Summary
In this article we will review how a reader should evaluate a screening study. A clinical problem involving screening mammography is presented. We then outline criteria to determine whether screening is appropriate for a given condition. A search for relevant articles is described followed by an outline of the steps used to appraise a screening study critically. An applicable study is examined in detail for such things as the quality of randomisation and outcomes measured. The results of this study are then applied to a patient considered for screening.

Keywords: screening studies, literature evaluation

A reader’s guide to the evaluation of screening studies

Craig Earle, Paul C Hébert

Screening is the process of trying to detect an illness before it becomes symptomatic. Patients may be screened with history, physical examination, or laboratory testing to exclude a specific disease. It is done with the belief that early diagnosis of the condition will lead to more effective treatments and improve the patient’s long-term health. It is also hoped that screening and early intervention may be more cost-effective than instituting therapy once a disease is well established. We will outline characteristics of diseases appropriate for screening and the principles of assessing screening articles using breast screening as an example (box 1).

Is screening appropriate?

When considering screening, the first step is to decide whether the disease in question, the screening test, and the available interventions are appropriate for screening. Box 2 outlines criteria establishing the feasibility of screening:

Clinical problem

You are a family physician with a busy practice in a large urban centre. One of your regular patients, a healthy 42-year-old woman named Mrs B, comes to see you for a routine checkup. Toward the end of the visit, she asks your opinion on whether she should have yearly mammograms to make sure she does not have breast cancer. She explains that a close friend has just been diagnosed with this condition. After further questioning, you find that she has no risk factors predisposing her to cancer of the breast. Her breast examination is normal. You explain that screening for breast cancer is recommended for women over the age of 50 but you are not sure if mammograms are useful for women of her age. You conclude by asking her to return in one week. At that time, you will provide her with a more definite answer.

Box 1

Does the burden of disease warrant the screening test?
Breast cancer is a significant health problem both in terms of its frequency and severity. An average woman has a one in nine lifetime risk of developing breast cancer, and it is the commonest overall cause of cancer death in women.1 Given the prevalence and severity of this condition, a screening programme is worth considering.

Does the natural history of the disease include a ‘detectable preclinical phase’?
Following the ‘biologic onset’ there must be a period of time, although clinically silent, the disease might be picked up by an appropriate test (figure 1). For breast cancer this would be the time between the initial mutation and the point at which a mass or a metastatic lesion is detected. The preclinical phase should also be the time of most effective intervention. If treatment is equally effective or ineffective when the disease becomes obvious, nothing is gained by screening.2 For example, testicular cancer is highly curable even at very advanced stages. Screening all young men for serum markers to detect this tumour before it became palpable would not be likely to change the overall mortality from the disease. On the other hand, a patient with Alzheimer’s disease, a condition with no useful treatment, would not benefit from early diagnosis either. Lastly, a fulminant condition such as type I diabetes which is symptomatic from the outset would not be amenable to screening because the preclinical phase does not exist.

Is the screening test sensitive and specific?
Mammography has been found to have a sensitivity of approximately 80% and a specificity as high as 99%.3 Unfortunately, the sensitivity of a test is usually gained at the expense of specificity, and vice-versa. Screening tests are usually highly sensitive. The aim is to detect all possible cases, while accepting that there will be false-positive results.

The predictive value of screening parallels the prevalence of disease in the population being tested. As a result, there are many more false positives and negatives when younger women undergo mammography given that the incidence of breast cancer is directly proportional to age (figure 2).4 Predictive values are important because a false-positive screening result would undoubtedly cause your patient unnecessary anxiety. In addition, the burden and risks of follow-up investigations such as surgical biopsy are significant.4 On the other hand, a false-negative mammogram may be inappropriately reassuring and cause the patient to ignore a palpable lump which may develop between mammograms. This may result in a delayed diagnosis and more advanced disease before the administration of therapy.5 In this circumstance it would have been better to receive no screening.
Figure 1. Timeline of breast cancer development. The detectable preclinical phase is the time in which a disease may be diagnosed by a screening test or programme. If the intervention is more effective during this period, screening may improve survival or other relevant outcomes.

Criteria to evaluate a disease for screening

- does the burden of disease warrant the screening test?
- does the natural history of the disease include a 'detectable preclinical phase'?
- is the screening test sensitive and specific?
- is the screening programme feasible (time, cost, discomfort)?
- is there an intervention which can improve outcome?

Box 2

<table>
<thead>
<tr>
<th>Screening test characteristic</th>
<th>Formula</th>
<th>Breast cancer prevalence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1/100</td>
</tr>
<tr>
<td>Prevalence</td>
<td>( \frac{(a + c)}{(a + b + c + d)} )</td>
<td>1.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>( \frac{a}{(c + e)} )</td>
<td>80.0</td>
</tr>
<tr>
<td>Specificity</td>
<td>( \frac{d}{(d + b)} )</td>
<td>99.9</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>( \frac{a}{(a + b)} )</td>
<td>45.0</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>( \frac{d}{(c + d)} )</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Figure 2. Example of how predictive values are dependent on the prevalence of the disease. Sensitivity and specificity are characteristics of a test or evaluation programme which are intrinsic to the test and therefore unaffected by the prevalence. Predictive values, however, are significantly affected by prevalence.
left to the discretion of their family doctor. Endpoints, including both overall mortality and mortality due to breast cancer, were ascertained annually by mailed questionnaires and monitoring cancer registries.

Evaluating screening studies

Once having tracked down a relevant article, we must decide whether the results of the trial are valid. This is the process of critical appraisal. If the conclusions of the paper are not true because of a systematic flaw in the trial, it cannot help in clinical decision making. The criteria used for assessment of screening studies are listed in Table 1.

**What were the objectives of the trial?**

Ideally, the research hypothesis and trial objectives should be explicitly stated in the introduction. One should determine whether it was designed to evaluate the ‘efficacy’ of a particular screening test under the best of circumstances or to look at the ‘effectiveness’ of applying it in the real world (Table 2). The authors describe the CNBSS as an efficacy trial of whether screening can reduce cause-specific mortality from breast cancer. The goal of assessing efficacy was compromised, however, by the non-compliance of some patients and concerns about mammographic quality.

**Were the candidates well described?**

A precise, simple and reproducible description of study patients is essential to determine if the right patients were chosen for screening. Once the internal validity of the study has been confirmed, a precise description of study patients will help us decide if the results also apply to our patient. The patients in the CNBSS were well described. They were recruited from the community by media publicity, physician education, and personal invitation.

Efficacy trials such as the CNBSS try to assemble highly compliant cohorts at high risk of the disease. However, this approach may yield a very select group of patients that is quite different from the reader’s patient population, making the results less generalisable. CNBSS participants were not radically different from the general Canadian population although, on average, they were better educated and of higher socioeconomic status. Both factors may have increased the overall risk of breast cancer in the study group. The main exclusion criteria were a history of breast cancer or mammography in the preceding 12 months.

### Table 1 Criteria for assessing a screening study and their application to the Canadian National Breast Screening Study (CNBSS)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>CNBSS (Part I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What were the objective of the study?</td>
<td>to assess the efficacy of mammography in premenopausal women</td>
</tr>
<tr>
<td>2. Were the candidates well described?</td>
<td>yes</td>
</tr>
<tr>
<td>3. Is the study randomised?</td>
<td>yes</td>
</tr>
<tr>
<td>4. How well was the randomisation done?</td>
<td>well done</td>
</tr>
<tr>
<td>5. Was the screening procedure well described?</td>
<td>yes</td>
</tr>
<tr>
<td>6. Was the follow-up &gt; 80%?</td>
<td>intention-to-treat</td>
</tr>
<tr>
<td>7. How were non-compliers analysed?</td>
<td>overall and cause-specific mortality</td>
</tr>
<tr>
<td>8. What outcomes were measured?</td>
<td>not applicable</td>
</tr>
<tr>
<td>9. If a positive study, were both clinical and statistical significance considered?</td>
<td>yes</td>
</tr>
</tbody>
</table>

### Table 2 Study design features of a screening study

<table>
<thead>
<tr>
<th>Study characteristics</th>
<th>Efficacy</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research question</td>
<td>Does the screening test affect outcome under ideal conditions?</td>
<td>Does the screening test work under real-world conditions?</td>
</tr>
<tr>
<td>Patient selection</td>
<td>Directed selection to identify high risk, highly compliant patients</td>
<td>Assembles a population representative of community screening</td>
</tr>
<tr>
<td>Inclusion criteria</td>
<td>Strict</td>
<td>Lax</td>
</tr>
<tr>
<td>Intervention</td>
<td>Tightly controlled: optimal equipment under optimal circumstances</td>
<td>Loose: test performed as it would be in the community</td>
</tr>
<tr>
<td>Endpoints</td>
<td>Disease-specific, or ‘surrogate’ endpoints, eg, number of early-stage tumours detected</td>
<td>Patient-specific, eg, all-cause mortality rate or quality of life</td>
</tr>
<tr>
<td>Analysis</td>
<td>By treatment received, ie, non-compliers and crossovers are excluded from the analysis</td>
<td>By intention-to-treat, ie, non-compliers are included to reflect the feasibility of the test</td>
</tr>
</tbody>
</table>
Is the study randomised?
The CNBSS randomly allocated patients to one of the study groups using standard techniques. In large studies, randomisation usually results in an equal distribution of patients with both known and unknown risk factors. Randomisation also avoids 'self-selection bias'; it has been observed that patients who volunteer for a screening arm tend to be more health-conscious and, on average, remain healthier than those who do not.10

Randomisation is important in screening studies since the results of non-randomised trials are often influenced by 'lead-time' and 'length-time' bias. Lead time is the interval between the diagnosis of a disease at screening and the time at which it would become clinically obvious due to symptoms. By detecting disease one year earlier with screening, for example, our patient would seem to live one year longer from the time of diagnosis, though she may die at exactly the same time if intervention is not effective. She will not have lived longer, but will only have lived longer with the diagnosis (figure 3). Lead-time bias is eliminated by comparing mortality rates rather than length of survival after diagnosis.11

With 'length-time bias,' tumours that are slow growing will have a longer period in which they are clinically silent yet detectable by the screening test. Therefore, they have a greater chance of being picked up by the screening test than faster-growing tumours. This is especially evident at the initial testing as patients with more aggressive tumours will often have already become apparent and removed from the population being screened. As a result, the more indolent tumours left over get picked by the screening (figure 4). Because they are more slowly progressive than those picked up clinically, patients identified by screening will appear to live longer even though there has been no change in the natural course of their disease.11

How well was the randomisation done?
The best way to assess the quality of randomisation is to examine the size and baseline characteristics of the two groups. The goal is to assemble two identical cohorts, so that the only difference between them that could affect the outcome is the screening programme. The two groups in the CNBSS were similar except that the majority of node-positive tumours detected at initial screening ended up being assigned to the mammography group. Of 22 women who ultimately proved to have breast cancer with four or more positive nodes, 19 were assigned to the mammography screening arm of the study. The investigators have been unable to explain how this imbalance occurred except by chance. The result is that the two cohorts did not start off with an equal risk of breast cancer-related death. This might have diminished any survival advantage screening could have provided to the group. When there is an imbalance in known risk factors an attempt should be made to adjust for the maldistribution. This is usually done by multivariate analysis. Using this approach, the investigators attempt to determine if there was an effect of screening while adjusting for the imbalance in baseline risk factors. Excluding women with palpable abnormalities from the start may have avoided the problem altogether.

Was the screening procedure well described?
The authors should provide enough detail about the screening methods that the procedures used in the study may be replicated in other settings. Screening techniques were extensively scrutinised and well described in the CNBSS. Results of the quality assurance done as part of the study were published in several other articles referenced in the paper. Although there has been some criticism of mammographic quality in the trial, the CNBSS achieved rates of cancer detection similar to other comparable studies.12

### Box 3

<table>
<thead>
<tr>
<th>Some definitions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias: a systematic error introduced in the design or conduct of a study</td>
<td></td>
</tr>
<tr>
<td>Length-time bias: slower disease processes detected more often than more aggressive disease making a screening test appear to prolong survival</td>
<td></td>
</tr>
<tr>
<td>Lead-time bias: the disease is detected earlier but survival remains unchanged. Thus, the screening appears to prolong survival</td>
<td></td>
</tr>
<tr>
<td>Efficacy: the evaluation of an intervention under optimal conditions</td>
<td></td>
</tr>
<tr>
<td>Effectiveness: the evaluation of an intervention under 'real world' conditions</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3 How lead time affects survival time after screening. Shaded areas indicate length of survival. Patient A illustrates the natural history of a disease. Patient B had a screening test which led to an early diagnosis. Without effective treatment she appears to live longer but dies at the same time as she would have if she had not been screened. This is lead-time bias. Patient C has had an early diagnosis with effective intervention resulting in improved survival.11
Was the follow up greater than 80%?
It is important that the authors report on as many of the patients who were randomised as possible, since those who drop out or are lost to follow-up often do not do so by chance. Their characteristics may have influenced the results. In the CNBSS follow-up in both groups was greater than 90%. If follow-up is incomplete a sensitivity analysis should be done. This would entail re-analysing the data assuming all of the lost patients had one outcome, then reworking it under the assumption that they had the opposite outcome. This form of sensitivity analysis is called the ‘best case/worst case’ scenario. If the sensitivity analysis does not significantly affect the results of the trial, the conclusions are probably valid.

How were non-compliers analysed?
In many respects non-compliance (the failure of patients to adhere to the study protocol) is comparable to losses to follow-up. Both are consequences of undertaking research in humans and may lead to inaccurate or inappropriate conclusions. In conducting a screening study, investigators should make every effort to minimise protocol violations (non-compliance and contamination). However, non-compliance invariably occurs. If compliance rates are less than 90% the authors should provide the appropriate analysis and discuss how non-compliance may have influenced the study results. If compliance rates are less than 80%, study conclusions should be considered suspect.

In efficacy trials non-compliers may be removed from the analysis and patients analysed by treatment received. In effectiveness trials, however, all patients should be considered in the analysis according to the randomisation assignment. Despite claiming to be an efficacy trial, the CNBSS analysed patients who did not adhere to the study protocol as part of the group to which they were assigned. Non-compliance, or ‘contamination,’ was a significant problem as 26% of control women underwent mammography outside the trial. Thus, the control arm had some form of screening during the study period. In addition, only 85% of women assigned to receive mammography did so. Both arms in the study were contaminated to some extent. Although a compliance rate of 85% in women assigned to receive mammography is very good when compared to similar studies, 10–15% of women who did not continue mammography dilute differences that could exist between the two strategies. This is a major weakness of the trial.

In a study of ‘real life’ effectiveness, on the other hand, non-compliers should be analysed with the group to which they were assigned. This is called ‘intention-to-treat’ analysis. It gives an indication of how well the manoeuvre being studied was tolerated by patients under real world conditions. If non-compliers are not analysed in an effectiveness trial, the authors should acknowledge how this might affect the results.
What outcomes were measured?

In an efficacy trial the important outcomes are related to the biology of the disease in question, such as the percentage of T1 tumours detected. This contrasts with effectiveness studies in which the outcomes are important to the patient such as overall mortality or quality of life. To estimate the importance of the outcome being measured one can ask whether a significant difference would be likely to change clinicians’ practice or a patient’s decision to participate in screening.

In addition to disease-specific outcomes like the stage and nodal status of the tumours picked up, the CNBSS also examined patient-specific outcomes including overall and cause-specific mortality rates. Ascertainment of cause of death appeared thorough and free of bias. Data was also collected on the number of benign (unnecessary) biopsies of masses generated by the screening program.

After seven years of follow-up, the screened group showed no decrease in cause-specific mortality. In fact, the trend was toward increased mortality. Thirty-eight women in the screened arm of the study died of breast cancer compared to 28 in the control group. This yields a breast cancer death rate of 14.7 per 10 000 person-years in the screened group versus 10.4 in those receiving usual care. Overall mortality rates were also comparable between the two groups.

In a negative trial, was the power assessed?

A negative study result suggests that there is no difference between the two groups examined. There are other possible explanations, however. If the trial was too small it may not have had the power to detect a difference. A sample size should be calculated a priori using estimates of the expected frequency of disease in the population being studied and the difference in outcome anticipated. Studying a patient group at a lower risk for the disease than predicted is therefore another reason for a negative trial. Similarly, if the intervention did not work as well as expected, or if there was an unforeseen biological effect that reduced the number of events in the study group, the planned power would not be achieved.

In the CNBSS a large trial was needed because few breast cancer deaths were expected to occur within the five years of study as this age group has a relatively low breast cancer incidence and good prognosis. The sample size was calculated to have a 90% power to detect a 40% reduction in breast cancer deaths at five years, similar to that seen for women older than 50 in the HIP trial. At five years, however, the number of deaths from breast cancer was in fact less than predicted. It was too low to achieve the planned power. The trial was therefore extended for two more years.

As a result, critics have felt that the CNBSS’s conclusions have been based on too short a period of follow-up. No trial to date has been able to show a benefit from screening mammography for any age group before seven years of follow-up, and even these reductions were so small that they were not statistically significant. The investigators plan to continue following patients and report again after 10 years. For this reason many consider the results of the CNBSS as preliminary.14

Do the results of the study apply to your patient?

For a study to be useful to us in clinical practice the results must be both valid and applicable to our patients. The CNBSS is a well-conducted study which concurs with the current literature. No screening trial has been able to show a conclusive survival advantage from mammography in the subgroup of women under 50 years old. This may be because mammography is less sensitive in the pre-menopausal breast,2 or perhaps screening should be more frequent to detect younger women’s faster-growing tumours.3 We judge the results valid, recognising that further follow-up is needed to allow the data to mature.

For the results to be applicable to our patient, the study population should be biologically similar enough to her that the results can be extrapolated. Either the average baseline characteristics of the group should be comparable to our patient, or she should belong to a well-described subgroup in the trial in which the study conclusions are still valid. Our patient has many of the same baseline characteristics as the study group.

Armed with your evidence you see Mrs B in follow-up the next week and explain the results of your search. You tell her that mammography does not add anything to the clinical breast exam in screening younger women for cancer, probably because mammograms are less sensitive in the denser breasts of pre-menopausal women. Mrs B accepts the conclusions. She will continue seeing you for her periodic examinations. She agrees to continue to practise breast self-examination on a monthly basis, and will postpone her first mammogram.

Key points

- several disease characteristics such as the burden of disease, an appropriate detectable preclinical phase, and an effective treatment should be present prior to screening patients or populations
- a screening test should also be accurate, safe and cost effective
- a screening test or programme is best evaluated using a randomised clinical trial design.

Box 4