診率)の影響も含めてadvanced adenomaとがんの発見率を比較した研究で，免疫法群で受診率，発見率，陽性反応適中度が高く，化学法より優れていることが明確に示された[6]．免疫法は化学法に替わるべき検診法として各国で導入が始まっている．

d）便潜血検査の検診間隔と対象年齢に関する証拠(表2, 3)

化学法のRCTにより検診間隔は2年まで死亡率減少効果が確定している．なおミネソタ研究では1年間隔では33%，2年間隔では20%の低下であり，罹患の減少も含め，1年間隔でより効果が高いことが示されている．免疫法では2年間隔で行われた研究もあり，化学法のエビデンスと合わせ2年間は有効性が認められたといえる．海外では2年間隔の検診を推奨する国が多い．

有効性のある年齢については，RCTではおもに45～74歳を対象としている(表3)．この年代についてもっとも強い証拠があるといえる．わが国では対象年齢に上限はないが世界的には年齢上限を定めており，精検の内視鏡の前処置など負担が大きいので高齢者には不利益が懸念され，わが国でも今後，設定すべきと考えられる．

2．内視鏡による検診(表4)

a）S状結腸内視鏡検査　有効性の証拠のレベル

以前から症例対照研究2研究により硬性S状結腸内視鏡(RS)による検診のRSが届く直腸・

表 4　大腸内視鏡検診の有効性に関する主な証拠

年	報告者	Journal	研究デザイン	スクリーニング法	RR(95% CI) 罹患	死亡
2010	Atkin	Lancet Oncol	RCT	Sigmoidoscopy	全大腸 0.77(0.70-0.84) Distal 0.64(0.57-0.72)	全大腸 0.69(0.59-0.82)
2011	Segnan	JNCI	RCT	Sigmoidoscopy	全大腸 0.82(0.69-0.96) Distal 0.76(0.62-0.94) Proximal 0.91(0.69-1.20)	全大腸がん 0.78(0.56-1.08) Distal 0.73(0.47-1.12) Proximal 0.85(0.52-1.39)
2012	Schoen	NEJM	RCT	Sigmoidoscopy*	全大腸 0.79(0.72-0.85) Distal 0.71(0.64-0.80) Proximal 0.86(0.76〜0.97)	全大腸 0.74(0.63-0.87) Distal 0.50(0.38-0.64) Proximal 0.97(0.77〜1.22)
2014	Holme	JAMA	RCT	Sigmoidoscopy/ sigmoidoscopy＋FOBT	全大腸 0.80(0.70-0.92)	全大腸 0.73(0.56-0.94)
2008	Baxter	Ann Intern Med	CCS	Colonoscopy	/	全大腸 0.69(0.63-0.74) Left-Sided 0.39(0.34-0.45) Right-Sided 1.07(0.94.1.21)
2013	Doubeni	Ann Intern Med	CCS	Colonoscopy	0.29(0.15-0.58) Right- sided 0.36(0.16-0.80) Left-sided 0.26(0.06-1.11) (進行がん罹患率)	/
2013	Nishihara	NEJM	Cohort	Colonoscopy	全大腸 0.44(0.38-0.52) Distal 0.24(0.18-0.32) Proximal0.72(0.57-0.92	全大腸 0.32(0.24-0.45) Distal 0.18(0.10-0.31) Proximal 0.47(0.29-0.76)

S 状結腸がん死亡率のリスク低下が示唆されていたが，最近，Flexible Sigmoidoscopy(以下 Sigmoidoscopy)を 1 回だけ行う検診の有効性評価の大規模 RCT 3 研究と，3〜5 年間隔で行う 1 研究が報告された．これら質の高い RCT のうち，2 研究において 26〜31% の死亡率減少効果が示され，3 研究において 18〜23% の罹患率減少効果が示されている．英国の研究では罹患率減少効果は 10 年に渡って認められ，効果は長期間持続すると考えられる．

b）全大腸内視鏡検査(CS)

全大腸内視鏡検査(colonoscopy：CS)の有効性に関する質の高い研究はなかったが，比較的最近になってようやく中等度程度の質の症例対照研究などが報告され，死亡率減少効果が示唆されていた．これらの研究では CS のスクリーニングは左側の大腸がんのリスク低下には寄与するが，右側の結腸がんのリスクは低下させないことが示唆されている．一方，最近報告されたコホート研究は非常に質の高いもので，大腸内視鏡検診受診が 68% 大腸がん死亡リスクを下げたことが報告された．とりわけ直腸，S 状結腸がんではリスク低下は 82% と効果は大きかった．この研究では右側結腸がんのリスクも低下する結果であるが，リスク低下の程度は左側に比べ小さかった．これらの結果は左側と右側で大腸がんの biology が異なることや，CS の感度が右側のがんに対して低いことなどを示唆する．

c）CS による検診の不利益

CS を将来，対策型検診として検討する際に必要な不利益のデータについては，前投薬や下剤による前処置について死亡例が報告されている．CS 自体による偶発症は 1998 年から 2002 年までの約 300 万例の検査で 0.069%(2,038 例)，死亡は 0.00088%(26 例)と高くはないとも言えるが，これらの報告は，大規模専門施設からのものであるため，実際より過小評価の可能性が高い．海外では詳しく客観性の高い調査報告があり，大腸穿孔の頻度の報告値は screening CS についておよそ 0.1〜0.3% と日本より高い．

d）内視鏡検診のエビデンスの位置づけ

Sigmoidoscopyのエビデンスは確立し，CSの有効性も確実と考えられる．しかし，将来の対策型検診への導入には大腸穿孔など主要な不利益の実態把握が不可欠である．さらに深部結腸がんのリスクについて遠位のがんより効果が低い可能性が指摘されており，今後の研究課題である．またCSの処理能力は精検法に限っても必ずしも十分ではなく，現状では専門施設においての実施にとどまる．ただし，他のがん検診とは異なり，唯一，推奨できる任意型検診法であり，FOBTによる定期的な検診に加え，50～60代で一度行うことは積極的に勧められる．

3．その他―CT―colonography(3D-CT)

死亡率をエンドポイントとした研究はない．海外でNational CT colonography Trial(米国ACRIN研究　2007)など，腺腫の診断能を見たいくつかの前向き研究があり，大きな腺腫については内視鏡の感度とそれほど差がないとされる．しかし日本で行われている3D-CTについての研究報告はまだない．日本では画像表示法や前処置，糞便の画像処理のための造影剤が異なるのでその評価が別途必要である．とはいえ，海外の方法については精検法としては一定の評価は得られたと言える．

スクリーニング後の治療やフォローアップなどに関連するエビデンス

1．ポリープ切除の効果[7]

FOBTの検診で検診を提供された群において大腸がん罹患率が減少するエビデンスが示されているが，その要因は頻繁に行われるポリペクトミーであるとされてきた．米国National polyp studyの長期観察から，ポリペクトミーが大腸ポリープ患者の大腸がん死亡リスクを減少させる効果も示された．これらから大腸がん検診の効果は大腸がんの早期発見のみならず，ポリープ切除の効果にもよることが強く示唆される．

2．ポリープ切除後，大腸内視鏡後の大腸がんや腺腫のリスク[8,9]

検診では便潜血検査陽性者の30～40%に大腸ポリープが発見され，その一部が切除の対象となる．また従来，ポリープ切除後の患者は基本的には内視鏡でフォローアップされるが，1年毎のフォローアップなど頻繁な内視鏡検査が行われてきた．この方針に関連するエビデンスとしては，初回内視鏡の所見別にリスクが異なり，腺腫なし，6 mm未満の腺腫のみあり，6 mm以上の腺腫あり，粘膜内がんありで10 mm以上の腺腫あるいは粘膜内がんのリスクが高くなり，初回腺腫なしでは1年後のリスクは0.1%にとどまることが後ろ向き研究ながら報告されている．

RCTであるNational polyp studyにおいて，切除後3年でのフォローと1，3年後のフォローでは浸潤がんのリスクに差はないと示されたことは重要なエビデンスである．まだ最終結果として論文にはなっていないがわが国のJapan polyp studyからも同様の結果が得られているという．これらにより，切除後3年間は毎年のフォローに比べ浸潤がんのリスクは上昇しないことが示されたといえる．またコホート研究により，初回，腺腫がない場合には5年間，浸潤がんリスクは上昇しないことも示唆されている．さらにsigmoidoscopyに関するRCTでは，1回行った群で対照群に比べ10年間罹患率(浸潤がんの)が有意に低下することが示され，これまでの症例対照研究と合わせ，内視鏡検診の効果は長期にわたって持続することも明らかとなったと言える．フォローアップの間隔年数としては，浸潤がんのリスクが上昇しない期間の最大値が基準となることから，従来のような短期間でのフォローアップは必要はないと考えられる．

以上から，ごく一部の非常にハイリスクな症例を除けば切除後3年以上は開けられること，その後はFOBTの検診に戻すことが合理的と考えられる．そのようなリスク要因についてより

詳細な根拠を得るために，内視鏡によるフォローの間隔とリスクに関する研究の重要度は高い．また，前述のように右側がんに対する内視鏡後のリスクが左側がんとは異なる可能性についても研究課題として注目される．

今後行うべきこと―日本で大腸がん検診の成果をあげるために[3]

大腸がん検診はきちんと行うことで，その死亡率を国レベルで低下させる成果が十分期待できる．海外，特に欧州を中心に，乳がんと子宮がん検診で国レベルの死亡率を減少させた仕組みは組織型検診という方法であり，その骨子は科学的根拠のある検診を徹底的に精度管理して行うというものである[4]．わが国ではこれまでがん検診の成果があがっていないが，それはこの組織型検診の条件から逸脱している状況に原因がある．がん検診全般に精度管理の基盤を作らずにやりっ放しの検診が横行している状況である．大腸がん検診はその科学的根拠が確立しており，精度管理がポイントになる．そもそも，精度管理の体制ができていなければ検診は行うべきではないというのが検診の原則であり，それが組織型検診の背景である．日本はまだ成果をあげられる水準まで検診の理解が進んでいないのである．

文　献

1) 厚生労働省人口動態統計．
2) 斎藤　博，他：死亡率低下を目指した大腸がん検診の将来像．日本臨牀 72:15-21，2014
3) 斎藤　博，他：大腸がん検診のエビデンスと今後の研究の展望．日消誌 111:453-463，.2014
4) 斎藤　博，他：がん治療エッセンシャルガイド．南山堂，190-197，2012
5) 斎藤　博：大腸癌検診の発見率，偽陰性率はどのくらいですか？　大腸癌 FRONTIER 2:94-97，2009
6) Van Rossum LG, et al.:Random comparison of guaiac and immunochemical fecal occult blood test for colorectal cancer in a screening population. Gastroenterology 136:82-90, 2008
7) Winawer SJ, et al.:Randomized comparison of surveillance intervals after colonoscopic removal of newly diagnosed adenomatous polyps. N Engl J Med 328:901-906, 1993
8) Imerilale TF, et al.:Five-year risk of colorectal neoplasia after negative screening colonoscopy. N Engl J Med 359:1218-1224, 2008
9) Matsuda T, et al.:Five-year incidence of advanced neoplasia after initial colonoscopy in Japan:a multicenter retrospective cohort study. Jpn J Clin Oncol 39:434-442, 2009

著者連絡先(〒104-0045)　東京都中央区築地 5-1-1
国立がん研究センターがん予防・検診研究センター検診研究部　斎藤　博

特別連載

日本のがん対策の新しい動き
―科学的根拠に基づいたがん対策を進めるために―

がんの早期発見分野におけるがん対策進捗管理指標と進捗状況

奥山絢子[*1]　東　尚弘[*1]　斉藤　博[*2]　雑賀公美子[*2]
町井涼子[*2]　松田和子[*2]　若尾文彦[*3]

*1 国立研究開発法人国立がん研究センターがん対策情報センターがん政策科学研究部
*2 国立研究開発法人国立がん研究センターがん予防・検診研究センター検診研究部
*3 国立研究開発法人国立がん研究センターがん対策情報センター

はじめに

平成24年6月に第二期のがん対策推進基本計画（以下，基本計画）が策定され，わが国のがん対策における方向性が示された．基本計画では，1）がんによる死亡者の減少，2）全てのがん患者とその家族の苦痛の軽減と療養生活の質の維持向上，3）がんになっても安心して暮らせる社会の構築の三つが全体目標として掲げられている．この全体目標を達成するために，就労や小児がん等の9つの分野別施策と個別目標が明記されている．さらに基本計画では，新たに目標の達成状況の把握とがん対策を評価する指標の策定を行うこと，そして個々の取り組むべき施策が個別目標の達成に向けてどの程度の効果をもたらしているのか評価することについても言及された．そこで，厚生労働省研究班「がん対策における進捗管理指標の策定と計測システムの確立に関する研究」（代表若尾文彦）では，これまでがん対策の進捗状況について把握するための指標の策定に取り組んできた[1]．本稿では，研究班のがん対策の進捗管理指標策定の一貫として取り組んできたがんの早期発見分野についての指標とその現状について報告する．

1　基本計画におけるがんの早期発見分野の個別目標

がんの診断や治療の進歩により，一部のがんでは早期発見・早期治療を行うことでがんの死亡率を減少させることが可能となった．がん検診の目的は，集団におけるがんによる死亡を減少させることであり，単に多くのがんを発見することが目的ではない．がんの中には，自然に退縮するものや進行が非常に遅いため死亡の原因につながらないものが，一定割合存在しているため，それらのがんを発見しても，がん死亡の減少に結びつかないばかりか，無駄な検査や治療，そしてそれに伴う有害事象や精神的苦痛，経済損失を発生させることとなり，がん検診を論ずる上では不利益に関しても利益と同時に検討することが重要であるとされている．

基本計画では，がんの早期発見の個別目標として，1）5年以内に全ての市区町村が精度管理・事業評価を実施するとともに，科学的根拠に基づくがん検診を実施すること，2）がん検診の受診率を5年以内に50%（胃，肺，大腸がん検診は，当面40%）を達成すること，3）がん検診の項目や方法について国内外の知見を収集して検討し，科学的根拠のあるがん検診を実施することを掲げている．つまり，わが国におけるがんによる死亡を減少させるためには，科学的根拠に基づくがん検診を，適切な精度管理下で実施し，多くの国民がそれらのがん検診を定期的に受診することで初めてその効果が期待できるといえる．そこで，研究班ではこれまでの「がん検診のあり方に関する検討会」[2]での議論を踏まえ，がんの早期発見分野に関する進捗管理指標と現状について整理した（表1）．

2　わが国のがん検診制度

現在，がん検診は，健康増進法に基づく健康増進事業として市区町村で実施されている住民検診に代表される対策型検診と人間ドックなどの任意型検診とがある．対策型検診は，地域住民等集団における死亡率の減少を目的として行われる検診であり，その対象は個人の検診受診の希望によらず健常な住民全てを対象としている．このため，対策型検診ではがん検診の有効性（＝死亡率減少）を示す科学的根拠が必要であり，また不必要な精密検査が行われることによる受診者の身体的・精神的な負担や侵襲的な処置に伴う有害事象を避けることが重要となる．一方で，任意型検診は個人を主眼に置いての死亡リスクを減少させることを目的としており，基本的に検診を受けるか否かは個人の選択となる．任意型検診においても，本来は科学的に有効性が認められている検診を行うべきであるが，まだ有効性が未確立である検診についても個人の希望により実施される場合もある．国のがん対策の評価においては，個人の希望によって実施されている検診よりも，制度的な側面に着目する必要がある．よって，研究班では健康増進法に基づく市区町村におけるがん検診事業の実施状況およびその精度管理状況についての指標を整理した．なお，国民のがん検診受診率については，就業している国民は，労働安全衛生法に基づく職域での健康診断受診時に合わせてがん検診を受診する場合や個人の人間ドックで受診する場合があることから，住民検診の記録だけでは把握することができない．そこで，国民のがん検診受診率については，国の基幹統計である国民生活基礎調査を用いて評価した．

表1　がんの早期発見分野における進捗管理指標

基本計画の個別目標	平成28年達成目標	指標
市区町村の科学的根拠のあるがん検診の実施	0%	1a：指針に基づかないがん検診を実施している市区町村の割合
	100%	1b：指針に基づくがん検診を実施している市区町村の割合
市区町村のがん検診の精度管理・事業評価の実施	100%	2：「事業評価のためのチェックリスト」を実施している市区町村の割合
	100%	3：精検受診率，精検未把握率，精検未受診率，精検未受診・未把握率，要精検率，がん発見率，陽性反応適中度
	—	4：がん検診のコールリコール（個別受診勧奨・再勧奨）を実施している市区町の割合
がん検診の受診率	50%（胃がん，肺がん，大腸がんは当面40%）	5：がん検診受診率

3 進捗管理指標と現状

国民全体のがん死亡を減らすためには，1）科学的根拠に基づくがん検診を，2）適切な精度管理下で実施し，3）多くの国民がそれらのがん検診を定期的に受診することが重要となる．そこで，がんの早期発見分野の進捗管理状況について，これら三つの段階別に整理した．

1 ◉ 市区町村における科学的根拠に基づくがん検診の実施状況

先述のように，がん検診の対象は，健康な人であり，がん検診におけるがんの早期発見・早期治療という利益の他に，偽陽性，偽陰性，偶発症，受診者の身体的・心理的な負担等の不利益について考慮する必要がある．厚生労働省は，「がん予防重点健康教育及びがん検診実施のための指針（平成20年4月厚生労働省健康局長通知）」を定め，市区町村における科学的根拠に基づくがん検診を推進している（表2）[3]．そこで，この指針に基づくがん検診および指針以外の検診の実施状況について，厚生労働省の「市区町村におけるがん検診の実施状況調査」の結果を用いて現状について整理した．また，市区町村における指針に基づくがん検診の精度管理状況については，後述する指標2：「事業評価のためのチェックリスト」を実施している割合および指標3：精検受診率，精検未把握率，精検未受診率，精検未受診・未把握率，要精検率，がん発見率，陽性反応適中度を用いて整理した．

くり返しとなるが，がん検診は健康な人を対象としており，がん検診におけるがんの早期発見・早期治療という利益の他に，偽陽性，偽陰性，偶発症，受診者の身体的・心理的な負担等の不利益を考慮することが重要である．さらに，がん種によっては成長が緩慢なもの，途中で進展が滞るものや消退する病変もあり，こうした病変を発見し，治療することは，受診者にとって利益とならないばかりか過剰診断・過剰治療をまねくことにつながる．そこで，ここでは科学的根拠に基づくがん検診の実施だけでなく，がんによる死亡率減少効果が不明確である指針に基づかないがん検診の実施状況についても併せて整理をした．ただし，今後研究が進むにつれて科学的根拠に基づくがん検診についての指針は見直される可能性がある．

指標1a：指針に基づかないがん検診を実施している市区町村の割合

指針に基づかないがん検診（指針以外のがん種のがん検診）を実施している市区町村の割合は，平成19年度では62.9％であったのに対し，平成24年度では77.3％と増加傾向にあった．特に，前立腺がんのPSA検査を実施していると答えた市区町村が74.9％と多数を占めた（平成24年度がん検診実施状況）．

指標1b：指針に基づくがん検診を実施している市区町村の割合

集団検診，個別検診のいずれかにおいて指針に基づくがん検診を実施している市区町村の割合は，平成24年度がん検診の実施状況をみると胃

表2　がん予防重点健康教育およびがん検診実施のための指針で推奨される検診

種類	検査項目	対象者	受診間隔
胃がん検診	問診，胃部X線検査	40歳以上	年1回
肺がん検診	問診，胸部X線検査，喀痰細胞診	40歳以上	年1回
大腸がん検診	問診，便潜血検査	40歳以上	年1回
乳がん検診	問診，視診，触診，乳房X線検査（マンモグラフィ）	40歳以上	2年に1回
子宮頸がん検診	問診，視診，子宮頸部の細胞診，内診	20歳以上	2年に1回

がん(胃X線検査)が99.1%,肺がん(肺X線検査)が96.0%,肺がん(喀痰細胞診)が85.8%,大腸がん(便潜血検査)が99.9%,乳がん(乳房X線検査)が99.0%,子宮頸がん(細胞診・従来法)が90.8%とほぼ全ての市区町村で指針に記載されているがん検診が実施されていた.

2◎市区町村におけるがん検診の精度管理実施状況

がん検診の実施において期待されるがん死亡率減少効果を得るためには,科学的根拠に基づく検診を適切な精度管理下で実施することが求められる.平成20年の厚生労働省がん検診事業の評価に関する委員会の報告書「今後の我が国におけるがん検診事業評価の在り方について」[4]では,がん検診の精度管理を「目標と標準の設定」「質と達成度のモニタリング・分析」および「改善に向けた取組」の三段階に整理し,精度管理の指標と目標を設定し,モニタリングし,改善を目指すこととされた.がん検診の精度管理においては,がん死亡率の減少効果(アウトカム指標)を評価するためには相当な時間を要することから,「技術・体制的指標」と「プロセス指標」の側面から把握することとした.

・技術的・体制的指標

指標2:「事業評価のためのチェックリスト」を実施している市区町村の割合

「事業評価のためのチェックリスト」は,平成20年の厚生労働省がん検診事業の評価に関する委員会の報告書「今後の我が国におけるがん検診事業評価の在り方について」[3]で示されており,その実施状況については厚生労働省研究班による「市区町村におけるがん検診チェックリストの使用に関する実態調査」[5]で報告されている.この調査の対象は,健康増進事業に基づく集団検診を「がん予防重点健康教育及びがん検診実施のための指針」[3]に基づいた検査方法で行っている市区町村である.平成26年度調査では,1,736市区町村を対象に調査された(回収率80.5%).

結果,チェックリストの総合実施割合*注釈は,胃がんが71.8%,肺がんが71.9%,大腸がんが71.3%,乳がんが69.7%,子宮頸がんが67.8%であった.受診者数,要精検率,精検受診率,発見率,陽性反応適中度については,「性・年齢階級別」「検診機関別」「受診歴別」に集計している市区町村は,年々増加傾向にあった.精検受診率について「性・年齢階級別」に集計している市区町村はどのがん種においても約77%,発見率の「性・年齢階級別」に集計している市区町村は,約63%であった.検診機関の質を担保するため委託検診機関の選定時に約57～60%の市区町村では仕様書を取り交わしていた.しかし,その仕様書に必要最低限の精度管理項目4が記載されている市区町村は,近年増加傾向にあるものの依然42～44%にとどまっている.

・がん検診のプロセス指標

指標3:精検受診率,精検未把握率,精検未受診率,精検未受診・未把握率,要精検率,がん発見率,陽性反応適中度

胃がん(胃X線検査),肺がん(胸部X線検査と喀痰検査(高危険群のみ)の併用),大腸がん(便潜血検査),乳がん(視触診とマンモグラフィの併用),子宮頸がん(細胞診)について,精検受診率,精検未把握率,精検未受診率,精検未受診・未把握率,要精検率,がん発見率,陽性反応適中度について表3に示す.胃X線検査のように市区町村において達成されるべき最低限の基準(許容値)3を達成しているものもあるが,達成されていないものもある.また,ここで示した数値は,全国平均値であり市区町村によって許容値を達成している市区町村もあれば達成できていない市区町村がある点に留意する必要がある.受診率,精検未把握率等について,正確に市区町村の割合を算出するためには,少なくとも前述の「事業評価のためのチェックリスト」の項目が実施されている必要がある.つまり,がん検診における精密検査把握率が高まることで初めて各市区町村のがん発見率や陽性反応適中度が正確に把握可能となる.近年,「事業評価のためのチェックリスト」の実施割合は向上してきており,ようやくがん検診のプロセス指標の動向について検討できる段階に入ったともいえ,がん検診のプロセス指標については,今後の動向を把握していく必要がある.