# The Evaluation of Cancer Screening

## Concepts and Outcome Measures

Stephen W. Duffy, MSc[a,*], Robert A. Smith, PhD[b]

**KEYWORDS**

• Screening • Cancer • Evaluation • Methodology

**KEY POINTS**

- Cancer screening evaluation is a specialist area of healthcare evaluation, requiring specific skills and methods.
- Evaluation may have different purposes, including proof of principle, quality control of screening services, or assessment of innovative screening technology.
- Methods of evaluation will depend on both the purpose and the primary object of screening (prevention or early detection).

### INTRODUCTION

Before considering evaluation of cancer screening, we should probably describe what cancer screening is, and before that we should define medical screening more generally. An eloquent and useful definition of screening has been given by Wald as "… the systematic application of a test or enquiry to identify individuals at sufficient risk of a specific disorder to warrant further investigation or direct preventive action, amongst persons who have not sought medical attention on account of symptoms of that disorder."[1]

The preceding definition is clearly very general and can cover a wide range of investigations, conditions, and mechanisms of action. However, one point on which it is very specific is the population to which the screening is applied: persons who have not sought medical attention on account of symptoms of that disorder. If a test is applied to persons who have sought medical advice for symptoms, this is not screening, it is diagnosis.

Cancer screening encompasses a wide range of investigations, aims, and mechanisms of achieving those aims. Major potential screening investigation strategies include:

[a] Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK; [b] American Cancer Society, 250 Williams Street, Atlanta, GA 30303, USA
* Corresponding author.
*E-mail address:* s.w.duffy@qmul.ac.uk

Med Clin N Am 104 (2020) 939–953
https://doi.org/10.1016/j.mcna.2020.07.002 **medical.theclinics.com**

- Imaging: for example radiographic mammography for breast cancer, low-dose computed tomography for lung cancer.
- Examination of exfoliated cells: examples include cervical smears, now largely replaced or being replaced by human papillomavirus testing.
- Visual examination: examples include unassisted visual examination of the skin for atypical nevi or melanoma, and more invasive approaches including colonoscopy for colorectal cancer.
- Biomarkers: for example circulating markers of disease, such as prostate-specific antigen (PSA) for prostate cancer.
- Palpation: examples include clinical examination of the breasts, and digital rectal examination.

This list is by no means exhaustive, but gives an idea of the range of potential cancer screening tests. There is similarly a range of clinical aims and mechanisms. For instance, mammography screening for breast cancer is aimed at detecting breast cancer at an earlier stage when treatment is more likely to be successful, compared with when breast cancer is diagnosed symptomatically.[2] Colonoscopy and sigmoidoscopy, on the other hand, are aimed primarily at detecting precancerous adenomas, removing them, and thus preventing them from progressing to cancer at all.[3]

One further point to note with respect to cancer screening: in general, the screening test does not diagnose the cancer. It generally identifies those who need further investigation. To take the example of breast cancer screening, a positive screening mammogram is not a diagnosis of breast cancer. In the National Health Service Breast Screening Programme in the United Kingdom, on average, 1 in 5 women recalled for further investigation following a suspicious screening mammogram actually have breast cancer. The test is not expected to distinguish perfectly between those who do and do not have the disease, but the extent to which it does is an important ingredient in evaluation.

In this article, we review the main tools available for evaluation of cancer screening, in terms of the following:

1. Proof of principle: does the screening prevent mortality or significant morbidity from the cancer?
2. Service evaluation: is a routine service screening program (ie, screening in the community) delivering the expected clinical benefits?
3. Program quality: is a screening program meeting standards of test accuracy, punctuality, minimization of screening side effects, and so forth?
4. Innovation: should an existing screening program change to a new technology?

## CANCER SCREENING EVALUATION TECHNIQUES: PROOF OF PRINCIPLE

### *Randomized Trials of Screening*

Cancer screening as a public health activity is not a case-finding exercise. Its role is to prevent premature mortality or significant morbidity from the cancer in question. A major task of cancer research is to design studies that will inform policy makers as to whether it does so. Let us first take the case in which the screening aims to detect cancer, but at an early stage, when treatment is more likely to be successful in preventing death from the disease.

At this point, we briefly mention the 2 classic biases, *lead time bias* and *length bias*, which screening reviews perennially discuss, but which have been known about for decades.[4] With respect to lead time, if screening is successful in detecting cancer early, it necessarily confers an increase in the time from diagnosis to death, that is,

an increase in survival time. This would occur whether or not the screening prevented or delayed death from the disease in question. Length bias refers to the phenomenon whereby comparison of outcomes between screen-detected and symptomatic cancers is biased by the fact that less aggressive tumors are likely to grow more slowly and therefore have a longer window of opportunity for screen detection.

It should be noted here that the preceding does not mean that lead time is a bad thing: for screening to be effective, *lead time is essential*. Nor does it mean that survival analyses and comparison of screen-detected with symptomatic cancers are uninformative: it simply means that they do not prove that screening works in principle.

So how *do* we establish the effectiveness or otherwise of cancer screening interventions in principle? As with most medical interventions, the design of choice is the randomized trial: we randomize one population to receive the intervention (or rather be offered the intervention) and another to usual care. If, as for mammography screening for breast cancer, or fecal occult blood testing for colorectal cancer, the aim is to detect cancer at an early stage and prevent death from the disease, then the appropriate trial endpoint is death from the disease, offset by the total populations randomized to each group, whether the intervention group members took up the offer of screening or not, and regardless of whether the persons randomized developed the cancer in question or not. The time origin should be the point of randomization (not the point of diagnosis of cases, as in survival analysis).

This basic design avoids the classic biases mentioned previously, and as the comparison is of the randomized groups whether or not they were actually screened, it avoids self-selection issues. An example is the Swedish Two-County Trial of mammographic screening. The subjects were randomized to the offer of regular mammography screening, or not, over a period of approximately 7 years, and followed up for a total of 29 years for mortality from breast cancer.[5] Results are shown in **Table 1**. The table shows a significant 31% reduction in breast cancer mortality with the offer of screening. The investigators converted this to an absolute effect of 1 breast cancer death prevented per 1344 mammographic examinations, or per 414 persons screened 3 times over a period of 7 years.[5] We remark that this was the final follow-up of the Two-County Trial, cited here as it is most relevant to the calculation of absolute benefit. The relative benefit has remained constant since the initial publication of mortality results in 1985, which informed screening policy in many coiuntries.[6]

It should be noted that although the randomized trial design as described avoids the anticonservative biases of lead time and length bias, it is inherently conservative. In the first instance, substantial noncompliance with screening dilutes the effect. In the Two-County Trial, compliance was relatively high, approximately 85%, but this still means that the effect of screening is diluted by the 15% who did not receive screening and presumably did not receive any mortality reduction as a result. Thus, although randomized trials measure the efficacy of screening, they often do not provide an

**Table 1**
**Primary result of the Swedish Two-County Trial of mammographic screening at 29-year follow-up**

| Trial Group | Subjects Randomized | Breast Cancer Deaths | RR (95% Confidence Interval) |
|---|---|---|---|
| Intervention | 77,080 | 351 | 0.69 (0.56–0.84) |
| Control | 55,985 | 367 | 1.00 (–) |

*Abbreviation*: RR, relative risk.

accurate estimate of the effectiveness of screening among the population that actually undergoes screening.

A second issue is that to measure the full benefit of screening, a trial would have to follow up the entire study population to death, which is not feasible and would not deliver a sufficiently timely result. However, there needs to be a minimum follow-up period; put in stark terms, the duration of the trial has to be long enough for cancers in the control group first to come to symptomatic attention and thereafter to cause death. This is illustrated in **Fig. 1**, which shows a cancer in the intervention group of a trial of effective screening, and its equivalent cancer in the control group. In both cases, the tumor is "born" in year 2. In the intervention arm, it is detected before any symptoms by screening in year 4, treated successfully, and the host goes on to live for 21 years afterward and dies of other causes in year 25. In the control group, the corresponding cancer is diagnosed symptomatically in year 7, treatment is unsuccessful and the host dies in year 11. The point is that this represents a cancer death prevented by screening, but with only 10 years of follow-up from randomization it would not be observed. The longer the follow-up, the fewer such unobserved benefits.

There is a related cause of underestimation of benefit. Screening trials in general offer the intervention for a relatively short period of time, usually less than 10 years, and in some cases less than 5.[7] The cancers diagnosed during this screening phase, in both trial groups, are followed up thereafter for death, specifically from the cancer in question. Under the principle of randomization, without any screening, we would expect rates of diagnosis in both trial groups to be parallel over time. However, in the presence of screening in one arm and usual care in the other, some cancers that subsequently cause death are diagnosed in the control group after the end of the screening phase, but during the screening phase in the intervention group due to lead time. Deaths from these cancers will be included in the intervention group but deaths from their counterparts in the control group will not be included. This will bias the result against the screening. Duffy and Smith[8] showed that this bias can be partially corrected by offering an exit screen to the control group contemporaneously with the final screen of the intervention group. This design was used in the Swedish Two-County and Gothenburg Trials.[5,9]

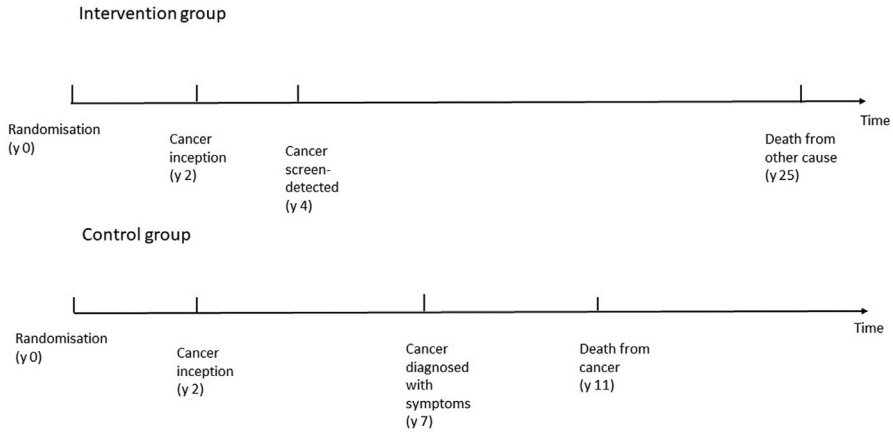


**Fig. 1.** Illustration of potential timescale of prevention of cancer death in a trial of screening.

When screening for a precursor lesion with the objective of preventing cancer altogether, the randomized trial remains the design of choice. The 2 main differences are that the exit screen of the control group is no longer necessary, and of course the endpoint is cancer diagnosis rather than cancer death, as in the UK Flexible Sigmoidoscopy Trial for prevention of colorectal cancer.[3]

### *Deaths from Other Causes*

It is sometimes argued that screening should show a significant effect on all-cause mortality to inform policy,[10] or it is implied, such as when one reads in a systematic review that a 30% reduction in disease-specific mortality was observed, "*but there was no reduction in all-cause mortality.*" Consideration of a simple example shows that the focus on all-cause mortality is ill-considered. Let us take the example of ovarian cancer, which might be responsible for approximately 4% of all deaths in a typical middle-aged female population. Suppose the effect of the offer of ovarian cancer screening is to reduce ovarian cancer mortality by 20%, without affecting deaths from other causes. In a very large trial with 100,000 all-cause deaths expected in the control group, the expected number of deaths in the study group would be 99,200 ($0.04 \times 0.2 \times 100{,}000 = 800$). Thus, the expected all-cause mortality relative risk would be 0.992 with a 95% confidence interval of 0.9834 to 1.0008; that is, even with 100,000 expected all-cause deaths in each arm, the error bars completely swamp the effect of the screening. A study with 300,000 all-cause deaths expected in each arm would arguably be powered for this effect. Does this mean that to evaluate ovarian cancer screening, we need a trial with 12 million women, 6 million in each arm and follow-up such that 5% in each arm die from any cause? No, it means that the effect of ovarian cancer screening on all-cause mortality is essentially unverifiable. The answer is surely to have cause-specific death from the cancer in question and from the sequelae of screening or treatment for that cancer as the endpoint, and to adopt very rigorous cause of death determination policies, with a high rate of autopsy if necessary.

One can see the absurdity of the advocacy of all-cause mortality if one applies the philosophy in other areas, such as seat belt legislation or road speed restrictions, traveler vaccinations, migrant animal quarantine, and so on.

The use of all-cause mortality is sometimes advocated on the grounds of objectivity. Apart from the fact that human judgment is required in all areas of medicine and health, one might comment that its use discards so many things that we know. These include that only those with a cancer can die of it, only those irradiated can die of a radiation-induced disorder, and so on. The way to effective evaluation is to use our knowledge in trial design, not to throw it away. The evaluator also can use the technique of excess mortality analysis, which compares the excess deaths in cancers diagnosed in the intervention group over the death rate in the population at large, with the corresponding excess morality in cancers diagnosed in the control group, without classifying deaths by cause at all. This was used in the overview of Swedish breast screening trials, and confirmed the reduction in cause-specific mortality.[11]

In noting that all-cause mortality is an inappropriate endpoint in public health interventions generally, Sasieni and Wald[12] acknowledge that the question of whether the intervention under investigation does increase the risk from other causes of death is a valid one. In the first instance we assess whether the screening has the desired effect on the primary endpoint, death or incidence of the cancer in question. It is then reasonable to ask whether the screening is associated with an increased risk of other causes of death. The difficult question is how to elicit this?

First, one would check whether there was a significant or suggestive effect of the intervention on deaths from all other causes than the cancer in question, in the entire population. One might be tempted to do the same for a substantial number of individual causes of death, but this would be mistaken. If one tested 20 causes of death, one would expect one spurious result at the 5% significance level. Instead, consider the preceding remarks about what we know. For example, advocates of use of all-cause mortality for breast screening trials suggest that the reduction in breast cancer deaths may be compensated for by deaths from more frequent use of radiotherapy in screen-detected cancers or as adverse effects of treatment in larger numbers of cancers treated due to larger numbers of cases found in the screened arm.[13] To address this issue, the obvious answer is to compare deaths from other causes between the 2 groups **in the cancer cases only**, or to carry out an excess mortality analysis, as in the Swedish overview.[11]

As a general strategy in a trial of screening to prevent mortality from an individual cancer, we would therefore suggest the following:

1. Compare mortality from the cancer targeted.
2. Compare mortality from all other causes combined.
3. Compare excess mortality from the cancers diagnosed.
4. Testing for differences between groups as a whole in specific causes of death should be done only for plausible, protocol-specified hypotheses.

### *Other Endpoints Including Overdiagnosis*

Other endpoints addressed in screening trials include incidence of the cancer in question, rates of various treatment modalities, and psychological outcomes. It is beyond the scope of this article to specify detailed analyses for these, but some observations should be made here.

One adverse effect of screening that has received considerable attention is overdiagnosis. The most common definition of this is the diagnosis of a histologically confirmed cancer as a result of screening that would not have been diagnosed in the patient's lifetime if screening had not taken place. In the past, this has been estimated by crude comparison of incidence between the intervention and control groups.[13,14] There are major problems with this approach. These include the phenomenon of lead time.[15] An excess may be observed between intervention and control groups, but a portion of this will be due to cases diagnosed in the intervention group whose counterparts in the control group will be diagnosed in the future but have not been diagnosed yet. This portion represents early diagnosis, not overdiagnosis, but it is often included in the latter.

This is eloquently illustrated by the European trial of PSA screening for prostate cancer.[16–18] In this trial, 77,890 men were randomized to periodic PSA testing and 89,353 to usual care. **Table 2** shows incidence results at 11, 13, and 16 years' follow-up. The excess number of cancers in the intervention group reduces over time, as the control group "catches up" by diagnosis of cancers that would have been detected years earlier in the intervention group. It should be noted that the numbers of prostate cancer deaths prevented increases with follow-up time, so that the numbers of excess cases per life saved at the 3 follow-up points are respectively 41, 22, and 18. This illustrates that too short an observation period will underestimate the benefits and overestimate the harms of screening.

This excess due to lead time can also induce an artificial excess in treatment modalities. The implication is not that these comparisons cannot be made, but that they should be either mathematically adjusted for lead time or at the very least, interpreted in the light of lead time.

**Table 2**
**Prostate cancer incidence in the European trials of prostate-specific antigen screening by follow-up time**

| Follow-up | Study Group | Subjects | Prostate Cancer Cases (Rate/1000) | Excess Cases in Intervention Group |
|---|---|---|---|---|
| 11 y | Control | 89,353 | 4307 (48) | — |
| | Intervention | 77,890 | 5990 (77) | 2251 |
| 13 y | Control | 89,353 | 6107 (68) | — |
| | Intervention | 77,890 | 7408 (95) | 2111 |
| 16 y | Control | 89,353 | 7732 (87) | — |
| | Intervention | 77,890 | 8444 (108) | 1668 |

Also, as noted previously, some trials have an exit screen of the control group. Even when analysis is limited to trials that did not (nominally) screen the control group at the close of the screening phase, and consider long follow-up for which lead time is less of an issue,[14] there remain methodological issues of design of the trials, which detract from the validity of a simple comparison of incidence.[19]

There is a final reason why incidence from the randomized trials may not be useful for estimating absolute rates of overdiagnosis or of cancers treated by certain modalities. The trial populations are unlikely to be representative of the general population targeted for screening by routine services, and due to the timescale issues mentioned previously may be characterized by incidence rates of past decades. Although relative benefits in terms of mortality may usually be generalized, absolute rates of incidence cannot. For further suggestions with respect to overdiagnosis, see the next section.

## SERVICE SCREENING EVALUATION

Here we consider the task of assessing whether a routine screening program is delivering the expected benefit, and the estimation of one of the major publicly expressed concerns about screening, overdiagnosis.

### *Estimation of Benefit*

The major benefit of screening is either reduction in mortality from disease, as in breast cancer screening, or reduction in incidence, as in endoscopic screening, aimed at detecting and removing adenomatous polyps to prevent progression to invasive colorectal cancer, or cervical screening, aimed at detecting and removing cervical intraepithelial neoplasia and thus preventing progression to invasive cervical carcinoma. A similar range of observational methodologies is available for both. Essentially there is a choice of cohort or case-control approaches, with different tactical methodological choices available within both.

For cohort approaches, assuming that mortality or incidence can be ascertained, 2 issues have paramount importance. The first is to have a source of a counterfactual estimate of what the mortality or incidence would have been if the screening had not taken place. The second is the need for to ensure accurate ascertainment of exposure to screening.

To obtain counterfactual estimates in a nonrandomized setting, there are sometimes geographic comparator groups available, as when Copenhagen introduced mammography screening before the rest of Denmark.[20] More often, however, data are available

only on a single region or country when the screening was introduced universally in a relatively short period. In this case, we have the choice of historical comparison (before-after), or contemporaneous comparison of those accepting the offer of screening with those not doing so. The former is confounded with temporal changes; for example, in treatment of the disease, and the latter is prone to self-selection bias, whereby those who choose to be screened are at different risk of dying of the cancer in question than those who do not. Methods are available to deal with both, including comparison of those unscreened before the screening with those unscreened due to declining the offer of screening in the screening era, and formal mathematical correction for the self-selection bias.[21,22] The main message here is to be aware of these potential biasing features and adopting design or analytical methods to reduce their effect.

For the second issue, the problem is not simply to ensure accurate data on screening invitation and attendance, although this is clearly necessary. It also requires linkage of mortality with data on date of diagnosis. Consider a screening program for prostate cancer that starts in the year 2000 and a prostate cancer death in 2004. The survival figures for prostate cancer mean that in all probability that cancer was diagnosed before 2000, that is, before screening was available. To correctly classify exposure to screening in observational cohorts, a powerful tactic is to define the endpoint as "refined," or incidence-based mortality, that is, deaths from cancers diagnosed *after* the introduction of screening.[22] More recently, an interesting variant on this has been used in both breast and prostate cancer, that is, the incidence of cancers subsequently proving fatal within a certain period of diagnosis.[23,24] This has the added advantage of correctly classifying the exposure status of the population denominator at the relevant time, that is, the diagnosis year, in addition to the exposure status of the cases with the endpoint.

The case-control approach essentially works as follows: cases are persons with the endpoint (for example, death from breast cancer, diagnosis of invasive cervical carcinoma), and controls are persons without the endpoint, matched for age, sex, and opportunity for screening. The cases and matched controls are then compared with respect to screening exposure before the diagnosis dates of the cases. The rationale is that if screening is preventing deaths or diagnoses, the cases will be characterized by lesser screening exposure history than the controls. This design is often less resource-intensive and facilitates straightforward individual classification of screening exposure, but is equally subject to potential self-selection bias and may have other biases related to retrospective identification of cases and ascertainment of exposure.

The case-control evaluation has frequently been used for breast cancer screening,[25] but arguably, the paradigmatic example is the UK case-control evaluation of cervical cancer screening.[26] In this study, 1305 cases of invasive cervical cancer were compared with 2532 age-matched disease-free controls. This study showed no benefit of screening in women younger than 25 and demonstrated a longer-lasting protection of a screen at older ages. This informed the age limits and interscreening intervals in the national program in the United Kingdom. The program changed the lower age limit to 25 and instituted 3-yearly screening for women younger than 50 years old, and 5-yearly for older women.

The case-control approach is therefore an attractive and potentially powerful one. However, researchers adopting this approach should be aware of a number of complicating factors:

- Only screening before the date of diagnosis of the case is relevant. Thus, controls are given a pseudodiagnosis date, equal to the date of diagnosis of their matched case.[26]

- Self-selection bias, and methods for correction for this.[27]
- Screening opportunity bias: if the screen at which a case is detected is included as exposure, the result underestimates the benefit of screening, whereas if it is excluded, the result overestimates the benefit.[28] The true effect will be likely between the two, and a sensitivity analysis may be done by adding a potential lead time to the pseudodiagnosis date of the controls.[29]
- Ascertainment issues: there may be differential identification of cases and controls by screening history. The remedy for this is high-quality cancer registration and screening data, and vigilance against the possibility of ascertainment bias.

### *Overdiagnosis*

Potential adverse effects of screening include discomfort or embarrassment from the test, radiation exposure from the test or subsequent examinations, investigations following suspicious screening results in screenees who turn out not to have cancer, anxiety about cancer, and overdiagnosis. The last of these has received most attention in recent years. As noted previously, a common definition is the diagnosis as a result of screening of cancer that would not have been diagnosed in the host's lifetime if screening had not taken place. Because it is not possible to distinguish histologically a truly nonprogressive cancer from one that is progressive, rates of overdiagnosis commonly are estimated by comparing incidence rates in a group that underwent screening with a group that did not.

With the preceding definition, at least some overdiagnosis must occur in the case of screening to detect frank cancer at an earlier stage. We cannot have successful screening without lead time, and due to competing risks of death from other causes, we cannot have lead time without overdiagnosis. Some who have undergone screening will die shortly thereafter unexpectedly, whereas others' deaths were anticipated, and yet a referral for screening was made without consideration of the lack of potential benefit.

Overdiagnosis can be expressed in several different ways, which will give different impressions to both health professionals and potential screenees.[30] It is probably fair to say, however, that to the cancer scientist, the interesting measure is the proportion of screen-detected cancers that are overdiagnosed, whereas to the person invited to screening and the provider of screening, the more relevant measure is the absolute population risk of an overdiagnosed cancer.

Various approaches have been adopted to estimate overdiagnosis in the context of screening services in routine health care. Many of these depend on comparison of incidence of cancer in the context of a screening program, compared with a counterfactual incidence, estimated from historical data. This has 2 major problems, best considered in the context of mammography screening. The first is that in the late twentieth century, when mammography programs were being set up in many countries, incidence of breast cancer was on the increase because of changes in reproductive behavior, body habitus, and other risk factors. The second is the issue of lead time mentioned previously. Puliti and colleagues[31] showed that studies that failed to take account of these complicating factors obtained high estimates of overdiagnosis and studies that took account of them resulted in low estimates, the latter being more reliable in theory and more plausible in practice.

Another point to note is that when overdiagnosis estimation is driven primarily by incidence of disease, long-term observation is necessary. Consider the example of the prostate screening trial in **Table 2**. The longer the observation period, the smaller the excess. This also can be illustrated by **Fig. 1**. The cancer in the intervention group is clearly not overdiagnosed, because its corresponding cancer in the control group

was diagnosed 3 years later. However, if we had only 5 years of observation, the cancer in the intervention group would be considered as excess and potentially overdiagnosed.

It is possible with detailed screening data, to posit a statistical model of overdiagnosis, involving a heterogeneous tumor population regarding capability of progression to symptomatic disease.[32] This, however, is mathematically and computationally complicated, and it is not unusual for the statistical estimation tool to fail to find a plausible or precise estimate.[32]

The important principle to bear in mind is that simple comparison of incidence in a screened (or invited) population with that in an unscreened population, as described previously, is not a valid estimate of overdiagnosis. Accurate estimation requires taking account of underlying incidence trends and lead time, and requires long-term observation.

In the case of screening for precursor lesions, it is not clear that overdiagnosis is a meaningful concept. Many cervical or colorectal precursors would never have become cancer if left untreated, so in that sense it could be argued that they are overdiagnosed. However, the diagnosis of an adenoma in the colon or a case of cervical intraepithelial neoplasia of grade 2 do not incur treatment beyond well-tolerated excision in an outpatient setting, and do not have the same life-changing effects as a diagnosis of cancer. Thus, overdiagnosis in this context is not of significant public health interest.

## ASSESSMENT OF QUALITY OF SCREENING

### *The Screening Test*

The primary measures of a screening test's quality are its sensitivity and specificity. These are best shown by example. **Table 3** shows the results of fecal immunochemical testing (FIT) in 3211 subjects with a threshold of 10 μg of hemoglobin per gram of feces.[33] All subjects also underwent colonoscopy, which was treated as the gold standard of diagnosis in this study, and were classified as positive for advanced neoplasia (cancer or advanced adenoma) or not.

The sensitivity of the test is the probability of a positive result in those who actually have the disease. Of the 311 subjects with advanced neoplasia, 163 had a positive FIT result. This gives an estimated sensitivity for FIT at this threshold of 163/311 = 52%.

The specificity of the test is the probability of a negative result in those who truly do not have the disease. Of 2900 subjects free of disease in this study, 2552 had a negative FIT result. Thus the specificity is estimated as 2552/2900 = 88%.

In addition to this, we can calculate the positive predictive value of the test, that is, the proportion of subjects with a screen positive result who actually have the disease. In this case, it is estimated as 163/511 = 32%.

**Table 3**
**Results of fecal immunochemical testing (FIT) in relation to colonoscopy findings**

| | Colonoscopy Result (Gold Standard) | | |
|---|---|---|---|
| FIT | No Advanced Neoplasia | Advanced Neoplasia | Total |
| < 10 μg/g | 2552 | 148 | 2700 |
| ≥ 10 μg/g | 348 | 163 | 511 |
| Total | 2900 | 311 | 3211 |

Clearly, a good screening test should have both a high sensitivity and a high specificity. The positive predictive value will depend on the prevalence of the disease in the population tested. A higher-risk population will show a higher positive predictive value.

Intuitively, a higher threshold for FIT in the preceding example would improve specificity at a cost of a loss of sensitivity. Similarly, a lower threshold would yield higher sensitivity and lower specificity. When the test can be expressed as a continuum, as in this case, the combined positive and negative accuracy can be described by a receiver operating curve (ROC). This is a plot of the sensitivity against 1-specificity for the points on the continuum. **Fig. 2** shows the ROC curve for estimated risk of lung cancer from the Liverpool Lung Project, which was used to determine eligibility for low-dose computed tomography screening for lung cancer in the UK Lung Screening Trial.[34]

The diagonal line shows what one would expect if the test had no diagnostic value at all. Clearly the closer the curve is to the left-hand side and the top of the area, the closer the sensitivity and specificity are to 100% and the better the test. The accuracy can be summarized by the area below the curve, in this case 0.72, and again, the closer this area is to 1, the better the test. However, the reduction to a single dimension of accuracy is not necessarily useful. If the diagnostic workup is invasive and potentially harmful, or limited by staff capacity, one might require a minimum specificity and tolerate whatever sensitivity this entailed. On the other hand if the consequences of missing a case were crucial, one might specify a minimum sensitivity. In these cases, a single summary statistic of accuracy is not of particular use.

### *Quality of a Screening Program*

A health care provider delivering a screening program will wish to monitor the quality of that program. The parameters monitored may pertain to the diagnostic quality of the

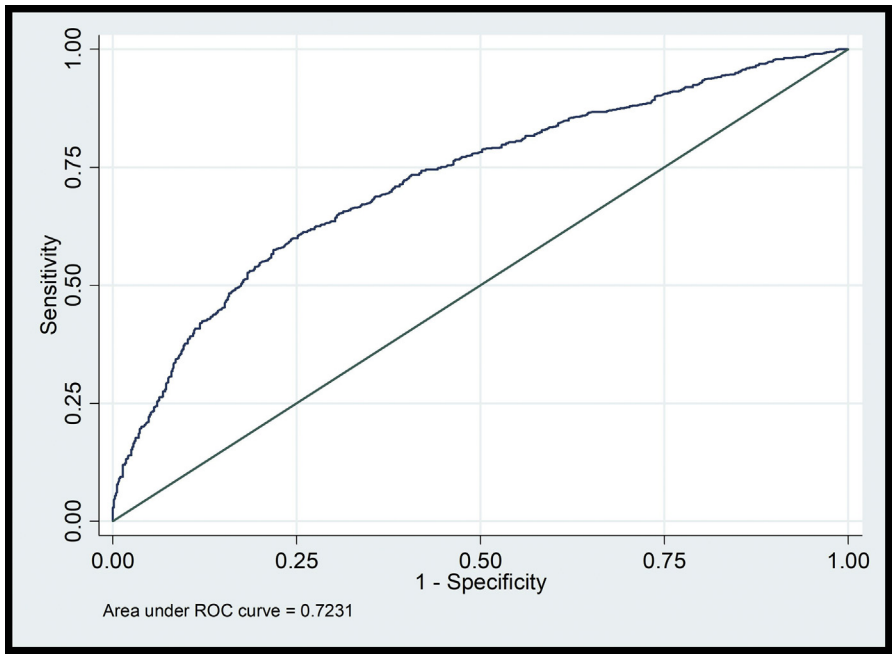


**Fig. 2.** ROC curve of estimated risk from the Liverpool Lung Project Model.

**Table 4**
**Selected standards and achievable targets for the UK's NHS Breast Screening Program**

| Parameter | Standard to Be Achieved | Achievable Target |
|---|---|---|
| Coverage of target population aged 50–70, % | ≥70 | ≥80 |
| Number referred for assessment (first screen), % | <10 | <7 |
| Number referred for assessment (subsequent screens), % | <7 | <5 |
| Screen result notification within 2 wk, % | >95 | Not stated |
| Time to assessment within 3 wk, % | >98 | 100 |
| Benign surgical biopsy rate (first screen) | <1.5/1000 | <1/1000 |
| Benign surgical biopsy rate (subsequent screens) | <1/1000 | <0.75/1000 |
| Proportion of cancers with preoperative diagnosis, % | ≥90 | ≥95 |
| Standardized cancer detection ratio[a] | >1.00 | >1.40 |
| Interval cancers within 12 mo | < 0.65/1000 | Not stated |

[a] Age-standardized to expected rates from the Swedish Two-County Trial on which the UK program is based.

screening provided, efficiency of the service, or both. Table 4 shows selected parameters monitored in the UK's National Health Service (NHS) Breast Screening Program, with the standards to be achieved (failure to do so generating remedial action) and achievable targets for units to aim at.[35]

As can be seen, some of the standards are aimed at quality of the screening test (minimum standardized detection ratios), some at quality and acceptability of the program delivered (minimum times to results and assessment appointments), and some to both (maximum percentages recalled for assessment). These are a small sample of many standards applied to the program, which include technical radiographic and radiologic parameters, staging of cancers diagnosed, separate invasive and noninvasive cancer detection rates and standardized detection ratios of invasive cancers of size smaller than 15 mm.

The point of standards such as these is to achieve very high-quality, maximizing benefits and minimizing harms. These are ethically as well as practically important, as screening is not a service requested by the patient. It is offered to the healthy, asymptomatic population as a public health measure, and therefore must be able to guarantee minimum performance for those who take up the offer.

## INNOVATIONS TO EXISTING SCREENING PROGRAMS

Diagnostic technology moves faster than the research community can evaluate, thus there is a need for rapid evaluation of innovations in screening practice or technology. As noted previously, the strongest evidence for the efficacy of screening comes from a randomized trial with the clinical endpoint that the screening is intended to prevent as the outcome variable. However, cervical screening has been widespread for decades without such evidence and is generally agreed to have prevented very large numbers of cervical cancer cases and deaths. Also, when proof of principle has been established by one early detection modality, does a potential improvement brought about by a new test need evidence of the same research design?

From changes that have already been made to screening programs, there is an evident consensus that it does not. The changes from film to digital mammography,

**Table 5**
**Detection of 59 cancers by 2D mammography and integrated 2D and DBT in the STORM trial**

| 2D Mammography Result | 2D + DBT Result | | |
|---|---|---|---|
| | Positive | Negative | Total |
| Positive | 39 | 0 | 39 |
| Negative | 20 | 0 | 20 |
| Total | 59 | 0 | 59 |

*Abbreviations:* DBT, digital breast tomosynthesis; 2D, 2-dimensional.

guaiac fecal testing to immunochemical testing, and from Pap smear to human papillomavirus testing, have all been made in numerous screening programs without the necessity for a randomized trial with death from or incidence of cancer as the endpoint.

When a technological innovation is proposed, there are a number of designs available, but one of the most powerful is the split-sample design, in which all participants receive both the standard screening test and the innovation. The name refers to the use of the design in evaluating a new blood test, in which each participant's blood sample is split into 2 aliquots, 1 to receive the old test, 1 the new. The advantages of this design include the improved precision of within-screenee over between-screenee comparisons, which in turn confers the required statistical power with a smaller study size, and the inbuilt control for personal confounders and center effects. The latter can be particularly important in screening trials in which the participating centers may achieve different screening accuracies.

A good example is the STORM trial of integration of digital breast tomosynthesis (DBT) into breast cancer screening.[36] In this study, 7292 women in Trento and Verona, Italy, were screened with both 2-dimensional digital (2D) mammography, and integrated 2-dimensional mammography and DBT. There were 59 cancers detected. The detection modes of these are shown in **Table 5**. No cancers were detected by 2D mammography alone and not by integrated 2D + DBT. Of the 59 cancers, none were detected by 2D alone, 39 were detected by both modalities, and 20 by integrated 2D + DBT alone.

The formal statistical comparison in this design is between the disagreements, that is, of the 20 cases detected by integrated 2D + DBT alone versus the zero cases detected by 2D alone. This shows a greater detection rate for integrated 2D + DBT, which is highly statistically significant. To achieve 90% statistical power for this difference in detection rates with a comparative trial in which half the subjects received one modality and half the other, would require 37,000 screenees in all, a 5 times greater study size.

Thus, although other designs are available, the split-sample study should always be considered when evaluating new screening technology or other changes to existing programs. It can be incorporated pragmatically within the program, it is fast, efficient, and usually more affordable than alternative designs.

## SUMMARY

The preceding has summarized the major considerations and methodological approaches for evaluation of cancer screening. If there is one overriding message for the reader to appreciate, it is that cancer screening evaluation *is not easy*. One cannot approach the subject with only the usual epidemiologic tools. In particular, the shift in

the timescale of tumor diagnosis, treatment and potential progression or recurrence adds a degree of complexity to the task. There are, however, clear principles, and sometimes clearly superior approaches, such as the split-sample design for innovations to existing screening programs. However, as in other walks of life, if we find we are getting quick and easy answers, we should always ask ourselves: are we doing something wrong?

## REFERENCES


1. Wald NJ. Guidance on terminology. J Med Screen 1994;1:76.
2. Smith RA, Duffy SW, Tabar L. Breast cancer screening: the evolving evidence. Oncology (Willston Park) 2012;26:479–81.
3. Atkin W, Wooldrage K, Parkin DM, et al. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomized controlled trial. Lancet 2017;389(10076):1299–311.
4. Cole P, Morrison AS. Basic issues in population screening for cancer. J Natl Cancer Inst 1980;64:1263–72.
5. Tabar L, Vitak B, Chen THH, et al. Swedish Two-County Trial: impact of mammographic screening on breast cancer mortality during three decades. Radiology 2011;260:658–63.
6. Tabár L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985;1(8433):829–32.
7. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395–409.
8. Duffy SW, Smith RA. A note on the design of cancer screening trials. J Med Screen 2015;22:65–8.
9. Bjurstam NG, Björneld LM, Duffy SW. Updated results of the Gothenburg Trial of Mammographic Screening. Cancer 2016;122(12):1832–5.
10. Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. J Natl Cancer Inst 2002;94(3):167-173.
11. Larsson LG, Nyström L, Wall S, et al. The Swedish randomized mammography screening trials: analysis of their effect on the breast cancer related excess mortality. J Med Screen 1996;3(3):129–32.
12. Sasieni PD, Wald N. Should a reduction in all-cause mortality be the goal when assessing preventive medical therapies? Circulation 2017;135:1985–7.
13. Gøtzsche PC, Jorgensen KJ. Screening for breast cancer with mammography. Cochrane Database Syst Rev 2013;(6):CD001877.
14. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet 2012;380:1778–86.
15. Duffy SW, Parmar D. Overdiagnosis in breast cancer screening: the importance of length of observation period and lead time. Breast Cancer Res 2013;15:R41.
16. Schröder FH, Hugosson J, Roobol MJ, et al. Prostate-cancer mortality at 11 years of follow-up. N Engl J Med 2012;366:981–90.
17. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate cancer mortality: results of the European Randomized Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. Lancet 2014;384:2027–35.
18. Hugosson J, Roobol MJ, Månsson M, et al. A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer. Eur Urol 2019;76(1):43–51.

19. Njor SH, Garne JP, Lynge E. Over-diagnosis estimate from The Independent UK Panel on Breast Cancer Screening is based on unsuitable data. J Med Screen 2013;20:104–5.
20. Olsen AH, Njor SH, Vejborg I, et al. Breast cancer mortality in Copenhagen after introduction of mammography screening: cohort study. BMJ 2005; 330(7485):220.
21. Tabar L, Yen MF, Vitak B, et al. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. Lancet 2003;361:1405–10.
22. Swedish Organised Service Screening Evaluation Group. Reduction in breast cancer mortality from organised service screening with mammography: 1. further confirmation with extended data. Cancer Epidemiol Biomarkers Prev 2006;15(1):45–51.
23. Tabár L, Dean PB, Chen TH, et al. The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. Cancer 2019;125:515–23.
24. Kelly SP, Rosenberg PS, Anderson WF, et al. Trends in the incidence of fatal prostate cancer in the United States by race. Eur Urol 2017;71:195–201.
25. Paap E, Verbeek AL, Botterweck AA, et al. Breast cancer screening halves the risk of breast cancer death: a case-referent study. Breast 2014;23:439–44.
26. Sasieni P, Adams J, Cuzick J. Benefit of cervical screening at different ages: evidence from the UK audit of screening histories. Br J Cancer 2003;89:88–93.
27. Duffy SW, Cuzick J, Tabar L, et al. Correcting for non-compliance bias in case-control studies to evaluate cancer screening programs. Appl Stat 2002;51: 235–43.
28. Massat NJ, Dibden A, Parmar D, et al. Impact of screening on breast cancer mortality: the UK program 20 years on. Cancer Epidemiol Biomarkers Prev 2016; 25(3):455–62.
29. Walter SD. Mammographic screening: case-control studies. Ann Oncol 2003; 14(8):1190–2.
30. Njor SH, Paci E, Rebolj M. As you like it: How the same data can support manifold views of overdiagnosis in breast cancer screening. Int J Cancer 2018;143(6): 1287–94.
31. Puliti D, Duffy SW, Miccinesi G, et al, EUROSCREEN Working Group. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. J Med Screen 2012;19(Suppl 1):42–56.
32. Ryser MD, Gulati R, Eisenberg MC, et al. Identification of the fraction of indolent tumors and associated overdiagnosis in breast cancer screening trials. Am J Epidemiol 2019;188(1):197–205.
33. Brenner H, Qian J, Werner S. Variation of diagnostic performance of fecal immunochemical testing for hemoglobin by sex and age: results from a large screening cohort. Clin Epidemiol 2018;10:381–9.
34. Field JK, Duffy SW, Baldwin DR, et al. UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. Thorax 2016;71(2):161–70.
35. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/774770/Breast_draft_standards_V1.7.pdf. Accessed April 9, 2020.
36. Ciatto S, Houssami N, Bernardi D, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. Lancet Oncol 2013;14(7):583–9.