

Rejoinder

David A Freedman, Diana B Petitti and James M Robins

We reviewed¹ the critique of mammography^{2–5} and much of the underlying technical literature. We found that the critique depended on misreadings of the data and the literature. The clinical trials of mammography, although no doubt imperfect, show that screening saves lives. Other reviewers concur, including Nyström *et al.*,⁶ Health Council of the Netherlands,⁷ US Preventive Services Task Force,^{8,9} and the International Agency for Research on Cancer (IARC).¹⁰ According to IARC, for example:

The possibility of the introduction of bias into the results of the studies of screening with mammography alone by a range of methodological factors was considered. The available evidence suggested that none, if any, bias was present that could have had a sufficiently large effect to affect the overall rate ratios appreciably. (ref. 10, p. 174)

There is *sufficient evidence* for the efficacy of screening women 50–69 years by mammography as the sole screening modality in reducing mortality from breast cancer. (ref. 10, p. 179)

The commentaries do not lead to any changes in these conclusions.

Gøtzsche¹¹ has raised many arguments; some are old, some are new, but none are convincing. To give the flavour, we discuss several of the points, starting with this one. Gøtzsche maintains that he had no preconceived ideas about screening when he reviewed the mammography literature. This reiterates claims made elsewhere:

We had no opinion on the effect of screening when we started to review the literature.¹¹

We had no *a priori* opinion on the effect of screening for breast cancer when we were asked in 1999 by the Danish Institute for Health Technology Assessment, the National Board of Health, to review the randomised trials. (ref. 3, p. 20)

The claims are not supported by the record. In 1997, he published a letter in the *Lancet* discussing colon cancer screening studies and the ethics of screening:¹²

The studies also raise a pertinent ethical issue: do we wish to turn the world's healthy citizens into fearful patients-to-be who, in the not too distant future, might be asked to deliver, for example, annual samples of faeces, urine, sputum, vaginal smear, and blood, and undergo X ray and ultrasound examination with all it entails in terms of psychological morbidity and the potential for harm because of further testing and interventions due to false positive findings?

He makes much of the apparent increase in the incidence rate of invasive cancer for women age 50 and over in the US, considering this to be proof of over-diagnosis due to screening.

However, there have been substantial changes over time in the risk factors for breast cancer, including age at menarche, age at first pregnancy, and number of pregnancies. For example, the fertility rate dropped from about 3.4 in the early 1960s to 1.8 in the late 1970s, coming back up to 2.4 in the early 1990s.¹³ Until risk factors have been accounted for, changing incidence rates are difficult to interpret, which may explain some differences of opinion in the literature.^{10,14,15} There are other problems with the argument. Breast cancer rates for older women were going up well before screening became widely recommended. Moreover, there is no corresponding increase for women below the age of 50: Figure IV-3 in the source document¹⁶—unlike Gøtzsche's Figure 1—displays both sets of rates for comparison.

With respect to the Health Insurance Plan (HIP) trial, Gøtzsche imputes to us the view that numbers 'are not consistent, not even numbers of randomized women', and suggests that we found a discrepancy of 434. Not at all. The figure of 434 is reported by Gøtzsche and Olsen (GO).² We explained,¹ as did Miller,¹⁷ that numbers in the HIP tables do not refer to the population randomized, but to the population after exclusion criteria were applied. In the screening group, women diagnosed with breast cancer prior to randomization were detected at first screen; however, in the control group and the refused-screening group, such women were detected at death or recurrence, and excluded after detection.^{18,19} That is the source of the 'discrepancy'. We show¹ that the data are internally consistent and close to what is expected on an actuarial basis. We conclude that bias in the HIP estimates is negligible (also see below).

Gøtzsche asserts, incorrectly, that 'many more women with breast cancer prior to randomization were excluded from the study group than control group ...'. There is no response to our discussion. He even withdraws a previous near-retraction,²⁰ claiming instead:

the lead investigator [Sam Shapiro] admitted that even more than 20 years after the study started, some prior breast cancer cases among the controls were unknown to the investigators and should have been excluded.

Shapiro actually said:

Of course, some prior breast cancer cases are still unknown to us among the study group women not screened and the controls that should be excluded. However, these cases are believed to be few in number and are unlikely to affect our comparisons. (ref. 18, p. 75)

In other words, some women in the control group and refused-screening group were still alive and disease-free in 1985. These women do not appear in the counts of deaths or breast cancers. The counts are not biased. However, the women are included

in the tally of person-years at risk, which is therefore slightly inflated, creating a tiny bias *against* mammography.^{1,17}

The original critique²⁻⁴ also claimed imbalance of baseline characteristics, using data from Table 4.1 in Shapiro *et al.*¹⁹ However, GO failed to notice that the data only cover subsamples of entrants during the first year of the HIP trial, and used the wrong sample sizes when calculating *P*-values. If the right sample sizes are used, differences are well within the range of chance variation. Gøtzsche concedes the point but says:

The papers are very confusing We took account of these reduced sample sizes when we calculated our *P*-values which are therefore correct, if the text is to be believed. However, the table header speaks of women entering the study in 1964, and not of all women as the text does.

The HIP text is consistent with the table, and quite clear:

in December 1963 the randomized controlled trial started Information obtained through a survey of subsamples of the total study and control women with entry dates through 1964 indicates that these two groups were highly comparable (table 4.1). (ref. 19, pp. 17, 35).

In the Two-County trial, Gøtzsche claims that bias in classification of deaths 'can be very large', citing differences between data from the Two-County investigators and the Swedish death registry. However, a difference by itself hardly demonstrates bias. The Two-County investigators might be wrong; so might the death registry: and some differences of opinion are only to be expected. Indeed, after review of the medical records, the Two-County investigators believe they are determining cause of death more accurately than the death registry (Stephen Duffy, personal communication; Duffy also finds little statistical evidence for systematic differences between Two-County investigators and the death registry on breast cancer mortality rates in the active study population [ASP] or the passive study population [PSP], OR = 1.18, 95% CI: 0.88, 1.58). At a more basic level, the data cited by Gøtzsche show that screening has a significant impact on the death rate from breast cancer, under a wide variety of rules for counting deaths.

In their original critique, GO focused on 'discrepancies' in reporting, especially of breast cancer deaths.²⁻⁴ We resolved all the discrepancies that we checked. Gøtzsche now protests that there are other such discrepancies. The chief one seems to be in lines 1 and 2 of his Table 5, with 26:18 deaths in the two arms of the trial at 10.4 years of follow-up and 22:16 at 12.5 years. Line 1 is from Nyström and Larsson;²¹ line 2, from Tabár *et al.*²² But there are some crucial differences of detail in the way these two groups tabulate the data: (1) they define age differently (one group uses exact age, the other uses year of birth); (2) each group has its own 'endpoint committee' for determining cause of death; (3) Nyström and Larsson count deaths according to the 'followup model', while Tabár *et al.* use the 'evaluation model', which gives smaller numbers. These differences are explained by Nyström and Larsson, in the paper Gøtzsche cites.

The original critique also placed great emphasis on total mortality as the only valid endpoint for screening trials.²⁻⁴

Nyström *et al.*⁶ demonstrated a 2% reduction in total mortality (RR = 0.98; 95% CI: 0.96, 1.00). Gøtzsche responds that 'Nyström and colleagues did not claim that the 2% reduction in total mortality they found was statistically significant.' Under the circumstances, this is a distinction without a difference. For example, (95% CI: 0.959, 0.999) conveys essentially the same information as (95% CI: 0.961, 1.001). The former is 'statistically significant', the latter is not.

We turn to the Canadian National Breast Screening Study (CNBSS).^{23,24} In CNBSS1 (age 40-49), there were 22 advanced cancers (4+ nodes involved) detected by physical examination at first screen: 17 in the treatment arm and 5 in the control arm. This difference is statistical evidence that high risk women were 'steered' to treatment during assignment.^{25,26} Gøtzsche¹¹ (and Miller²⁷) reply that the excess is a post-randomization bias: diagnostic workup was less thorough for women in control, because there was no radiography.

This argument is implausible, as we show now. In paraphrase, if during the physical examination of the breasts an abnormality was detected, the participant was referred to the National Breast Screening Study review clinic, where it was the study surgeon's role to examine the participant and to decide whether further diagnostic procedures were indicated: in the control group, these procedures could include diagnostic mammography, aspiration, or biopsy.²³

Thus, Gøtzsche is suggesting that the review process missed something like $(17-5)/17 = 70\%$ of the advanced cancers (4+ nodes involved). But these tumours generally exceeded 1 cm in size. Indeed, a majority exceeded 2 cm (Table 1). These are not small tumours. They would not be easily missed on referral. Moreover, nothing in Gøtzsche's argument is specific to cancers with 4+ nodes involved. Were the argument correct, we would also expect a large excess in other invasive cancers detected by physical examination at first screen. However, 48 of these were in the treatment group and 55 in control. (ref. 24, p. 1470)

Gøtzsche is right to note an error in our calculation of the reduction in death rates in the Two-County study. The proportionate reduction in mortality due to the invitation to

Table 1 The Canadian National Breast Screening Study, women age 40-49. Number of advanced cancers (4+ nodes involved) detected by physical examination at first screen, by size (mm), for each arm of the trial

	Treatment	Control
<9	0	0
10-14	1	0
15-19	5	0
20-39	8	4
≥40	1	1
Unknown	1	0
Total	16	5

The 'Treatment' column is computed by us from Appendix Table 3 in Miller *et al.*²⁴ subtracting 'Mammography Alone' from 'All'. There are discrepancies between Appendix Table 3 in Miller *et al.*²⁴ and Table 7 in Miller *et al.*²³: one cancer has been reclassified from 4+ nodes to 1-3 nodes, and there is an additional cancer with unknown nodal status. (ref. 24, p. E-307).

screening is $(6.5-3.9)/6.5 = 40\%$ in Kopparberg, and $(5.7-4.3)/5.7 = 25\%$ in Östergötland.

We turn to the other discussants. Berry²⁸ says that:

arguments over the credibility of the randomized screening mammography trials are red herrings. The overall results of the trials suggest that screening very likely reduces breast cancer mortality.

He claims this has always been his position, and asks why he has 'been painted as being anti-screening'. This question is answered by his testimony before the US Senate (28 February 2002).

At the January 2002 PDQ [Physicians' Data Query] Board meeting we considered an article authored by Drs Ole Olesen and Peter Gotzsche of the Nordic Cochrane Collaborative and that appeared in *The Lancet* in October 2001. This article critiqued the randomized trials that have been conducted to evaluate the benefits of screening mammography and cited a number of deficiencies and flaws. Many of these were known previously and there was little original information in the review. However, it served to put the trials' deficiencies into perspective and led us to re-evaluate the credibility of the trials. We decided to revise our breast cancer screening statement and to refer to the Olsen-Gotzsche article The current version of the statement indicates that the benefits of screening are uncertain. Therefore, in a sense the revision will be minor. However, we plan to indicate that the existence of benefit itself is uncertain.

Berry has been far too critical of trials that demonstrate the efficacy of mammography, including HIP and Two-County.²⁸⁻³⁰ A change in perspective would be welcome.

Baum's main point³¹ is that the three trials we discuss—HIP, Two-County, and CNBSS—do not meet the CONSORT standard. (For reasons that remain unclear, he thinks CNBSS comes closer than the other two.) Most of what is known in medicine derives from studies that would fail Baum's standard. In an imperfect world, the relevant question seems to be this: what do the existing data say about mammography? Baum rejects the question, but we think that the data make a strong case for the efficacy of screening.

His other points are equally unpersuasive. For example, contrary to Baum's view, the HIP and Two-County data have generally been analysed by the intention-to-treat principle. In Two-County, for instance, the ASP comprises all women who were invited to be screened, whether or not they accept screening. The PSP comprises the women who were not invited to screening. The ASP is compared to the PSP. This is intention to-treat.

For another example, we say that lead time bias is due to the fact that screening speeds up detection. He calls this a misunderstanding: 'lead time bias ... simply means extending the period of observation so that a woman with a screen detected cancer may enjoy a longer period of post diagnosis survival ...' But the period is longer only because the cancer is detected earlier. Why is this a point of contention?

Miller²⁷ generally agrees with us on the HIP trial. In the Kopparberg segment of the Two-County trial, he does 'not have absolute certainty that the clusters in Kopparberg were balanced'. Absolute certainty is rarely to be found, but considerable reassurance is provided by Nyström *et al.*³² and Duffy.³³ On CNBSS, the publications by Miller and his co-workers fail to resolve our doubts. In the end, we cannot agree that CNBSS should be considered superior in quality to HIP or Two-County.

Baum and Miller correctly point to difficulties in extrapolating from the study population to any larger population of interest, especially when risk factors and treatment options are changing. This is an issue worth considering in most clinical trials. The issue will not be resolved even by the strictest application of Baum's principles, because resolution depends on judgement about the similarity between the population of interest and the population randomized, not on details of trial methodology or reporting. This judgement is easier for Two-County than for many other trials, because the study population is a natural one—women aged 40–74 at baseline, living in two Swedish counties.

Baum and Miller ask about the clinical utility of mammographic screening in today's circumstances. At present, many critics of screening seem to be attacking efficacy instead of discussing harder questions of clinical relevance. Accepting the evidence that demonstrates internal validity of HIP and Two-County—and in consequence the efficacy of screening—would make it more feasible to have a productive discussion of utility.

References

- 1 Freedman DA, Petitti DB, Robins JM. On the efficacy of screening for breast cancer. *Int J Epidemiol* 2004;**33**:43–55.
- 2 Gøtzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000;**355**:129–34. Discussion, *Lancet* 2000;**355**:747–52.
- 3 Olsen O, Gøtzsche PC. *Screening for Breast Cancer with Mammography (Cochrane Review)*. Oxford: Update Software. The Cochrane Library, Issue 4, 2001.
- 4 Olsen O, Gøtzsche PC. *Systematic Screening for Breast Cancer with Mammography*. 2001. <http://image.thelancet.com/lancet/extra/fullreport.pdf>
- 5 Olsen O, Gøtzsche PC. Cochrane review on screening for breast cancer with mammography. *Lancet* 2001;**358**:1340–42. Discussion, *Lancet* 2001;**358**:2164–68.
- 6 Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist LE. Long-term effects of mammography screening: Updated overview of the Swedish randomised trials *Lancet* 2002;**359**:909–19. Discussion, 2002;**360**:337–40.
- 7 Health Council of the Netherlands. *The Benefit of Population Screening for Breast Cancer with Mammography*. The Hague, 2002.
- 8 US Preventive Services Task Force. Screening for breast cancer: Recommendations and rationale. *Ann Intern Med* 2002;**137**:344–46.
- 9 US Preventive Services Task Force. Breast cancer screening: A summary of the evidence. *Ann Intern Med* 2002;**137**:347–67.
- 10 International Agency for Research on Cancer. *Breast Cancer Screening*. Lyon: IARC. IARC Handbooks of Cancer Prevention, Vol. 7, 2002.
- 11 Gøtzsche P. On the benefits and harms of screening for breast cancer. *Int J Epidemiol* 2004;**33**:56–64.

- ¹² Gøtzsche P. Screening for colorectal cancer. Letter. *Lancet* 1997; **349**:356.
- ¹³ US Census Bureau. *Statistical Abstract of the United States*. Washington, DC, 2000.
- ¹⁴ Tarone RE, Chu KC. Implications of birth cohort patterns in interpreting trends in breast cancer rates. *J Natl Cancer Inst* 1992; **84**:1402–10.
- ¹⁵ Liff JM, Sung JF, Chow WH, Greenberg RS, Flanders WD. Does increased detection account for the rising incidence of breast cancer? *Am J Public Health* 1991; **81**:462–65.
- ¹⁶ Ries LAG, Eisner MP, Kosary CL *et al.* (eds). *SEER Cancer Statistics Review, 1973–1999*. Bethesda, MD.: National Cancer Institute, 2002. http://seer.cancer.gov/csr/1973_1999/.
- ¹⁷ Miller AB. Screening for breast cancer with mammography. Letter. *Lancet* 2001; **358**:2164.
- ¹⁸ Shapiro S, Venet W, Strax P, Venet L, Roeser R. Selection, follow-up, and analysis in the Health Insurance Plan study: A randomized trial with breast cancer screening. In: Garfinkel L, Ochs O, Mushinski M (eds). *Selection, Follow-up, and Analysis in Prospective Studies: a Workshop*. National Cancer Institute Monograph 67. Bethesda, MD: National Cancer Institute, May 1985, pp. 65–74. Discussion, pp. 75–79. (NIH publication no. 85–2713.)
- ¹⁹ Shapiro S, Venet W, Strax P, Venet L. *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986*. Baltimore: Johns Hopkins, 1988.
- ²⁰ Gøtzsche PC. Screening for breast cancer with mammography. Author's reply. *Lancet* 2001; **358**:2167–68.
- ²¹ Nyström L, Larsson LG. Breast cancer screening with mammography. Letter. *Lancet* 1993; **341**:1531–32.
- ²² Tabár L, Fagerberg G, Chen HH *et al.* Efficacy of breast cancer screening by age: New results from the Swedish two-county trial. *Cancer* 1995; **75**:2507–17.
- ²³ Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Can Med Assoc J* 1992; **147**:1459–76.
- ²⁴ 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**:305–12.
- ²⁵ Mettlin CJ, Smart CR. The Canadian National Breast Screening Study: An appraisal and implications for early detection policy. *Cancer* 1993; **72**(Suppl.):1461–65.
- ²⁶ Tarone RE. The excess of patients with advanced breast cancers in young women screened with mammography in the Canadian National Breast Screening Study. *Cancer* 1995; **75**: 997–1003.
- ²⁷ Miller AB. Commentary: A defence of the Health Insurance Plan (HIP) study and the Canadian National Breast Screening Study (CNBSS). *Int J Epidemiol* 2004; **33**:64–65.
- ²⁸ Berry DA. Screening mammography: a decision analysis. *Int J Epidemiol* 2004; **33**:68.
- ²⁹ Berry DA. Benefits and risks of screening mammography for women in their forties: A statistical appraisal. *J Natl Cancer Inst* 1998; **90**:1431–39.
- ³⁰ Kolata G. Different conclusions from the same study. *New York Times*, 9 April 2002, p. D4.
- ³¹ Baum M. Commentary: False premises, false promises and false positives—the case against mammographic screening for breast cancer. *Int J Epidemiol* 2004; **33**:66–67.
- ³² Nyström L, Larsson LG, Wall S *et al.* An overview of the Swedish randomized mammography trials: Total mortality pattern and the representivity of the study cohorts. *J Med Screening* 1996; **3**: 85–87.
- ³³ Duffy SW. Interpretation of the breast screening trials: A commentary on the recent paper by Gøtzsche and Olsen. *The Breast* 2001; **10**:209–12.
- ³⁴ Hearst N, Newman TB, Hulley SB. Delayed effects of the military draft on mortality: a randomized natural experiment. *New Engl J Med* 1986; **314**:620–24.
- ³⁵ Sommer A, Zeger SL. On estimating efficacy from clinical trials. *Stat Med* 1991; **10**:45–52.
- ³⁶ Angrist J, Imbens G. Identification and estimation of local average treatment effects. *Econometrica* 1994; **62**:467–75.
- ³⁷ Chu KC, Smart CR, Tarone RE. Analysis of breast cancer mortality and stage distribution by age for the Health Insurance Plan clinical trial. *J Natl Cancer Inst* 1988; **80**:1125–32.

Appendix: Two statistical issues

This Appendix covers two specialized topics. The first is correcting for dilution of intention to-treat estimates by cross-over. Our Table 1 (part of which is reproduced here for ease of reference as Appendix Table 1) showed 39 deaths from breast cancer in the study group, compared with 63 in control.¹ The ratio is $39/63 = 0.62$, which measures the effect of the invitation to screening on the study group (intention-to-treat): death rates from breast cancer are multiplied by the factor 0.62. There is dilution, however, because only some of the women in the study group were screened. Correcting for dilution, we found the effect of screening—on those women who were actually screened—was to multiply death rates from breast cancer by the factor $(39-16)/(63-16) = 0.49$. Here, 16 is the number of deaths from breast cancer in the refused-screening group.

Gøtzsche¹¹ appears not to understand the statistical logic, contending that subtraction 'leads to biased estimates'. Our method for estimating the effect of treatment on the treated is well-known, and it is unbiased.^{34–36} Of course, sampling error should be considered; we return to that in a moment. However, bias is the issue raised by Gøtzsche, and to deal with bias, assume for the moment that the data in Appendix Table 1 are free of sampling error.

The control group of 31 000 women comprises two groups that cannot be directly observed: (1) 20 200 women who would have accepted screening had they been invited ('compliers'), and (2) 10 800 who would have refused. Indeed, 20 200/31 000 and 10 800/31 000 are the ratios observed in the study group.

The control group experienced 63 deaths from breast cancer. How many of these are contributed by group (2), the refusers? The answer is, 16. From Appendix Table 1, the death rate from breast cancer among the refusers in the treatment group is 16/10 800. This rate applies equally to the refusers in control: due to randomization, the refusers in treatment and control are balanced in risk, and sampling error is taken to be negligible. We infer that 63–16 breast cancer deaths were contributed by compliers in the control group. We see directly from Table 1 that $39-16 = 23$ breast cancer deaths were contributed by the compliers in the study group.

That is why the effect of screening on the compliers is to multiply their death rates from breast cancer by the factor $(39-16)/(63-16) = 0.49$, as we said initially. Subtracting 16 from 63 and 39, contrary to Gøtzsche's claim, creates no bias at all. Of course, subtracting other numbers, as in his illustration,¹¹ may well create bias. Finally, the effect

Appendix Table 1 Health Insurance Plan data. Group sizes (rounded), and deaths from breast cancer in 5 years of follow-up

	Group size	No. of deaths from breast cancer
Study		
Screened	20 200	23
Refused	10 800	16
Total	31 000	39
Control	31 000	63

of sampling error on the estimate can be determined, for instance, by the delta-method: the standard error is about 0.15.

Our next topic is the effect of length bias and lead time bias on statistical tests. Table 6 in our paper¹ reports the percentage of breast cancer cases in each arm of the HIP trial that have died. The time period (diagnosis within 7 years of entry) is chosen so that incidence has equalized.³⁷ Gøtzsche¹¹ says that we 'compare total mortality *among breast cancer cases* in the study group and the control group and find a significant difference'; and we consider the 'comparison fair since numbers of incident cases in the two arms have equalized'. But, he continues, 'this is a gross error that, regrettably, is very common in the screening literature'. Two reasons are given: (1) the incidence rates should not have equalized, and (2) the comparison is vitiated by length bias and lead time bias. We accept neither point. The first is circular: he thinks screening leads to over-diagnosis and does not save lives, so trials that fail to show over-diagnosis (or succeed in demonstrating a positive effect?) are flawed. With respect to (2), the object of waiting for equalization is to avoid bias.³⁷

To give some detail, the endpoint under consideration is death after a diagnosis of breast cancer. Suppose we use numbers of women randomized as the denominators rather than numbers of incident cases.

Then $z = 2.24$, $P = 0.025$. In this analysis, the bias is against screening. Two steps are needed for the demonstration.

1. All cancers in the treatment arm—whether screen-detected or not—are counted against treatment. Therefore, unless there is some breakdown in management of the trial, there will be at least as many cancers in the treatment arm as the control arm, up to random error. In other words, there are at least as many women on risk in the treatment arm as the control arm. (There was equalization for HIP since there were only 1–4 rounds of screening in the treatment arm.)
2. Because their cancers tend to be detected earlier, women in the treatment group are on risk longer than the controls, and a few more of them die for that reason alone.

In fact, as the reasoning shows, the analysis would be conservative even for a trial with over-diagnosis in the treatment arm.

We made a χ^2 test in our paper:¹ $\chi^2 = 5.8$, $p = 0.02$. This test takes the incidence numbers as given, and uses them (in effect) as denominators, picking up a little extra power. The gain is small because the sample is large and the incidence numbers are nearly equal. Conversely, the bias in favour of screening (if there is such a bias) is tiny.