-

Re: Benefits and Risks of Screening Mammography for Women in Their Forties: a Statistical Appraisal

The summary that appears on page 1415 [J Natl Cancer Inst 1998;90] describing the commentary by Donald Berry *(1)* is factually and importantly incorrect. Contrary to the summary, the statement issued by the Consensus Development Conference Panel on January 23, 1997, never acknowledged that there was a "statistically significant reduction in breast cancer mortality [for women ages 40–49 in the randomized, controlled trials of screening]." The panel did not even mention the new data (showing the benefit) that it had been convened to review. This failure was the reason that there was so much consternation at the end of the conference.

The panel also failed to notify women and their physicians that, if the data from the Canadian trial (which has been highly criticized for its poor design and execution) are not included, there is the same benefit for women in their forties as for those over the age of 50 years. They also never informed women that two of the trials showed statistically significant benefits of 35% *(2)* and 45% *(3)* in mortality reduction for these women.

The Journal is supposed to provide objective and factually correct scientific analysis. It should be careful about such loose statements, and it certainly should not engage in historical revisionism.

The Berry commentary itself *(1)* only adds to the confusion that has resulted in the controversy over screening women ages 40–49 years. His praise for the Canadian trial ignores the facts. Not only was that trial underpowered *(4)*, but its randomization lacked blinding *(5)*, it included women with obvious advanced breast cancer in a trial of screening, it had them examined clinically before random allocation so that the allocation—on open lists—could be subverted on the basis of the clinical findings *(6)*, and it had poor-quality mammography *(7)*. The fact that a review found no forensic evidence of subversion is meaningless, since the reviewers never interviewed those responsible for the allocations despite being well aware that such interviews were the only way to confirm subversion.

Dismissing the lack of power in the Canadian trial, Berry suggests that it was the most believable trial because of its size. If this is the case, then the meta-analysis of the combined eight studies that shows a statistically significant benefit should be even more believable. Instead, Berry incongruously argues that statistical significance is meaningless. This argument only diverts the focus from the question of the benefit of mammographic screening to the fundamental controversy among theoretical statisticians regarding conventional,

frequentist principles versus Bayesian analysis. That discussion does not belong here.

Berry's Bayesian conclusion relies on the basic observation that the trials differed in achieved mortality reduction. We fully agree. The Canadian study is a prime example of how poorly performed mammography can be of little value. Instead of suggesting that there is no benefit from screening, Berry should be urging women to avoid the type of poor-quality mammography found in the Canadian trial and to demand high-quality mammography.

His argument that demographic factors in the two groups in the Canadian trial were equivalent has little bearing on the subversion of the randomization process. It would take a shift of only a few dozen women with advanced breast cancer from the screened group to the control group to subvert the trial. Even a shift of several hundred women would not detectably alter the demographics in a trial that involved 50 000 women.

Berry's suggestion that statistical power is unimportant is inconsistent with his stressing the wide confidence intervals in the data. Lack of power due to inadequate sample size is the reason for wide confidence intervals. It will similarly affect a Bayesian analysis. Power is critical when the findings suggest that a study shows no benefit. A "negative" finding should never be taken as "proof" that there is no benefit, yet opponents of screening have made this inference despite the fact that the lack of power was due to the inappropriate use of unplanned subgroup analysis.

The lack of statistical significance in unplanned, underpowered subgroup analyses of the early data has been used, inappropriately, to conclude that the benefit is small and inconsequential. This is not true. An underpowered negative study is only evidence that the study was not big enough. None of the studies by themselves, or even combined, were big enough in the early years of follow-up to be meaningful. Adding the other problems, including noncompliance and contamination, it is remarkable that, with longer follow-up, the trials show a statistically significant benefit for screening women ages 40–49 years.

DANIEL B. KOPANS
ELKAN HALPERN

REFERENCES

(1) Berry DA. Benefits and risks of screening mammography for women in their forties: a statistical appraisal. J Natl Cancer Inst 1998; 90:1431–9.

(2) Andersson I, Janzon L. Reduced breast cancer mortality in women under age 50: updated results from the Malmo Mammographic Screening Program. Monogr Natl Cancer Inst 1997; 22:63–7.

(3) Bjurstam N, Bjorneld L, Duffy S, Smith T, Cahlin E, Erikson O, et al. The Gothenberg Breast Cancer Screening Trial: preliminary results on breast cancer mortality for women ages 39–49. Natl Cancer Inst Monogr 1997;22:53–5.

(4) Kopans DB, Halpern E, Hulka CA. Statistical power in breast cancer screening trials and mortality reduction among women 40–49 years of age with particular emphasis on the National Breast Screening Study of Canada. Cancer 1994;74:1196–203.

(5) Kopans DB, Feig SA. The Canadian National Breast Screening Study: a critical review. AJR Am J Roentgenol 1993;161:755–60.

(6) Tarone RE. The excess of patients with advanced breast cancer in young women screened with mammography in the Canadian National Breast Screening Study. Cancer 1995;75:997–1003.

(7) Yaffe MJ. Correction: Canada Study [letter]. J Natl Cancer Inst 1993;85:94.

NOTES

Affiliation of authors: Department of Radiology, Ambulatory Care Center, Massachusetts General Hospital/Harvard Medical School, Boston.

Correspondence to: Daniel B. Kopans, M.D., Department of Radiology, Ambulatory Care Center, Massachusetts General Hospital/Harvard Medical School, 15 Parkman St., 2nd Floor, Rm. 219, Boston, MA 02114.

RESPONSE

The first three paragraphs of the Kopans and Halpern letter refer to the Journal summary and the Consensus Development Conference Panel rather than directly to my commentary. However, I want to correct two wrong impressions they convey. First, indeed, the panel report did not refer specifically to the data presented at the conference. Dr. Kopans has made this point on many occasions. As I replied to him at the conference, the panel heard all the data and considered them with diligence and utmost care. The standards by which National Institutes of Health Consensus Panels operate preclude reference to specific studies in the Panels' reports. A panel's task is to synthesize the available information and to issue its conclusions.

Regarding the second Kopans and Halpern complaint about the panel, indeed, the Canadian trial has been criticized. And Dr. Kopans is its harshest critic. The panel considered all such criticisms. It did not choose to separate out the Canadian trial in its report, and it did not choose to separate out other trials, all of which have flaws. The point Kopans and Halpern make about women in their forties receiving the same benefit as women more than 50 years of age if the Canadian trial is excluded is misleading. The incidence of disease is substantially lower among younger women. Therefore, even if the risk reduction were the same in the two groups, the absolute benefit would be smaller in younger women.

In their criticism of my commentary, the authors repeatedly raise the issue of the Canadian trial. My comments were not in praise of the trial but in defense of some of the criticism that has been levied against it. Criticism is healthy and even essential in science. But some criticisms of this trial are irrational and are obviously motivated by the critics' dislike of its results. Especially curious is the suggestion that the Canadian trial was underpowered. It was the largest trial of women in their forties, and, therefore, it has the greatest power. Why don't Kopans and Halpern criticize the Gothenburg and Malmö trials for being underpowered?

The authors' logic in their fifth paragraph escapes me. They say that, if the Canadian trial is the most believable, then a meta-analysis of the eight trials should be even more believable, and that I "incongruously" argue that statistical significance is meaningless. The two issues are unrelated. There are two points relevant to these issues that Kopans and Halpern overlooked in my commentary. Perhaps I did not make them sufficiently clear, and I appreciate the opportunity to do so here. First, my comments about statistical significance have nothing to do with whether one takes a Bayesian or a frequentist approach to statistics. Suppose some very large studies showed a statistically significant reduction in mortality of 1% due to screening and the life expectancy of a woman screened over the decade would be increased by 1 hour. I trust that no one would recommend screening under these circumstances. Statistical significance addresses the question of whether there is benefit, but it is nearly irrelevant in a woman's decision problem.

The other point the authors missed is

the assumption underlying my meta-analysis of the eight studies. Previous analyses, using Mantel–Haenszel, assume that mortality reduction is the same in all of the studies, i.e., that the studies are homogeneous in this sense. I dropped this assumption and let the data themselves address whether or not the assumption was appropriate. Allowing for the possibility of heterogeneity gives a rather different conclusion about significance; i.e., the estimated reduction in mortality is the same, but the associated uncertainty is substantially greater. Whether one makes the assumption of homogeneity has nothing to do with whether one takes a Bayesian or a frequentist approach. Addressing heterogeneity is somewhat easier to do and easier to interpret in the Bayesian approach, and the Bayesian approach has the advantage—one I exploited—of providing predictive probabilities for the results of a future trial.

Regarding the role of locally advanced breast cancer in the Canadian trial, I refer Kopans and Halpern to my commentary, and I will not belabor the point here. I have no disagreement with their penultimate paragraph. There is no question that bigger studies make for stronger conclusions. However, I never claimed that power is unimportant. What I said was: "After a study is completed, power calculations are irrelevant. We know the trial sample size and the trial results. Power is calculated assuming potential but fictitious benefits."

The authors' final paragraph contains a statement that belies their bias: "An underpowered 'negative' study is only evidence that the study was not big enough." A negative study has another possible explanation, i.e., that the null hypothesis of no benefit may be true!

In their final paragraph, the authors fail to recognize an important issue addressed in my commentary. After considering statistical issues regarding uncertainty based on the information from the trials, I took the perspective that the true mortality benefit was in fact the 18% reduction observed in the trials. I translated this benefit into absolute terms, my goal being to help women compare the estimated benefits with the risks and come to a conclusion about whether screening is appropriate for them.

DONALD A. BERRY

NOTE

Correspondence to: Donald A. Berry, Ph.D., Institute of Statistics and Decision Sciences, 223 Old Chemistry Bldg., ISDS, Box 90251, Durham, NC 27708-0251 (e-mail: db@isds.duke.edu).

EDITOR'S NOTE

On page viii of the National Institutes of Health Consensus Development Conference Statement: Breast Cancer Screening for Women Ages 40–49, January 21–23, 1997 *(1)*, the following sentence can be found: "Summary data in five of eight RCTs [randomized controlled trials] show a trend toward reduced breast cancer mortality only after a follow-up of 10 or more years, with the decrease estimated at 16 percent (with confidence intervals from 2 percent to 28 percent)." Since the reported confidence interval excludes zero, this finding constitutes ". . . evidence of a statistically significant reduction in breast cancer mortality after a follow-up of 10 years or more. . . ." Thus, we believe the In This Issue summary critiqued by Kopans and Halpern is accurate.

REFERENCE

(1) National Institutes of Health Consensus Development Panel. National Institutes of Health Consensus Development Conference Statement: breast cancer screening for women ages 40–49, January 21–23, 1997. J Natl Cancer Inst Monogr 1977;22: vii–xviii.