Breast Cancer Screening: A 35-Year Perspective

Suzanne W. Fletcher

* Correspondence to Dr. Suzanne W. Fletcher, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 133 Brookline Avenue, 6th Floor, Boston, MA 02215 (e-mail: suzanne_fletcher@hms.harvard.edu).

Accepted for publication January 31, 2011.

Screening for breast cancer has been evaluated by 9 randomized trials over 5 decades and recommended by major guideline groups for more than 3 decades. Successes and lessons for cancer screening from this history include development of scientific methods to evaluate screening, by the Canadian Task Force on the Periodic Health Examination and the U.S. Preventive Services Task Force; the importance of randomized trials in the past, and the increasing need to develop new methods to evaluate cancer screening in the future; the challenge of assessing new technologies that are replacing originally evaluated screening tests; the need to measure false-positive screening test results and the difficulty in reducing their frequency; the unexpected emergence of over-diagnosis due to cancer screening; the difficulty in stratifying individuals according to breast cancer risk; women’s fear of breast cancer and the public outrage over changing guidelines for breast cancer screening; the need for population scientists to better communicate with the public if evidence-based recommendations are to be heeded by clinicians, patients, and insurers; new developments in the primary prevention of cancers; and the interaction between improved treatment and screening, which, over time, and together with primary prevention, may decrease the need for cancer screening.

breast neoplasms; early detection of cancer

Abbreviation: USPSTF, U.S. Preventive Services Task Force.

INTRODUCTION

My introduction to breast cancer began in 1962, during the first week of medical school. Special grand rounds were held for new students. A woman with breast cancer (one who I now know had a tumor that was estrogen-receptor positive) had been diagnosed many years before; she had been treated sequentially with mastectomy, ovariectomy, adrenalectomy, and pituitectomy. When first diagnosed, she wondered whether she would live to see her son’s bar mitzvah; at the rounds, she looked forward to his impending wedding. The presentation taught students stages of breast cancer and treatments. There was not a word about screening.

A year later, in 1963, the Health Insurance Plan of New York, the first randomized trial of cancer screening, began to evaluate mammography and clinical breast examination. The trial, first results of which were reported in 1971 (1), set the standard for conducting the 8 succeeding randomized trials of breast cancer screening over the next 45 years. Altogether, randomized trials of breast cancer screening have involved more than 650,000 women. Screening for no other cancer has received such intense study; even so, no other cancer screening has produced such heated controversy. Multiple reviews and conclusions have been published, and interest has been strong not only in the medical literature but also among the lay public.

This paper is not another review. Rather, I focus on the larger lessons that research on breast cancer screening has uncovered. Over the years, the scientific controversy has spurred discovery and new thinking as investigators explored multiple ways of analyzing the accumulating data. This process has led to many of the lessons I discuss. My professional career serendipitously has spanned the period from pre-breast cancer screening to the present time—giving me a front-row seat to observe and a chance to participate (Table 1). Most of the lessons that emerged apply to not only breast cancer but other cancers as well. Along the way, I suggest some next steps that are needed. Finally, I consider possible long-term future directions for cancer screening.
THE CANADIAN TASK FORCE ON THE PERIODIC HEALTH EXAMINATION

The Canadian Task Force on the Periodic Health Examination (now the Canadian Task Force on Preventive Health Care) was established in 1976 (2), 5 years after the first published results of the Health Insurance Plan of New York trial. Formed at the request of the Conference of Deputy Ministers of Health of Canada, the Task Force evaluated the periodic health examination and made recommendations for office-based preventive practices. Its report was to, not from, government. The 10 Task Force members all came from academic settings, primarily clinical departments. Several members were experts in clinical epidemiology, epidemiology, and/or biostatistics.

From the perspective of 35 years, the importance of the Task Force’s work lay not so much in its specific recommendations as the approach it hammered out to evaluate evidence about office-based preventive services. In doing so, it examined previous recommendations and considerations, particularly those of Frame and Carlson (3–6) and Sackett and Holland (7). Most importantly, the Task Force based its recommendations on evidence in the medical literature.

Fifty-seven of the 78 health conditions the Task Force considered in depth involved screening (10 for cancer). The Task Force defined screening as an activity in asymptomatic persons “making use of procedures by which unselected general populations are classified into 2 groups: one with a high probability of being affected by killing or disabling conditions, unhealthy states or unhealthy behaviors, and the other with a low probability” (8, p. 13). Screening procedures included history-taking, physical examination, laboratory testing, and procedures such as radiography.

The Task Force recommended replacing the untargeted complete “annual examination” with a highly targeted examination aimed at preventing specific conditions, packaged according to the age and sex of the patient. This approach has now become the standard for preventive services, not only in Canada and the United States but in many other countries as well.

In deciding what conditions to target, the Task Force formalized several important methodological contributions to screening. First, when considering the evidence for screening for a given condition, Task Force members searched the health literature for answers to 3 questions, an approach that still constitutes the bedrock for evaluating screening (Table 2): 1) How great is the burden of suffering caused by the condition being sought? 2) How good is the test used to detect the condition during screening? and 3) How effective is the resulting treatment or preventive intervention?

The second major contribution was the Task Force’s approach to evidence. It recognized that evidence about effectiveness of prevention varied in terms of scientific rigor. Evidence from a well-designed and conducted randomized trial was given more weight in final recommendations than evidence from a cohort study, which in turn was stronger than the opinion of a clinical expert. The Task Force formally incorporated the strength of evidence into its recommendations by developing a grading system, from I (evidence from at least one well-conducted randomized trial) to III (expert testimony).

The third major contribution was to assign an overall grade to the recommendation for screening each condition considered. Grades ranged from A (good evidence that the condition be specifically considered in a periodic health examination) to C (not enough good evidence to recommend whether to include or exclude the condition in a periodic health examination) to E (good evidence to recommend that the condition be excluded from the periodic health examination). The graded recommendation was to incorporate all the information uncovered about the 3 questions of effectiveness, burden, and test quality; however, in practice, recommendations were dominated by strength of the evidence about effectiveness of treatment or prevention after screening. Table 3 shows the distribution of recommendation grades. In several cases, the Task Force indicated that screening should receive high priority for future research.

THE U.S. PREVENTIVE SERVICES TASK FORCE

The U.S. Preventive Services Task Force (USPSTF), an independent panel of experts in primary care, prevention, and research methods, was begun in 1984. Supported since 1998 by the Agency for Healthcare Quality and Research, it is charged by law to review the scientific evidence and make recommendations for clinical preventive services (9). Interaction with the Canadian Task Force has included...
meetings considering methodological issues and publication of a book on prevention (10).

The US Task Force adopted the basic approach of the Canadian Task Force, publishing its first Guide to Clinical Preventive Services in 1989 (11), with regular updates, now available on the Internet. Over the past 27 years, the Task Force has further developed the scientific approach for evaluating preventive and screening interventions and has standardized methods for reviews and recommendations. It works with an Agency for Healthcare Quality and Research–funded Evidence-based Practice Center, which conducts most reviews of evidence. The Task Force has routinely updated its approach (12), incorporating systematic reviews, meta-analyses, and modeling into its review methods; considering how to estimate both the certainty and the magnitude of net benefit of preventive maneuvers (13); and considering how to standardize the process of review when there is insufficient evidence (14). Membership of the Task Force was enlarged to include nurses, and members of interested outside groups are invited observers of the meetings. Rigorous peer review of each update was instituted, and draft recommendations are posted electronically for a period of public comment. In sum, the USPSTF took the beginning efforts of the Canadian Task Force and has continuously incorporated newly developed scientific methods that must undergird evaluation of screening.

### RANDOMIZED TRIALS

Both the Canadian and US task forces emphasized the importance of randomized controlled trials in their assessments of and recommendations for cancer screening. Breast cancer screening has been assessed in multiple randomized trials, and, even though much controversy has ensued, there is widespread agreement that the results found breast cancer screening to be effective for women of certain ages. The most recent meta-analysis found that breast cancer mortality reduction among women invited to screening was 15% for women aged 39–49 years, 14% for women aged 50–59 years, and 32% for women aged 60–69 years, with corresponding numbers needed to invite to screening to prevent 1 breast cancer death of 1,904, 1,339, and 377, respectively (15). With time, randomized trials of screening have been

#### Table 2. Key Questions Asked by the Canadian Task Force on the Periodic Health Examination When Considering Screening

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Good evidence to consider the condition in a periodic health examination)</td>
<td>1. How great is the current burden of suffering caused by the condition to be sought by screening, in terms of severity and frequency, for both the individual and society?</td>
</tr>
<tr>
<td>B (Fair evidence to consider the condition in a periodic health examination)</td>
<td>2. How good is the test used to detect the condition during screening, in terms of</td>
</tr>
<tr>
<td>C (Poor evidence regarding inclusion of the condition in a periodic health examination, and recommendations may be made on other grounds)</td>
<td>3. How effective is the resulting treatment or preventive intervention?</td>
</tr>
<tr>
<td>D (Fair evidence to recommend exclusion of the condition from the periodic health examination)</td>
<td>“(D)oes the available treatment, preventive or therapeutic, instituted as a result of carrying out the periodic health examination, do more good than harm to those patients to whom it is offered?” (8, p. 16).</td>
</tr>
<tr>
<td>E (Good evidence to recommend exclusion of the condition from the periodic health examination)</td>
<td>Effectiveness of an intervention depends on efficacy (whether it does more good than harm for patients who follow instructions) and compliance (the extent to which patients follow instructions).</td>
</tr>
<tr>
<td></td>
<td>“Effectiveness of treatment begun during asymptomatic phases of a health condition must be superior to that of treatment begun only when symptoms occur” (8, p. 16).</td>
</tr>
</tbody>
</table>

#### Table 3. Recommendations of the Original Canadian Task Force on the Periodic Health Examination (2)

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>All Conditions</th>
<th>Screening for Cancers, No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Good evidence to consider the condition in a periodic health examination)</td>
<td>8 10</td>
<td>1 (breast)</td>
</tr>
<tr>
<td>B (Fair evidence to consider the condition in a periodic health examination)</td>
<td>17 22</td>
<td>2 (cervical, colorectal)</td>
</tr>
<tr>
<td>C (Poor evidence regarding inclusion of the condition in a periodic health examination, and recommendations may be made on other grounds)</td>
<td>33 42</td>
<td>4 (stomach, oral, prostate, Hodgkin’s)</td>
</tr>
<tr>
<td>D (Fair evidence to recommend exclusion of the condition from the periodic health examination)</td>
<td>16 21</td>
<td>3 (lung, skin, bladder)</td>
</tr>
<tr>
<td>E (Good evidence to recommend exclusion of the condition from the periodic health examination)</td>
<td>4 5</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>78 100</td>
<td>10</td>
</tr>
</tbody>
</table>

*Epidemiol Rev* 2011;33:165–175
conducted for the other major cancers, including lung, colorectal, prostate, and ovarian. Along with breast cancer, these cancers account for half of all cancer deaths.

What about cancers with much lower incidence and mortality rates, such as melanoma and testicular cancer? Is it likely, or even feasible, to mount a large-enough randomized trial to determine the effectiveness of screening for such cancers? Because of the low incidence of these cancers, the numbers of people required would dwarf the numbers involved in trials of screening for the major cancers. For example, it has been estimated that a randomized controlled trial of screening for melanoma would require approximately 800,000 persons aged 50–74 years (more than the number of women involved in all the randomized trials of breast cancer screening combined) to determine whether screening decreases melanoma mortality by a third (16).

Randomized trials of screening take a long time before mortality results can be obtained, usually more than a decade. The trials are costly because of the large numbers of persons who must be followed over many years. Contamination threatens the validity of trial results, especially in the United States, where the population can obtain unvalidated screening tests easily. Contamination was a major problem in the National Cancer Institute–funded Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial; in the study evaluating the effectiveness of prostate cancer screening, up to 52% of the control group reported obtaining prostate-specific antigen testing outside the trial (17). Finally, randomized trials of cancer screening take so long that the development and introduction of new screening technologies threaten to make the results of a trial irrelevant to clinical practice by the time they are reported. The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial was begun before colonoscopy had become the dominant form of colorectal cancer screening in the United States and before a preliminary study suggested spiral computed tomography could be a better screening test than chest radiograph. In both cases, randomized studies had to be amended or begun to assess these new tests.

Once randomized trials have found that screening with a given test decreases mortality, must a new randomized trial be conducted for every newly developed test? For breast cancer, many new technologies, including digital mammograms, computer-assisted detection, and breast magnetic resonance imaging, have been developed and recommended for screening, after randomized trials demonstrated effectiveness of screening with film-screen mammography. No randomized study has been undertaken to determine whether screening with any of these new tests decreases breast cancer mortality over that of film-screen mammography. A 2001 report by the Institute of Medicine recommended that approval of new screening technologies should depend on evidence of improved clinical outcomes, but the report side-stepped the question of how this could be accomplished (18). The report also pointed out that, too often, new technologies were advocated for breast cancer screening after having been developed for diagnostic, not screening, purposes.

Cancer awareness has increased in the population over the last several decades. Cancers detectable by sight or touch (e.g., melanoma, testicular cancer, and breast cancer) may be noticed by affected individuals more often at curable stages today than was true several decades ago. Detection of these cancers outside of screening programs will make it even more difficult to demonstrate screening effectiveness in randomized trials.

Taken together, these problems raise the concern that it will be difficult for traditional randomized trials to continue to be the standard for evaluating cancer screening effectiveness. How to undertake faster, cheaper, more effective randomized trials for cancer screening deserves serious consideration. Some problems, such as contamination, could be handled with studies conducted in countries where contamination is not a problem because national health programs exist. Many countries contributing patients would decrease the cost for any one country when studying screening for less common cancers. It is also important to develop a consensus on the priority regarding when to conduct randomized trials. Is it more important to conduct a randomized study for a new technology for breast cancer screening (could comparative effectiveness studies suffice?) or for melanoma screening, for which no randomized trial has determined mortality effects?

What other methods of evaluation should be considered? The USPSTF has proposed that, when evidence from randomized trials is not available, nonrandomized studies such as cohort, cross-sectional, case-control, or quasi-experimental strategies may be necessary, despite the known problems of bias, especially confounding and lead time (14). In addition, the Task Force proposed 4 domains of information that should be considered: potential preventable burden, potential harms, costs, and current practice.

Multiple time series should be used more often in screening evaluations. The Papanicolaou smear was introduced for cervical cancer screening before randomized trials were used in medical research. Using the multiple time-series method, Canadian and Scandinavian researchers demonstrated that cervical cancer mortality decreased after population cervical cancer screening programs were introduced in different regions over different years (19–21). Multiple time-series data may be particularly useful in evaluating screening introduced across health systems or countries at different times. If the effect always follows introduction of an intervention, it is less likely that extraneous factors are involved. However, the multiple time-series method does not track interventions at the individual level. Also, confounding cannot be completely ruled out; for example, cancer screening might be accompanied by a new treatment that could improve mortality regardless of screening.

The Swedish 2-county study of breast cancer screening (22) and an Australian study of community-wide screening for melanoma (23) used cluster randomized trials. Another method, the stepped wedge cluster randomized design, combines cluster randomization with a one-way crossover design in which different clusters cross over from control to intervention at randomly determined different time points (24). With this method, ultimately the intervention is introduced to all clusters, but each cluster acts as a control for a given period of time. Carrying out the study in a setting with electronic medical records would facilitate collecting
relevant individual data. Use of a stepped wedge cluster randomized design has not been reported in evaluation of screening interventions, and statistical problems with changes in the composition of intervention and control groups, as well as the long time duration in randomized trials, would present analytic challenges. If these problems could be overcome, a stepped wedge cluster randomized design might be a way to evaluate screening when traditional randomized trials are not planned. For example, in Germany, a nationwide skin cancer screening program has begun (25). A stepped wedge cluster randomized design might have allowed rigorous evaluation of the intervention as it was implemented across the country.

After screening for a given cancer has been shown to be effective, rigorous evaluation of newly developed tests is necessary. At a minimum, the characteristics (sensitivity, specificity, safety, simplicity, cost, acceptability, and labeling) of any new test should be compared with those of the original test in the setting of community practice screening before the new technology is disseminated. For breast cancer, the sensitivity and specificity of digital mammography and film-screen mammography were compared in a large screening study (26), and a cost-effectiveness analysis was performed (27). No such comparison has been made for other new tests advocated for breast cancer screening of the general population. (A large, prospective Dutch study has compared magnetic resonance imaging with mammography and clinical breast examination among women at high risk of breast cancer (28).) When comparisons are not done, or they are performed in only small, selected groups of patients, the further out from original randomized trials, the less certain it is that screening with a new generation of technology is more, or even as, accurate as the test originally used. Rigorous comparison of new and old screening test characteristics alone does not determine the degree of overdiagnosis (refer to the Overdiagnosis discussion below).

FALSE-POSITIVE SCREENING TESTS

From the very beginning, the Canadian and US task forces considered the trade-off between benefit and harm in the decision to recommend screening. However, concern about harms was primarily for the treatment following screening. When the screening test itself was considered, specificity and safety were discussed, but specificity was considered primarily as test efficiency and cost, not harm. Test safety was thought of in terms of physical harm during the testing procedure, such as radiation effects of mammography or colon perforations during sigmoidoscopy. There was no discussion of possible harmful effects of false-positive results. In addition, the possibility of cumulative harm from screening tests that were to be repeated again and again was not explicitly considered. Studies of breast cancer screening have led the way in clarifying these concerns.

In the United States, 6.2%–18.8% (depending on age and time since a previous mammogram) of screening mammogram readings result in a recall for subsequent action other than future routine screening (29). Using this definition of an abnormal screening mammogram, most abnormal mammograms are not due to breast cancer; from age 40 to age 70 years, 91%–98.6% of abnormal mammograms are false positive (i.e., no breast cancer is diagnosed in the year after the abnormal mammogram). Follow-up testing after false-positive tests adds about 33% to the cost of breast cancer screening (30).

In the 1990s, researchers started studying the psychological and behavioral effects of experiencing a false-positive mammogram (31, 32). In the United States, women with false-positive results have higher levels of distress and anxiety and think more about breast cancer, but they also increase their subsequent use of screening mammography. In one study, physicians recorded patient anxiety in the medical records of 10% of patients after a false-positive mammogram; furthermore, health care visits, both breast related and nonbreast related, increased over the subsequent year (33). Nevertheless, a survey found that women viewed false positives as acceptable consequences of screening (34). Anxiety after false-positive tests has not been studied for most other cancers.

The chance that a woman will experience a false-positive mammogram over time is substantially higher than the 6%–19% of abnormal mammograms. In a US study, it was estimated that about half of women receiving annual mammograms would experience a false-positive mammogram over a 10-year period (30). A study in Europe found a cumulative incidence of 21% after 10 mammograms, about half that found in the US study (35). The cumulative incidence of experiencing a false-positive screening test has also been reported for lung cancer; after only 2 years, the cumulative false-positive rates were 15% for chest radiographs and 33% for low-dose computed tomography (36).

False-positive test results may increase in the future if newer technology increases sensitivity but worsens specificity (as clear with the example of lung cancer screening). Regarding breast cancer, studies have found that although sensitivity is higher, specificity is lower when breast magnetic resonance imaging is compared with mammography, even for patients at high risk of developing breast cancer because of genetic mutations (37). Use of more sensitive and less specific technology may be appropriate for patients at very high risk of developing cancer, such as those with BRCA mutations or untested women with first-degree relatives with BRCA mutations, but use of these tests can spread to populations at lower risk. The American Cancer Society has broadened its recommendation for screening breast magnetic resonance imaging to include women with a lifetime risk of breast cancer of 20% or greater based on risk models largely dependent on family history (38).

Decreasing the frequency of false-positive mammograms will be difficult in the United States. In Europe, cancer detection rates are similar to those in the United States, but frequency of false-positive mammograms is lower, probably because of differences in the medio-co-legal environment, guidelines for appropriate false-positive rates, and requirements for mammographers (39). Decreasing anxiety after a false-positive mammogram also appears difficult. In a randomized trial, efforts to educate women did not decrease anxiety among those with an abnormal mammogram.
had been abnormal. Although anxiety was highest for women who needed a breast biopsy, it was second-highest among women asked to return in 6 months, those for whom the mammographer was least concerned. Only those women for whom onsite reading and immediate follow-up were available had lower anxiety scores; many were not aware that their mammogram had been abnormal.

OVERDIAGNOSIS

I recall no early discussion of the concept that some cancers could remain dormant and never cause harm to patients, even though they were indistinguishable pathologically from potentially lethal cancers. The idea goes directly against the fundamental thesis of cancer screening—the earlier a cancer is found, the better the chance of cure. However, evidence began appearing that challenged the thesis. For some cancers, incidence of more advanced cancers did not fall commensurately as incidence of early-stage cancers increased. Furthermore, reports of long-term follow-up of some randomized trials of screening demonstrated an excess of cancers in the screening group that did not disappear over ensuing years when the number of cancers in the control group should have caught up. It was as if screening was causing cancer.

Overdiagnosis occurs when cancers are found on screening that will not cause death or symptoms if left alone; such cancers either regress or do not progress. Overdiagnosis in breast cancer screening was first suspected with the rapidly increasing detection of ductal carcinoma in situ, concurrent with the introduction of mammography. Although not invasive, ductal carcinoma in situ is associated with an increased risk of subsequent breast cancer. It was expected that the discovery and treatment of ductal carcinoma in situ, over time, lead to a decrease in incidence of invasive breast cancer, but invasive cancer continued to increase until 2003, when a small decrease coincided with large numbers of women stopping hormone replacement therapy after the Women’s Health Initiative Study reported harmful effects (41).

To what degree does overdiagnosis occur in breast cancer screening? A 2007 review of 8 studies found 3 reports based on randomized trials of screening and 5 on trends of breast cancer incidence before and after population-based screening programs were introduced (42). Estimates of overdiagnosis varied widely, from −13% to 84% of breast cancers detected. Major biases included different cancer risks in the screened and control populations, low compliance, contamination, offering screening to the control group before or during follow-up, and inappropriate adjustment for lead time. The authors concluded that the best method of measuring overdiagnosis is the cumulative-incidence approach, using data from randomized trials in which there are long-term results after screening has ended in the screened group and no screening in the control group. One study meeting these standards (but not taking into account noncompliance and contamination) found that the number of breast cancers diagnosed in the screened group was 10% higher than in the control group 15 years after screening ended (43).

Another issue in measuring overdiagnosis is the appropriate denominator that should be used. When overdiagnosis is calculated by subtracting the total number of breast cancers diagnosed in the control group from that in the screened group and dividing the result by the total number of breast cancers diagnosed in the control group, the result is a percentage of all breast cancers diagnosed, whether or not detected on screening (42), useful information for a screening program. However, if a woman wants to know how likely a cancer found on mammography represents overdiagnosis, Welch et al. (44) suggest that the correct denominator should be the number of breast cancers detected by screening in the screened group. Applying this approach to the above study, they calculated that 24%, not 10%, of cancers detected by mammography screening were the result of overdiagnosis.

Yet another reason for different estimates of overdiagnosis is that some include ductal carcinoma in situ and some do not. A recent National Institutes of Health State-of-the-Science Conference’s first recommendation was to develop and validate risk stratification models to identify patients with ductal carcinoma in situ who are at such low risk of subsequent adverse clinical outcomes (i.e., overdiagnosis) that they be followed with surveillance only (45). Overdiagnosis has been documented in screening for several other cancers, particularly in prostate cancer screening, with estimates of 24%–50% of cancers diagnosed after prostate-specific antigen screening due to overdiagnosis (46, 47).

Determining how to calculate overdiagnosis and how to avoid or account for biases may become more urgent as new randomized trials of breast cancer screening become less likely, while new technologies are likely to increase both sensitivity and overdiagnosis. Because multiple methods to calculate overdiagnosis and results are reported, a working group should be convened to consider the most valid and feasible methods to determine the degree of overdiagnosis in cancer screening.

RISK STRATIFICATION

There are increasing calls to develop prediction models that will stratify individual women by risk, to concentrate on those who are most likely to develop breast cancer, and to minimize harm to those least likely to benefit from screening (48, 49). The intuitive appeal to such an approach is overwhelming. Already, some stratification occurs in breast cancer screening, with different recommendations for groups of women according to age and, more recently, genetic mutation status. Can risk stratification help at the individual level of most women?

National Cancer Institute scientists have developed a popular breast cancer risk tool (50) that predicts a woman’s likelihood of having a breast cancer diagnosis in the next 5 years and up to 90 years of age, after she enters personal information about 8 risk factors. The prediction model works well at the population level—predicting how many breast cancers will occur in groups of women with similar risk factors (calibration)—but performs much less well at the level of individual women (discrimination), Figure 1
show that the curves for individual women at “high” and “low” risk of developing breast cancer over the next 5 years according to the prediction model overlap almost totally \((51)\); B. Rockhill Levine, Wake Forest University School of Medicine, personal communication, 2011). Efforts to identify additional risk factors to improve the risk prediction tool, including genomic information \((52)\), so far have shown limited success.

Why are useful risk prediction tools for individual women so difficult to develop in breast cancer screening? In 1985, Geoffrey Rose pointed out, “a large number of people at a small risk may give rise to more cases of disease than the small number who are at high risk. This situation . . . limits the utility of the ‘high-risk’ approach to prevention” \((53, p. 37)\). Most risk factors for breast cancer are modest. Wald et al. \((54)\) calculated that for a risk factor (or combination) to detect about half the individuals with a disease, with a 5% false-positive rate, a relative risk of about 200 is required.

Risk prediction tools for other cancers have not yet been as rigorously developed and tested as those for breast cancer. However, except for lung cancer or highly selected populations that do not account for most cancer occurrences, we already know that risk factors for common cancers are not large. Rose’s reasoning \((53)\) is likely to apply to most, not just breast, cancer.

Most evaluations of risk prediction models report calibration and discrimination statistics and/or sensitivity, specificity, and receiver operating characteristic curves. Such reports, important for researchers, are difficult for clinical understanding of the models. All reports should visually demonstrate how well a model separates curves of people who do and do not develop the cancer of interest (as in Figure 1), and should show absolute numbers as well as percentages.

A different approach to decreasing adverse effects in breast cancer screening while preserving most of the benefit is the one adopted by the US Task Force recommendation \((55)\). Effects of mammography screening were modeled under different screening schedules; compared with that for annual screening, most of the mortality benefit of breast cancer screening was preserved with biennial screens, while frequency of false-positive mammograms was cut in half. Overdiagnosis was also reduced.

**COMMUNICATION**

The original report of the Canadian Task Force was published in the medical journal of the Canadian Medical Association \((2)\), and an editorial was published in *Annals of Internal Medicine* \((56)\). The first USPSTF report was published as a book \((11)\). The US Department of Health and Human Services published *Clinician’s Handbook of Preventive Services* \((57)\), along with a “Put Prevention Into Practice Education and Action Kit.” There was no organized effort to communicate screening recommendations directly to the public.

Meanwhile, the American Cancer Society and other groups understood early that it was important to speak directly to the public about the need to find cancer early. Society publications and TV ads stressed that breast cancer strikes a large percentage of women over their lifetime, a percentage that kept growing over the decades. Then, too, feminism was on the rise in society, with a focus on the need for women to take back ownership of their bodies, including deciding about breast care. Many lay advocacy groups were formed and successfully pushed for increased breast cancer research funding. Discussing breast cancer in public became more acceptable, and the lay media increased its coverage of breast cancer and breast cancer screening. Breast cancer screening was the subject of a congressional committee hearing in 1997 \((58, 59)\). Fear of breast cancer among women was high. Yalom \((60)\) pointed out that, while for most of human history, the breast symbolized nurturing and sexuality, in the latter part of the 20th century the breast came to symbolize death and mutilation, as poignantly illustrated in Figure 2 (a full-color version of this figure is available on the *Epidemiologic Reviews* Web site [www.epirev.oxfordjournals.org]).

My own awakening to the change occurred in 1993 when I chaired a National Cancer Institute scientific workshop to review evidence about the effectiveness of breast cancer screening among women in their forties \((61)\). We concluded that, after 7 years of follow-up, randomized trials had not shown an effect of screening and that more follow-up time was needed. To my surprise, the workshop’s conclusions appeared on the front page of the *New York Times* \((62)\). Soon after, I met with representatives of several breast cancer advocacy groups. They expressed frustration and anger that our findings were so different from the messages they had been receiving up until that time. Several participants thought the scientific community was patronizing women
and emphasized that scientific information should not be withheld from the public. I came away thinking that we in the population sciences were not doing an adequate job of communicating with the public about their legitimate health concerns. In the succeeding years, a perfect storm has formed around breast cancer screening: women’s anxiety, political interests, and media emphasis have caught many cancer screening scientists, who knew little or nothing about communication with the public, totally unawares.

In the 1990s and the 2000s, scientific interest in informed and shared decision making for cancer screening took form (63). Research articles, systematic reviews, and editorials emphasized the need for a discussion of both benefits and harms of screening and patient involvement in screening decisions. In breast cancer screening, the need seemed especially acute for women in their forties because expert groups’ recommendations varied, and absolute mortality benefits were lower while frequency of false-positive mammograms was higher for this age group. Women indicated they wanted discussions about screening with their clinicians (64). Suggestions about the content and communication methods were made (65, 66). However, relatively little evaluation has occurred; in a 2009 systematic review of decision aids for people facing treatment or screening decisions, not one of the 55 randomized trials reviewed dealt with helping women decide about breast cancer screening (67). The science of communicating with patients about breast cancer screening—developing and testing methods to communicate the multifaceted and complicated information patients need to understand to make decisions, and testing the methods—should be a high-priority area of research.

Communicating with individual patients is not the same as communicating with patients millions at a time. The experience of several scientific bodies making evidence-based recommendations for breast cancer screening should be sobering to all scientists interested in evidence-based health care practice. The public anger over breast cancer screening recommendations of the 1997 National Institutes of Health Consensus Conference (58, 68) and, more recently, the 2009 USPSTF (69, 70), demonstrate dramatically the challenges population scientists face in communicating with the public. In both cases, the US Senate passed legislation to override the recommendations. The Task Force was criticized for rationing care, not including radiologists and oncologists, and protecting insurers. Among the lessons learned, Wolff points out the following: “Scientists are wise to banish politics from their recommendations but are unwise not to plan for the political reception that awaits them” (70, p. 163). Not to do so may jeopardize the very existence of scientific groups making health policy recommendations.

Reaction to breast cancer screening recommendations may be more heated than to other cancer screening recommendations, but unless a recommendation is to initiate or continue screening, the public reaction is likely to be negative. If we want educated and involved health care consumers, cancer screening investigators and guidelines leaders must learn how to speak on the public stage and to the media. For breast cancer screening recommendations, we have not decided that communicating with the public is a major part of the health policy task. We have not learned well enough the rules and methods for public communication. We need to engage more professional coaches to help us hone these skills. We also have not often enough applied our scientific expertise and set up experiments to learn what methods work best when addressing the public. We must begin to do so. Schools of public health, medicine, and nursing could further such efforts by working with schools of journalism to develop health-communication disciplines—for faculty and students.

THE LONG-TERM FUTURE OF CANCER SCREENING

Breast cancer screening with mammography has saved many lives and has helped knock the peak off breast cancer mortality in the Western world. Throughout this perspective, I have suggested steps that might improve breast, and other, cancer screening. Strengthening the evaluation of new screening technology characteristics and using new strategies to determine mortality effects of screening can better clarify the effectiveness of cancer screening. Better analysis of false-positive results and overdiagnosis can help clarify the hazards of cancer screening. Communication skills must...
improve so that individuals and groups who hope to benefit from cancer screening understand the trade-offs.

Breast cancer screening has proved far more complicated than originally envisioned. To save lives, millions of women have had to undergo repeated testing over decades; millions have experienced false-positive results, most experiencing at least some subsequent worry and requiring additional health care visits, and thousands have been diagnosed and treated for a breast cancer that would not have caused them harm. This is not the story for just breast cancer; the same story is occurring with other cancers as screening becomes more common. If false-positive testing and overdiagnosis cannot be controlled—and my own concern is that they are likely to worsen with new tests—large numbers of, perhaps most, people are likely to experience one or both of these adverse events for at least one cancer during their lifetimes. Because most risk factors for most cancers are (thankfully) small, concentrating screening on small groups is unlikely to lead to detecting most cancers. There should be a better way to conquer cancer.

Screening is secondary prevention. Over the past 2 decades, primary cancer prevention has increased, with lifestyle changes, immunizations, chemoprevention, and even prophylactic surgery. The largest reduction in cancer deaths in the United States is not due to screening but to decreases in smoking, a lifestyle change that health services research and public policies helped make possible. Immunization against hepatitis B and hepatitis C promise major global prevention of hepatocellular carcinoma. Tamoxifen and raloxifine reduce breast cancer incidence in women with elevated risks of breast cancer. Prophylactic mastectomy and ovariectomy dramatically decrease breast cancer incidence in women with deleterious genetic BRCA1 and BRCA2 mutations. All of these primary prevention approaches are likely to be more and more successful, and to expand to other cancers, as research progresses over the next several decades.

Therapy for several cancers, including breast cancer, is increasingly effective. The original Canadian and US task forces recognized the interaction between screening and therapy, that screening must be linked to effective treatment. At the extreme case, conditions were identified for which, early on, there was no or ineffective therapy even though a screening test existed (e.g., for acquired immunodeficiency syndrome). On the other hand, I recall no discussions considering the possibility that screening would not be necessary if treatments cured a patient regardless of the stage of the disease. Over time, as treatments evolve, it is possible that we can move from a situation in which screening is helpful to one in which it is superfluous. This scenario is now beginning to emerge in the cancer screening literature.

Testicular cancer is a case in point. Treatment for testicular cancer is highly effective, with 10-year survival rates of 95% (71). The US Task Force concluded that screening asymptomatic men is unlikely to produce additional benefits and recommends against screening for testicular cancer (72). (Its review also found no evidence about screening for testicular cancer.)

Regarding breast cancer, a 2005 modeling exercise suggested that the 20% decline in mortality since 1990 was due about equally to screening and improved therapy (73). A recent study of community breast cancer screening in Norway found similar results (74). Comparing breast cancer mortality in regions that introduced screening of women aged 50–69 years, along with multidisciplinary care, with regions in which screening implementation had not yet occurred but multidisciplinary care was available, researchers found a 10% reduction in breast cancer mortality—much smaller than the 15%–32% estimate the USPSTF calculated using data from randomized controlled trials (15). In the most recent randomized trial of breast cancer screening, the 10-year mortality results in the AGE trial were not statistically significant, with lower-than-expected mortality in the unscreened group (75). A 2005 population-based case-control study in the United States also found no significant effect of breast cancer screening (76). In all these studies, many possible explanations exist for the results, including several methodological limitations. Nevertheless, as therapy for breast cancer improves, the ability to demonstrate screening effectiveness is becoming more difficult.

As therapy for breast cancer continues to improve, when is it no longer useful to screen? One commentator has suggested that if the mortality benefit for screening women 50 years of age is 0.4 woman per 1,000 women screened over 10 years, mammography screening should no longer be an indicator of quality of health care (77). Others, including the USPSTF (78), recommend shared decision making, with the woman and her health care provider reviewing the benefits and harms of breast cancer screening before she makes a decision.

Screening has been important in the fight against cancer, but decades of research on breast cancer screening have shown us it is an imperfect tool—with the need for repeated screens on millions of people over decades, false-positive results, overdiagnosis, and substantial cost. Cancer screening is likely to continue to be important in the years to come. Nevertheless, progress in primary prevention and treatment should, over time, decrease the need for cancer screening. We should all look forward to the day when better strategies for primary care prevention and more effective treatments so reduce the need for cancer screening that it can be relegated to medical history.

ACKNOWLEDGMENTS

Author affiliation: Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts (Suzanne W. Fletcher).

The author thanks Dr. Robert H. Fletcher for reading the manuscript and for helpful suggestions.

Conflict of interest: none declared.

REFERENCES


