Commentary

Failure of Researchers, Reviewers, Editors, and the Media to Understand Flaws in Cancer Screening Studies

Application to an Article in *Cancer*

Donald A. Berry, PhD

Observational studies present inferential challenges. These challenges are acute in cancer screening studies, in which lead-time and length biases are ever present. These biases can make any study worthless. Moreover, a flawed study's impact on the public can be deleterious when its conclusions are publicized by a naïve media. Flawed studies can also make the public learn to be wary of any article or reports of articles claiming to be scientific. Here, the author addresses these and related issues in the context of a study published in *Cancer*. **Cancer 2014;120:2784-91.** *© 2014 American Cancer Society.*

KEYWORDS: cancer screening, observational studies, screening mammography, lead-time bias, length bias, randomized screening trials, efficacy of screening mammography by age.

INTRODUCTION

This issue of *Cancer* includes an observational study by Webb et al,[1] which I will call simply the Webb study. When the article appeared online, it was widely covered in the media. Typical headlines claimed "Study Suggests Earlier Mammography May Benefit Younger Women" and "Mammograms Should Begin at Age 40, Researcher Says." Over the last 20 years or so, many journals, including *Cancer*, have published articles promoting cancer screening. These articles have typically been overly optimistic about the benefits of screening. And, as in the case of the Webb study, some authors have challenged the guidelines of the US Preventive Services Task Force. The public learns to become wary of all studies and guidelines. This attitude is exacerbated by the media, whose motives are frequently more consistent with inflaming than informing.

In 2002, an erudite editorial in the *Chicago Tribune*[2] criticized the publication in *Cancer* of an earlier observational study. That earlier study also promoted screening mammography, and it too was widely covered in the media. The editorial could as well have been written in response to the publication and media coverage of the Webb study. It said in part:

> **. . .**Millions of women woke up 1 day last week to news that the last word had been written in the debate over mammography screening. Newspapers, TV and radio stations reported what many described as the definitive study, proving beyond doubt that getting routine mammograms sharply reduces a woman's chances of dying of breast cancer.
>
> **. . .**Too bad it wasn't true. Once again, many in the media fell victim to the well-intentioned authors of a new study without questioning the study's merits or calling an independent expert for a second opinion. Hey, it was published in a scientific journal—it must be right.
>
> **. . .**Most reports of the mammography study also failed to note it was published in the journal of the American Cancer Society—a strong believer in screening—and funded by the same group.

Corresponding author: Donald A. Berry, MD, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030; Fax: (713 563-4243); dberry@mdanderson.org

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas

See commentary, editorial, and related article on pages 2792-9, 2800-2, and 2839-46, respectively, this issue.

I acknowledge and congratulate the editor of *Cancer*, Dr. Fadlo Khuri, for inviting and publishing a commentary while recognizing that it would be critical of a journal under his purview. Dr. Khuri and the *Cancer* editorial board may well disagree with much of my criticism and with the harshness of my criticisms. Their willingness to have this commentary appear on pages they control represents a truly refreshing attitude for any journal and it bodes well for the future credibility and reputation of *Cancer*.

DOI: 10.1002/cncr.28795, **Received:** October 10, 2013; **Revised:** November 7, 2013; **Accepted:** November 25, 2013, **Published online** June 12, 2014 in Wiley Online Library (wileyonlinelibrary.com)

...The upshot was that, once again, women concerned about their health got confusing messages.

A purpose of the current commentary is to describe some of the flaws in the Webb study and in other observational studies of screening effectiveness. Drawing conclusions from observations is difficult, and the difficulties are especially severe in studies concerning screening's effectiveness. Although many published studies have methodological flaws, some are limited, and the conclusions have a grain of truth. Screening studies can be so flawed as to make the study meaningless or worse. Authors, reviewers, and editors may have been well intentioned as the *Tribune* suggests about the authors of the earlier article. But coverage by scientifically naïve media can make a study such as that by Webb et al detrimental to the health and well being of women. Some in the media are less well intentioned and are more interested in attracting readers and listeners by engendering controversy than they are interested in getting at the truth, which is a disservice to their audiences.

Lead-Time and Length Biases

To understand some of the flaws of observational studies of cancer screening, it helps to visit the early history of empirical research in screening. Hundreds if not thousands of published articles fueled the early enthusiasm for screening by comparing the outcomes of screen-detected cancers with those of cancers detected symptomatically. It was easy to demonstrate a statistically significant advantage in survival, for example: all that was required was to have a modest sample size.

Screen-detected tumors have better prognoses, partly because of screening's stage shift. And the differences are dramatic. However, over time, journal reviewers and editors have come to understand that these differences are subject to huge lead-time and length biases. Lead-time bias is easy to understand and communicate: labeling individuals earlier as having cancer means they will live longer after having been so labeled. However, it is difficult to tease out of the data whether any of this apparent benefit is real.

Length bias is much more subtle than lead-time bias but is even more important. Cancers that are slowly growing have a longer period between the time a tumor can be detected on a mammogram and when it will become evident from symptoms. This is called the tumor's *screening sojourn period*. The longer the sojourn period, the greater is the opportunity for the tumor to be identified by screening. So screening preferentially detects slowly growing tumors. Hence, screen-detected tumors have better prognoses quite apart from any benefit of early detection.

These are formal definitions: *Lead-time bias* is the increment in survival among screen-detected cases that equals the difference in time of detection and when the cancer would have been detected otherwise. *Length bias* is the increment in survival among screen-detected cases because of the over selection of slowly growing cancers.

Both lead-time and length biases contribute to screening's stage shift. What is surprising is that a screen-detected breast cancer has a better prognosis than a seemingly identical symptomatically detected cancer, including 1 that has the same stage of disease. So there is what might be called "within-stage shift." An implication is that the method of detection is an independent prognostic factor in breast cancer, and it will continue be an independent prognostic factor until our ability to determine prognosis based on molecular markers of tumors improves sufficiently (see Berry[3] and its references).

Of course, cognoscenti such as Sam Shapiro have always understood the difficulties of drawing conclusions about the benefits of screening from observational data.[4] Recognition of these difficulties gave rise to the attitude among researchers that essentially the only way to separate truth from bias was to randomize individuals to screening versus not screening. This attitude, in turn, led to designing and conducting randomized trials that addressed whether mammography reduces breast cancer mortality. Such trials are difficult to design and conduct, and all have imperfections, but they get as close to the truth as is possible. Readers interested in building a stronger background in *cancer screening* will benefit from the excellent edition with that title by Kramer et al.[5]

The Webb Study

Despite widespread acceptance in the research community of the importance of lead-time and length biases, their relevance is not appreciated by all authors, and apparently not by Webb et al. Theirs is an unusual type of observational study, but it is subject to observational biases. Webb and colleagues report a population-based study of 609 women at 2 hospitals in Boston who were proven to have died of breast cancer. These women's cancers had been diagnosed between 1990 to 1999, and the women had died in follow-up before or during 2007. Some of these 609 women had had regular screening mammograms before or at detection of their disease, and others had not been regularly screened. Webb et al draw conclusions about the benefits of screening, especially as these benefits relate to age, based on their observation that 71% of these women had not received regular screening, which they defined as at least biennial. This is despite the authors' claim that more women in Massachusetts participated in regular screening than did not.

Regarding the prevalence of screening, Webb and colleagues say that "a study at Massachusetts General Hospital estimated the rate of screening in women older than 40 years to be 80%, consistent with the Behavioral Risk Factor Surveillance System (BRFSS) survey (J. S. Michaelson, unpublished data)." However, they report neither the proportion of screening in BRFSS nor whether screening in BRFSS meant at least biennially, as it did in their study. Moreover, screening programs detect more cancers at the initial screen than at any particular subsequent screen. Defining regular screening in such a way that requires at least 2 mammograms means that cancers detected at first screen would be assigned to the nonregular-screening group.

Webb et al do not indicate the proportion of women in their full sample of 7301 breast cancer patients who had regular screening; and, indeed, assessing this proportion does not seem to have been part of their study. They do say that the BRFSS survey "may overestimate screening rates by as much as 20%." Indeed, to my knowledge, the best estimate for the prevalence of regular mammographic screening over this decade is 50%.[6] I use this estimate in the simulations discussed below, although this is almost certainly an overestimate for the Webb study in view of my above comments.

Before I address the roles of lead-time and length biases in the context of the Webb study, I will consider another fundamental issue.

Restricting Conclusions to Women Who Died of Breast Cancer

According to Webb et al, "Our method, a failure analysis, estimated the effect of screening mammography on mortality by review of the screening histories of women who died from breast cancer." Drawing useful inferences about the benefits of screening by conditioning the analysis on women who died of breast cancer is not logically possible.

To understand the logic of any argument, it helps to consider extremes, even if those extremes are unrealistic. Suppose that all mammographically detected cancers are overdiagnosed. Then there would be no breast cancer deaths among women with mammographically detected tumors. So apart from interval cancers this would suggest that screening is a panacea for breast cancer, whereas, in fact, it would be harmful.

Making a less extreme and more realistic assumption, suppose there is a modest to moderate level of overdiagnosis. This too will lead to an imbalance of deaths, with more than "expected" in the nonscreening group. And this statement is true whether or not screening has any benefit. Indeed, someone with logic symmetric to that of Webb et al, and equally defective, would conclude from the Webb study that screening is a disaster because it results in overdiagnosis with no compensating benefit.

I mentioned overdiagnosis, but the deficiencies in the Webb approach do not rely on assuming that there is any level of overdiagnosis associated with screening mammography. Lead-time and length biases (the latter being a generalization of overdiagnoses) are quite sufficient to exaggerate any imbalance, as I discuss below.

Failure analyses cannot enable quantifying risks and benefits, because they do not consider a relevant denominator. Moreover, no statistical conclusions are possible when using failure analyses of screening studies, nor do Webb and colleagues attempt to draw any formal statistical conclusions. But they do ask the reader to make a leap in logic as though their arguments actually enabled such a leap.

A failure analysis is appropriate in some scientific settings. For example, suppose hundreds of dead fish wash up onto the shores of a lake. What could cause such an unusual observation? Researchers may reasonably examine the fish to find out why they died. But suppose there is a nearby lake with half as many dead fish. A Webb-like conclusion is that a fish of a particular variety would be luckier to have been in the second lake, but such a conclusion cannot be drawn.

It is disappointing to see medical research involve an analysis that restricts consideration to patients who have died and that draws conclusions ignoring the patients who lived. Medical researchers have been taught that a diagnostic test's sensitivity is different from the test's positive predictive value. The former is the probability of a positive test result given that the patient has the disease in question. The latter is the reverse: the probability the patient has the disease in question given a positive test result. (Bayes theorem and an assumed disease prevalence, depending on the subject's characteristics, are required to disentangle the relationship.)

For the same reason, knowing only the proportion of patients whose tumors had been detected mammographically among those who died of breast cancer does not allow for calculating the proportion of patients who will die if their tumors are detected mammographically. More important, it sheds no light on the clinical question of whether a particular woman or women in a particular age group are less likely to die of breast cancer if they get screening mammograms.

I grant that this issue is subtle and that it is counterintuitive for many people. There is a tendency to make very wrong inferences in conditioning arguments. The consequences go beyond sleight-of-tongue mathematical "paradoxes." For example, in other disciplines, "reversing the conditional" has led to egregious and devastating mistakes in logic. One such reversal in legal settings is known as "the prosecutor's fallacy."[7,8] Suppose a criminal's blood that was spilled at the crime scene has rare genetic characteristics, occurring, say, in only 1 per 342 million people. Suppose that a suspect's blood is a match for the criminal's, a 1-in-342 million match. The prosecutor's fallacy is to conclude that the probability the suspect is innocent is 1 in 342 million. In a case in 2003 in the Netherlands, this fallacy with this actual probability led to the wrongful imprisonment of a surgical nurse. Lucia de Berk was convicted of murdering 7 patients who had died while having operations at which she was present.[9,10] She was fully exonerated at second appeal, but not until 7 years later, in 2010.

The 2 sets of flaws in the Webb study—the biases involved and the inappropriateness of failure analysis—are quite separate. They are independent, and either set of flaws is sufficient to negate the authors' interpretations from their study.

Simulating Deaths in a Cohort of Screen-Detected Cancers

The above arguments regarding biases and failure analyses may seem academic to some readers, but they are very real. To put quantitative teeth into some of these arguments, I simulated from a simple model of the Webb study. The point of the simulations was not to imitate Webb and colleagues exactly but to demonstrate what results would be possible when proceeding backward as they did, focusing only on those patients who died of their disease. The distinction between the Webb study and my simulations is that, in the latter, I know the benefit of screening exactly, because I constructed the simulations to have a very specific benefit: zero.

I made some assumptions in developing the simulation model. Like all modeling assumptions, they may be wrong. But modifying the assumptions of the model within a rather broad range will change the conclusions very little.

Although the simulations produce numerical conclusions, the actual numbers are not important. What is important is that it is easy to get numbers similar to those of Webb et al even in the case when screening has no benefit at all.

Another objective of these simulations is to clarify issues associated with lead-time and length biases. An advantage of simulations is that they allow for comparing the outcomes of screening versus nonscreening for the same "virtual women" by first assuming that the women get regular screening and then assuming that they get no screens. I assume equal numbers of breast cancers detected mammographically as detected symptomatically in the full study.

I simulated sojourn periods for breast cancers of women in the context of a screening program of biennial mammography. I then considered these same women when not screened (see below). Each tumor becomes mammographically detectable at the start of its sojourn period, and it becomes clinically detectable at the end of its sojourn period.

The distribution of sojourn time in breast cancer is unknown. Sojourn time varies, depending on factors such as age and breast density. I incorporated variability by assuming a distribution of sojourn times. In particular and consistent with assumptions of some other modelers, I assumed that sojourn times have an exponential distribution.[11] However, in making the point of these simulations, the assumption regarding any particular distribution of sojourn times does not matter. All that matters is that sojourn times vary from 1 tumor to another.

An exponential distribution is defined by its mean (which is the inverse of hazard rate). Which mean to assume? The mean sojourn time in breast cancer is not known precisely, and it depends on the population being considered. Typical estimates range from 4 years to 7 years.[12] I assumed a mean sojourn time of 5 years, although the exact value assumed has little effect on the conclusions.

Assuming an exponential distribution implies that all sojourn times are finite. With a mean of 5 years, almost every sojourn ends before the woman dies from causes other than breast cancer, especially in younger women. Hence, these simulations assume no overdiagnoses. We know that overdiagnosis exists, although its magnitude is uncertain.[13] Hence, these simulations have a major bias in favor of screening.

There is another important bias favoring screening in these simulations. The simulation model generates sojourn times that are positive, whereas some actual cancers have negative sojourn times. Namely, some cancers—especially in dense breasts—become symptomatic before they can be detected on a mammogram.

I generated sojourn periods starting at time points that were randomly distributed over a 30-year period, from year −20 to year 10. To emulate the Webb study, which considered breast cancers diagnosed over a particular decade (1990-1999), I restricted the simulation to those sojourn periods that ended after year 0. And, of these, I further restricted the simulation to those cancers that would have been detectable in a biennial screening program, with mammograms on the first day of years 2, 4, 6, and 8. The horizontal bars in Figure 1 represent the first 50 of the simulated sojourn periods.

The timings of the 4 screens in the program of biennial screening are illustrated in Figure 1 as vertical dashed red lines. The cancers detected by screening are those whose sojourn periods intersect 1 or more of the dashed red lines. Because the cancer will be detected at the first screening during its sojourn period, the only intersection that matters is the first. However, for completeness, Figure 1 illustrates entire sojourn periods.

Four of the 50 cancers in Figure 1 (represented by blue bars) do not intersect any of the dashed red lines. They are called interval cancers, because their sojourn periods are wholly between 2 adjacent screens. They were detected clinically before they could be detected mammographically.

Webb et al considered patients who died of breast cancer by 2007, analogous to year 17 in the simulations. Deciding which of the 50 sojourn times would result in breast cancer death by year 17 in Figure 1 requires an assumption. Sojourn times are roughly comparable to some number of tumor-doubling times. A sojourn time is longer when the tumor has a longer doubling time. Such tumors are relatively slow-growing and are less likely to be lethal. So patients who live longer tend to have longer sojourn times. The precise relation between sojourn time and survival time is not known, and factors other than sojourn time affect survival. Such factors include molecular characteristics of the tumor and therapy the patient receives. Therapy may depend on and may interact with the tumor's molecular markers and with the sojourn time itself.

I assumed a simple relation between sojourn time and survival. Namely, patients survive their breast cancer for 6 sojourn times from the time the tumor is initially detectable mammographically. (Moderate changes in the multiple 6 have no effect on the message of this example.) Implicit in this assumption is that breast cancer mortality does not depend on when the tumor was detected or, indeed, on whether it was detected by screening.

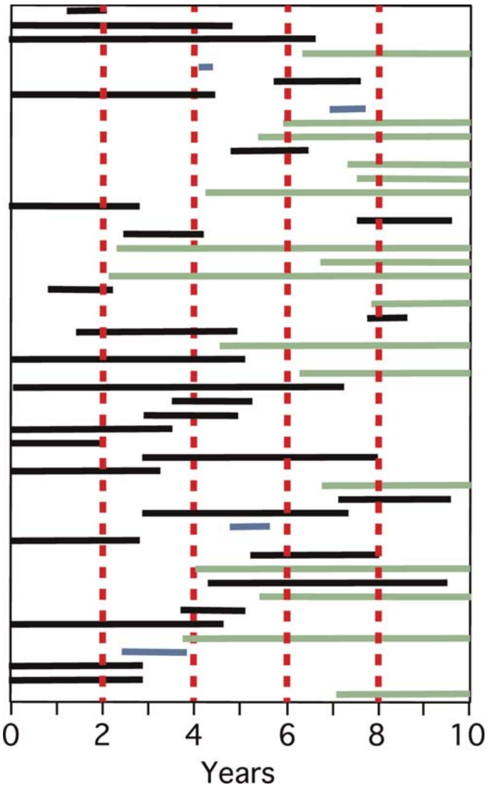

Figure 1. The horizontal bars show simulated sojourn periods for 50 breast cancers as described in the text. The vertical red lines indicate the biennial screening times at years 2, 4, 6, and 8. Intersections between vertical and horizontal lines are times at which the respective cancers would be detected by a mammogram. The blue bars represent interval cancers. The green bars represent cancers whose sojourn periods overlap year 10, so they would not have been detected during the 10-year period in the absence of screening. The black bars have right endpoints before year 10, so these represent screen-detected cancers that would have been detected before year 10 in the absence of screening.

Forty-six of the 50 simulated cases illustrated in Figure 1 were screen-detected and are represented as black or green bars (the distinction between the 2 colors is described below). The other 4 cases are the interval cancers mentioned above. Of the 46 screen-detected cancers, 9 were recorded as deaths by year 17 using the 6-sojourn-period assumption. All 4 of the interval cancers in Figure 1 were recorded as deaths by year 17. So there were 13 deaths among the 50 screened women. The proportion of interval cancers (4/13 = 31%) in this group, assuming regular screening, is similar to 34% of the 178 breast cancer deaths in the Webb et al screening group (see their Table 3).

Simulating Deaths in a Cohort of Cancers Detected in the Absence of Screening

Now, suppose the 50 patients represented in Figure 1 had not been screened. Assuming no effect of screening means that their outcomes would have been the same as if they had been screened. But only a subset of the tumors would now be detected between years 0 and 10: those with right endpoints of their sojourn periods before year 10. These 33 cancers are represented in Figure 1 by either a black bar or a blue bar. The other 17 cancers, those with sojourn periods represented by green bars in the plot, would not be detected by year 10 in the absence of screening. So these cancers would not be included in the nonscreening group. The green bars tend to be long, representing less aggressive tumors, and none of the respective 17 patients were recorded as deaths according to the 6-sojourn-period survival assumption.

To have the same sample size in a nonscreening group as in the screened group requires replacing the 17 sojourn periods in green with those generated in the same fashion as the others but having right endpoints earlier than year 10. Any deaths in these additional 17 patients will mean more deaths in the nonscreening group than in the screened group. The first simulation that I ran of these additional 17 patients had 15 additional deaths for a total of 28 in the nonscreening group. Simulations that had 7 to 12 deaths were more typical. Restricting to the deaths in the 2 groups, the proportion of the total in the nonscreening group is 20/(20 + 13) = 61% when there were 7 deaths among the 17 and 28/(28 + 13) = 68% when there were 15 deaths among the 17, similar to the percentage 71% reported by Webb and colleagues.

I emphasize that the particular numerical calculations are unimportant. What matters are the concepts. And recall that I made several conservative assumptions in this example. Modifying the simulations by 1) including overdiagnosed cancers, 2) including cancers with negative sojourn times, and 3) moving cancers detected at first screen from the regular-screening group would lead to a proportion of nonscreened patients among the deaths that is greater than 71%.

A distinction between breast cancers detected mammographically and those detected otherwise in this example is that the latter patients live for 5 sojourn times after detection, whereas the former patients live for 5 sojourn times plus the remainder of their actual sojourn period after detection. This remainder is the lead-time due to screening. Length bias results because the sojourn times for mammographically detected cancers are longer.

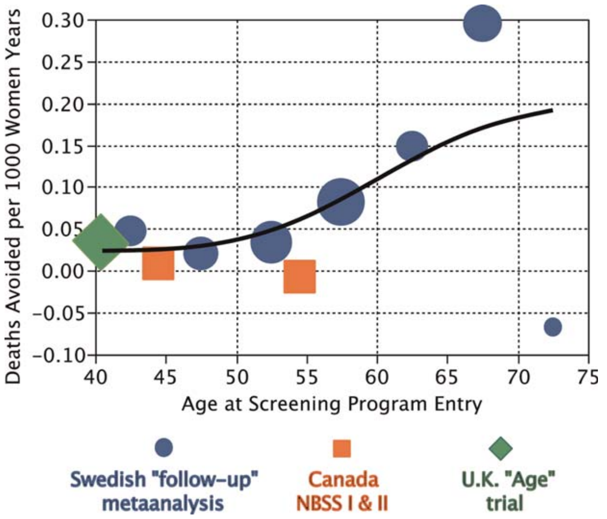

Figure 2. Deaths avoided per 1000 women years in the randomized trials are illustrated depending on age. Areas of the symbols in the plot are proportional to the respective numbers of breast cancer deaths. The smooth curve is a spline fitted to the data in the plot, weighted by numbers of breast cancer deaths. There were few women aged >70 years and, consequently, few breast cancer deaths in these women; the spline appropriately discounts the negative estimate of effect when restricting to this subset. The spline estimates at ages 45 years, 55 years, and 65 years are 0.026, 0.064, and 0.156, respectively. There are no obvious differences by country. NBSS I & II indicates Canadian National Breast Screening Studies 1 and 2.

Comparisons of the Simulations With the Webb Study Observations

By construction, there was no benefit of screening in this example. The greater number of deaths in the nonscreening group is pure bias. Similarly, the Webb study results are consistent with no benefit of screening, both overall and within particular age groups.

It was easy to build a model assuming no mortality benefit for screening that yields results similar to those of Webb et al. Of course, although these simulations assumed no benefit of screening, they do not indicate that screening has no benefit. They were designed only to demonstrate that Webb and colleagues' conclusions were logically flawed and that the study enabled no conclusions about the benefits of screening.

What Do the Randomized Controlled Trials Conclude Regarding Benefits of Screening by Age?

Webb et al claim that younger women benefit from screening more than older women. This spurious observation is caused by differential biases of the types described above, depending on age, but mostly because of different proportions of women getting screened in different age groups.

The randomized controlled trials (RCTs) conclude just the opposite effect of age from that of Webb and colleagues, and compellingly so. Although the RCTs have biases regarding the effectiveness of screening (see below), the biases affect the various age groups similarly. Figure 2 is adapted from Berry[14] and illustrates the evidence from the randomized trials regarding the effect of screening on breast cancer mortality, depending on age. The measure of effect is the number of deaths avoided by screening per 1000 woman-years in the screening program. The area of each symbol in Figure 2 is approximately proportional to the total number of breast cancer deaths in the respective trial or category.

The source for the results from the Swedish trials in Figure 2 (blue circles) is the meta-analysis by Nystrom et al,[15] who provided the results by 5-year categories of age (each symbol is set at the midpoint of the respective age category). The Swedish trials dominate in Figure 2, because those trials were relatively large, especially when combined into the meta-analysis. For reasons indicated below, the values plotted are from Nystrom and colleagues' "follow-up analysis."

The Canadian trials (Canadian National Breast Screening Studies 1 and 2) (Fig. 2, red squares) focused on 10-year categories.[16,17] The UK Age 40 trial[18] is indicated by a green diamond. Figure 2 does not include the US Health Insurance Plan (HIP) trial or the Edinburgh trial. Both were assessed as poor quality by the Cochrane Collaborative,[19] and the results from these 2 trials have little or no credibility. However, if these 2 trials were included in Figure 2, they would not be obvious outliers.

On the basis of deaths avoided by screening per woman-year, the estimated benefit from the randomized trials is 6 times as great for an individual aged 65 years as for an individual aged 45 years. This very clear effect of age is opposite from that claimed by Webb et al, as discussed above. In addition, Figure 2 indicates that the results across the 3 countries are quite concordant when accounting for age. This plot supports the hypothesis that screening reduces breast cancer mortality and also makes it clear that the benefit is minimal for younger women and strongest for women in their 60s.

Webb et al claim that "Meta-analyses of RCTs . . . underestimate effectiveness of screening mammography . . . ." They refer to compliance and contamination biases. These biases are real. But they overlooked a major bias in the opposite direction in the Swedish RCTs: "evaluation bias." There was a single screening mammogram in the control groups of most of the Swedish RCTs that was planned to occur when the last mammogram would have occurred had the woman been assigned to the screened group or, more accurately, to the group invited to be screened. The goal was to count only those deaths from cancers detected in both groups up until or at this last mammogram. Unfortunately, the timing of the control mammogram slipped by as much as 14 months. Evaluation bias results from having a longer period to count cancers that may result in breast cancer deaths.[20]

The Nystrom et al meta-analysis of the Swedish trials in 2002 appropriately adjusted for this evaluation bias by restricting consideration to deaths only from those cancers detected up until a particular date: December 31, 1996. They called this the follow-up analysis, which I used in making Figure 2. A follow-up analysis is more appropriate than an evaluation analysis because it eliminates evaluation bias. The follow-up analysis, as expected, is less optimistic about the reduction in breast cancer mortality caused by screening. For example, for women in their 40s, the reduction caused by screening was 20% using the evaluation analysis and only 9% (confidence interval, −9% to 24%) using the follow-up analysis.

Studies Published in Many Journals Have Failed to Appreciate Screening Biases

Cancer is hardly unique in publishing flawed studies of cancer screening. I have publicly criticized another journal for an egregious example in which researchers who failed to understand lead-time and length biases demonstrated that women in their 80s lived longer if their breast cancer was been detected mammographically.[21,22] In that case too, naïve media were duped. They advertised this study as demonstrating that women in their 80s live longer if they get mammograms and that "mammography benefits may have no age limit and that older women should consider being screened on a regular basis, generally every 1 to 2 years."[23]

Can We Learn Anything about Screening Effectiveness From Nonrandomized Studies?

The short answer is yes. But deriving something of value from observational studies of screening is very difficult. There are biases at every turn. For example, even if one is able to control for lead-time and length biases—no small task!—then selection bias looms. Women who get screened for cancer are different from women who do not. Some of these differences are known and measurable. Other differences that may be of even greater importance are neither measurable nor known.

One credible approach that does not use randomization is modeling. An example is provided in the report by Berry et al.[6] Those modelers used mathematics and simulations of individual patients (in a way that is more sophisticated than mine above). Some of the modelers incorporated natural history models of tumor growth. The modelers used comprehensive longitudinal databases and extensive cross-sectional databases from a variety of sources. These databases contained information about prevalence of disease; growth rates of tumors; the patterns of screening by age; characteristics of tumors depending on method of detection and patient characteristics, such as age; and population breast cancer mortality depending on these characteristics. The models also incorporated information about the use of adjuvant therapies and their benefits depending on patient characteristics. These models have been used by the US Preventive Services Task Force to aid in understanding whether the benefits of screening observed in the RCTs extend into the modern era of adjuvant therapy, the relative effects of annual versus biennial screening, and the effectiveness of screening depending on age.[24]

Epilogue

The study by Webb et al uses flawed arguments to draw inappropriate conclusions. The legitimacy of the study was promoted through broadcasts by segments of the media. Some media may have been duped because, “Hey, it was published in a scientific journal—it must be right.” But some of the guilty media knew better and went along because “the other networks are covering the article.” Both attitudes reflect poorly on the state of scientific reporting in the media.

The Webb study has negative implications regarding the level of science in editorial processes, but it also has unfortunate ramifications for the very cause the authors endeavor to promote. Evidence from randomized trials makes it clear that mammographic screening is an important option for women who are truly and fully informed about its benefits and risks. Promoting the benefits of screening based on misleading and flawed publications weakens its scientific basis and confuses the public.

FUNDING SUPPORT

This work was supported in part by the National Cancer Institute under grant 1U01CA152958-01.

CONFLICT OF INTEREST DISCLOSURES

The author made no disclosures.

REFERENCES

1. Webb ML, Cady B, Michaelson JS, et al. A failure analysis of invasive breast cancer. *Cancer*. 2014;120:2839-2846.
2. Editorial. Women and mammograms. Chicago Tribune. August 6, 2002. Available at: http://articles.chicagotribune.com/2002-08-06/news/0208060134_1_routine-mammograms-regular-mammograms-mammography. Accessed October 5, 2013.
3. Berry DA. The screening mammography paradox: better when found, perhaps better not to find. *Br J Cancer*. 2008;98:1729-1730.
4. Shapiro S, Strax P, Venet L. Evaluation of periodic breast cancer screening with mammography. Methodology and early observations. *JAMA*. 1966;195:731-738.
5. Kramer BS, Gohagan JK, Prorok PC, eds. Cancer Screening. New York: M. Dekker; 1999.
6. Berry DA, Cronin KA, Plevritis SK, et al; for the Cancer Intervention and Surveillance Modeling Network (CISNET). Effect of screening and adjuvant therapy on mortality from breast cancer. *N Engl J Med*. 2005;353:1784-1792.
7. Berry DA. DNA, statistics and the Simpson case. *Chance*. 1994;4:9-12.
8. Berry DA. The science of doping. *Nature*. 2008;454:692-693.
9. Buchanan M. The prosecutor’s fallacy. New York Times. May 16, 2007. Available at: http://buchanan.blogs.nytimes.com/2007/05/16/the-prosecutors-fallacy/?_r=0. Accessed October 5, 2013.
10. Goldacre B. Lucia de Berk—a martyr to stupidity. *The Guardian*. April 9, 2010: Available at: http://www.badscience.net/2010/04/lucia-de-berk-a-martyr-to-stupidity/. Accessed October 5, 2013.
11. Duffy SW, Nagtegaal ID, Wallis M, et al. Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. *Am J Epidemiol*. 2008;168:98-104.
12. Weedon-Fekjaer H, Vatten LJ, Aalen OO, Lindqvist B, Tretli S. Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results. *J Med Screen*. 2005;12: 172-178.
13. National Cancer Institute. Physicians Data Query: Breast cancer screening. Available at: http://www.cancer.gov/cancertopics/pdq/screening/breast/healthprofessional/page1/AllPages£Section_423. Accessed October 5, 2013.
14. Berry DA. Breast cancer screening: controversy of impact. *Breast*. 2013;22(suppl 2):S73-S76.
15. Nystrom L, Andersson I, Bjurstam N, et al. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. *Lancet*. 2002;359:909-919.
16. Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up. A randomized screening trial of mammography in women age 40 to 49 years. *Ann Intern Med*. 2002;137:305-312.
17. Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. *J Natl Cancer Inst*. 2000;92:490-1499.
18. Moss SM, Cuckle H, Evans A, Johns L, Waller M, Bobrow L. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years’ follow-up: a randomised controlled trial. *Lancet*. 2006;368:2053-2060.
19. Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet*. 2000;355:129-134.
20. Berry DA. Benefits and risks of screening mammography for women in their forties: a statistical appraisal. *J Natl Cancer Inst*. 1998;90: 1431-1439.
21. Badgwell BD, Giordano SH, Duan ZZ, et al. Mammography before diagnosis among women age 80 years and older with breast cancer. *J Clin Oncol*. 2008;26:1-8.
22. Berry DA, Baines CJ, Baum M, et al. Flawed inferences about screening mammography’s benefit based on observational data [letter]. *J Clin Oncol*. 2009;27:639-640.
23. American Society of Clinical Oncology. Cancer.Net. Available at: http://www.cancer.net/cancer-advances-women-80-and-older-benefit-mammography-few-are-screened. Accessed October 5, 2013.
24. Mandelblatt JS, Cronin KA, Bailey S, et al. for the Breast Cancer Working Group of the Cancer Intervention and Surveillance Modeling Network (CISNET). Effects of mammographic screening under different screening schedules: model estimates of potential benefits and harms. *Ann Intern Med*. 2009;151:738-747.