Screening for cancer presents inferential issues that are among the most difficult in science. Whether screening should be recommended and for whom is seldom clear, despite extensive experience in randomized trials, in observational studies, and in ordinary practice. “Finding early” is the easy part. Cancers found early by screening and by improvements in screening technology may or may not be worth the finding.

In general, improvements in screening mammography lead to finding more cancers and/or decreasing the number of false-positive results by improving sensitivity or specificity, respectively. A more important consideration is whether the finding improves the patient’s lot, that is, is identifying a particular cancer and treating it earlier doing the woman a favor? Perhaps, the cancer was not lethal, and without the new technology, it would not have been found for many years. A woman having such a cancer could have spent these years “cancer free” with no decrement in survival. Indeed, the cancer might never have been found otherwise, resulting in overdiagnosis. We know that the incremental cancers found by mammography are biologically different from symptomatic cancers and not just because of a stage shift (1–5). The problem is that we do not yet understand the biology of cancer well enough to know which cancers are important to find early and which can be ignored.

Consider the growth of a breast cancer from the first tumor cell: after perhaps 27 doublings, the tumor becomes detectable on a mammogram. After approximately 29 doublings—but with substantial variability—the tumor becomes symptomatic or detectable other than by a mammogram. So the “lead time” provided by screening mammography is on the order of two doubling times. (Sometimes there is no lead time, as when cancers found in screening programs are interval cancers—detected between mammograms.) Regardless of the lead time, the window for finding cancers via mammographic screening is small in comparison with the life of the tumor.

If a mammographically detected tumor has already metastasized, then finding it early has little or no incremental benefit—notwithstanding the apparent lead-time benefit. If the tumor has not metastasized and still would not be metastatic when found clinically, then again there would be little loss in the delayed detection. So the relevant question is whether the tumor becomes metastatic in a window that is about two doubling cycles wide, about 7% of the tumor’s preclinical life. Eventually, biology may enable an accurate assessment of the metastatic potential of any particular tumor. But present ability to determine which 20% of breast cancer patients in the United States will die of their disease is limited.

The difficulty translating a small increase in sensitivity into a decrease in mortality is evinced by the comparison of annual vs biennial screening. Screening yearly rather than every 2 years increases the number of cancers found and finds some tumors a year earlier, but the impact on mortality is small (6).

What happens when computer-aided detection (CAD) software is used with screening mammography? The second “pair of eyes” finds additional cancers, but it opens the lead-time window at most a small fraction of a doubling time. The possibility that a tumor would become metastatic precisely in any such short period of time is remote. So small is the fraction of cancers that would become metastatic in such a small interval—if there are any—that it would be difficult to impossible to detect an incremental mortality advantage of CAD even in a large randomized trial.

Moreover, improving sensitivity may preferentially find less aggressive tumors, or it might find more of those tumors that would have otherwise revealed themselves as interval tumors. Neither type of increments in sensitivity could have much of an impact on breast cancer mortality.

The other side of the sensitivity coin is specificity. Additional looking increases the rate of false-positive results along with the rate of true-positive results. The rate of false-positive results is relatively easy to assess in a comparison study. But the impact of false-positive results is less clear and indeed may vary considerably from one woman to another.

The important article by Fenton et al. (7) in this issue of the Journal updates a previous publication (8) on the impact of CADs in screening mammography programs of the Breast Cancer Surveillance Consortium (9) (http://www.seer.cancer.gov/studies/endresults/study25.html). A limitation of the earlier report was that there were relatively few CAD screens considered: 31,186, with 103 invasive cancers detected (10). In terms of the total number of screening mammograms, the present update is about twice as large as the earlier report, but it includes about 8 times as many CAD screens: 253,426, with 761 invasive cancers detected. Moreover, the update represents more mature use of CADs and therefore is likely to better represent steady state.

The Fenton et al. (7) study chronicles the performance of CADs with digitized film mammography. The conclusions are negative and qualitatively similar to those of the authors’ earlier report. The study is not randomized but compares like with like as well as possible under the circumstances of “real-world practice.”
The primary comparison is mammographic performance in those facilities that switched to CAD use, comparing before vs after the switch. Time and factors related to time are of course different. To at least partially control for these differences, the authors compare mammographic performance over time in those facilities that did not switch to CAD.

In their update, Fenton et al. (7) found a very small but statistically significant decrease in specificity (91.9% before CAD vs 91.4% after CAD) and a slight numerical but not statistical increase in sensitivity (79.7% before CAD vs 81.1% after CAD). This numerical increase is because the number of ductal carcinomas in situ diagnosed increased, which arguably is mostly over-diagnosis and will not likely translate into a mortality reduction. Curiously, the authors found a numerical decrease in sensitivity for detecting invasive cancers.

The positive predictive value of mammography was 3.6% after CAD, a statistically significant relative decrease compared with 4.3% in those same facilities before CAD. On the other hand, and somewhat paradoxically, in those same facilities, CAD led to a statistically significant 18% relative decrease in the rate of biopsies recommended, down from 1.27% to 1.04% per screen. However, “after adjusting for the Breast Cancer Surveillance Consortium registry, patient characteristics, HT use, and interpretation year, CAD use was not associated with statistically significant differences in the odds of biopsy recommendation (odds ratio = 0.99, 95% confidence interval = 0.93 to 1.05, \( P = 0.66 \)).”

The “real-world” study by Fenton et al. (7) is an example of comparative effectiveness research. Such studies can be more relevant for policy and practice than controlled efficacy studies. A case in point is the lack of improvement in sensitivity in the Fenton study. Perhaps, it reflects a flaw in their observational study, one that cannot be identified and therefore for which no adjustment is possible. But there are other plausible explanations. A standard way to conduct a controlled CAD study is to have readers consider mammograms without CAD, record their calls, and then add CAD, updating their calls. The independent “second read” increases the rate of cancers found. But suppose readers were to go directly to the CAD marks. (I do not know that this happens in the “real world,” but I might do it, especially if I had a large backlog of mammograms to read.) The reads apart from CAD would not be independent of CAD, and hence would not be second reads at all. The scenario would be similar to having single reads and the sensitivity would be little changed—as in the Fenton et al. study. Moreover, although studies indicate that CAD requires more time to read mammograms (11,12), there is some evidence that experienced readers require no additional time (12).

The platform of most modern CAD is digital mammography rather than digitized film mammograms, the latter being the case in the Fenton et al. study (7). CAD used with digital mammograms—as for example in the DMIST study (13)—may have different characteristics.

CAD is only as effective as its computer software. There is a notion in today’s world that computers are infallible. And Deep Blue’s winning ways in top-level chess competition has reinforced this notion. Reality is that it is not easy to program a robot to recognize patterns. The algorithms used may improve with time and experience. Indeed, a goal of artificial intelligence (AI) algorithms (as used by Deep Blue, for example http://en.wikipedia.org/wiki/Deep_Blue_(chess_computer)) is to enable a computer to “learn” over time through its experience and reinforcement of whether it was right or wrong. Although AI has the potential to enhance performance (14), I do not believe that any standard CAD software uses AI in an online fashion. An online AI approach would raise unusual (and interesting!) regulatory issues because such an algorithm could well call the same mammogram differently depending on when it is presented to the CAD device. And just as is true for human readers, two identical devices might call the same mammogram differently depending on their experiences (unless all devices of the same variety are connected and are updated with a common history).

In commenting on the earlier report of the Fenton et al. study, Hall (10) recommends “the conduct of larger, controlled studies of computer-aided detection that assess not only cancer diagnosis but also the gold standard: mortality. But such studies will be expensive, controversial, indeterminate, or quickly passé.” Indeed, these latter points argue against the recommendation. Moreover, as I have already suggested, any mortality benefit would be so small as to make its detection virtually impossible.

Hall asks rhetorically, “Will the results of Fenton et al. [the 2007 report] end the use of computer-aided detection in screening mammography?” and answers his own question: “Of course not.” He was right. According to Rao et al. (15), the proportion of screening mammograms in the United States using CAD increased from 39% in 2004 to 74% in 2008.

Why is CAD so popular? An obvious reason is that it is built into digital mammography equipment, which is increasingly common in the United States. Another is financial: In 2008, Medicare’s global reimbursement for CAD was $16.50 (15). Still another is that CAD marks are comforting to the reader, even though the comfort may be misplaced. In a related vein, relying on CAD marks—or the absence of same—in medical malpractice suits may be an effective defense, although the sword has two edges (16).

An argument for the use of CAD with film or digital mammograms is that it will get better over time. Fine. Researchers and device companies should work to make the software ever better. But this should happen in an experimental setting and not while exposing millions of women to a technology that may be more harmful than it is beneficial. In the meantime, economic incentives may stoke its continued proliferation.

References


**Affiliation of author:** Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX.