

## POINT–COUNTERPOINT

## On the efficacy of screening for breast cancer

David A Freedman,<sup>1</sup> Diana B Petitti,<sup>2</sup> and James M Robins<sup>3</sup>

Accepted 24 June 2003

**Background** ‘Mammography’ (screening for breast cancer by X-ray examination) came to be widely—although not universally—accepted in the 1980s when a number of clinical trials demonstrated a substantial reduction in risk. Early detection, before the disease spread, permitted therapy that was simultaneously less invasive and more effective. Questions that remained were largely about efficacy for younger women and optimal frequency for older women. The consensus was challenged in a series of papers by two researchers at the Nordic branch of the Cochrane collaboration, Gøtzsche and Olsen, who concluded that mammography does not save lives: instead, it exposes women to unnecessary surgical procedures.

**Methods** Qualitative review.

**Results** The basis for the Gøtzsche–Olsen critique turns out to be simple. Studies that found a benefit from mammography were discounted as being of poor quality; remaining negative studies were combined by meta-analysis. The critique therefore rests on judgements of study quality, but these judgements are based on misreadings of the data and the literature.

**Conclusion** The prior consensus on mammography was correct.

**Keywords** Mammography, breast cancer, screening, meta-analysis

The first large-scale clinical trial to demonstrate the efficacy of mammography was Health Insurance Plan (HIP) in New York,<sup>1–7</sup> followed by the Two-County study in Sweden.<sup>8–22</sup> There were about half a dozen other trials as well, some negative but most positive. In theory, if breast cancer begins as a local disease, then early diagnosis—before the disease spreads—should allow treatment that is less invasive and more effective;<sup>23</sup> there may also be a biological rationale from the perspective of systemic disease.<sup>24</sup> After an initial period of controversy, mammography gained general acceptance.<sup>23–29</sup> Some doubts remained, especially for younger women.<sup>30–32</sup> There were also questions about optimal schedules for screening and cost effectiveness,<sup>33,34</sup> but our focus is efficacy.

The consensus opinion was challenged by two researchers at the Nordic branch of the Cochrane collaboration, Gøtzsche and Olsen, who concluded that mammography does not save lives: instead, it exposes women to unnecessary diagnostic and

surgical procedures.<sup>35–38</sup> This opinion was based on a meta-analysis of the existing trials, where positive studies were eliminated as being of poor quality; the remaining two studies found negative effects. Thus, the critique hinges on the decision to exclude positive studies like HIP and Two-County. That decision was justified in turn by a literature review.

In this paper, we discuss HIP, Two-County, and the best-known of the negative studies, the Canadian National Breast Screening Study (CNBSS).<sup>39–45</sup> We summarize the trials and the critique. We briefly discuss other work faulting the positive studies.<sup>46,47</sup> Our paper addresses the major points raised by the critics, and a few of the minor ones. We find that the quality judgements behind the critique—and hence the meta-analytic results—are based on misunderstandings of the data. We see no reason to believe that CNBSS was superior in quality to HIP or Two-County. In our opinion, therefore, the critique has little merit. We conclude that the prior consensus on mammography was correct: screening does save lives. Others have reached similar conclusions on the central points,<sup>26–29,48–53</sup> although the critique has attracted some support.<sup>46,47,54–58</sup> The Swedish trials (including Göteborg, Malmö, and Stockholm) have been reviewed by Nyström *et al.*,<sup>53,59–61</sup> with commentary<sup>35–38,62–66</sup> and responses.<sup>53,67</sup> The Edinburgh trial and the Finnish National

<sup>1</sup> Department of Statistics, University of California, Berkeley, CA 94720–3860, USA.

<sup>2</sup> Kaiser Permanente, Southern California, Pasadena CA 91188, USA.

<sup>3</sup> Harvard School of Public Health, Boston MA 02115, USA.

Screening Program have been reviewed elsewhere, with comparison to the Swedish trials.<sup>27–29</sup>

## The Health Insurance Plan trial

HIP is the Health Insurance Plan of Greater New York, a group medical practice which had, in the 1960s, some 700 000 members in 31 medical groups. Medical records consisted largely of paper, with computerized summaries. Subjects in the experiment were 62 000 women age 40–64, members of HIP, who were assigned to treatment or control by systematic list sampling. ‘Treatment’ consisted of invitation to four rounds of annual screening—a clinical exam and mammography (two views, cephalocaudal and lateral). The control group continued to receive usual health care. Subjects were recruited during the period 1963–1966. There were 18 years of follow-up.

The analysis was by intention to treat rather than treatment received. This is conservative, and measures the effect of the invitation to screening rather than the effect of screening itself. (Biases in treatment-received analyses are discussed by Shapiro *et al.*<sup>6</sup>) The effect of screening is diluted because there were only four rounds of screening, and some women in the treatment group declined to be screened: 67% were screened at least once, 40% were screened four times (Table 3.1 in Shapiro *et al.*<sup>6</sup>). Because of this crossover, Shapiro *et al.*<sup>6</sup> refer to the treatment group as the ‘study group’.

Results from the first 5 years of follow-up are shown in Table 1 below.<sup>4,6,68</sup> The effect of the invitation is small in absolute terms:  $63 - 39 = 24$  lives saved. Since the absolute risk from breast cancer is small, no intervention can have a large effect in absolute terms. On the other hand, in relative terms, the 5-year death rates from breast cancer are in the ratio  $39/63 = 62\%$ . The absolute differential persists throughout the 18-year follow-up period, and is perhaps more marked if we take cases incident during the first 7 years of follow-up, rather than 5.

The effect of screening on the screened can be estimated<sup>69</sup> when—as in the HIP trial—there is crossover only from the treatment arm to the control arm: some women invited to screening refuse, while control women do not seek out screening. In Table 1, there were 23 deaths from breast cancer among women who were screened, 16 among women who were offered screening but declined, and 63 in the control group. We estimate that the control group includes 16 women who would have refused screening and who died of breast cancer. The effect of screening

on the women who accepted it was therefore to cut the 5-year death rate in half:

$$\frac{23}{63 - 16} = 0.49.$$

Numerator and denominator are unbiased estimates, but there is considerable statistical uncertainty due to the limited sample size. There is also a tacit assumption that the invitation to screening has no effect unless the woman takes it up. Similar estimates have been reported, based on data from the Two-County trial and population screening in Sweden,<sup>19,21</sup> with discussion.<sup>20,23,70–74</sup> Stronger assumptions are needed to analyse these data, which are partly experimental and partly observational; somewhat lower estimates have recently been suggested.<sup>22</sup>

We turn now to the critique. Gøtzsche and Olsen<sup>35–38</sup> have three main arguments against HIP which we discuss in turn.

- (1) Women with breast cancer diagnosed before randomization were differentially excluded from the screening group.
- (2) There was an imbalance in baseline characteristics.
- (3) There was differential bias in death certification.

### Differential exclusion of breast cancer cases

Women were assigned to study or control in alternation, so the two groups should be equal in size. However, the study group is a bit smaller and the differential changes a little from one report to another.<sup>35</sup> On p. 18 of Shapiro *et al.*,<sup>6</sup> the difference is:

$$\text{Study} - \text{Control} = 30\,131 - 30\,565 = -434.$$

According to Olsen and Gøtzsche:

This (differential) would be expected to create bias. If only 10% of these excluded breast cancer cases are added as breast cancer deaths after 18 years of follow-up, the breast cancer mortality becomes higher in the screened group than in the control group, since the difference in breast cancer mortality at that time was 44 deaths. (ref. 37, p. 6)

In essence, women with a prior diagnosis of breast cancer are at higher risk of death from that disease: differentially excluding them from the screening group therefore creates a bias favouring mammography. To assess this criticism, we consider the exclusion criteria in HIP, which can be summarized as follows: (i) change in medical coverage between randomization and first screen, or (ii) pregnancy at screening, or (iii) diagnosis of breast cancer prior to entry.

The first criterion excluded few women; bias from this source must be small. Furthermore, reasons for changing medical coverage (moving, changing jobs) seem at best weakly related to breast cancer risk, so the sign of the bias is uncertain. With respect to the second, pregnant women are also few in number (we estimate 100 in each arm, as does Raymond Fink, personal communication). Such women must generally have been in their early 40s, and therefore at lower-than-average risk of breast cancer. Perhaps two-thirds of them were excluded from the study group at screening, which creates a (small) bias against mammography. These two criteria operate only on the screening

**Table 1** Health Insurance Plan data. Group sizes (rounded), deaths in 5 years of follow-up, and death rates per 1000 women randomized

	Group size	Breast cancer		All other	
		No.	Rate	No.	Rate
<b>Study</b>					
Screened	20 200	23	1.1	428	21
Refused	10 800	16	1.5	409	38
Total	31 000	39	1.3	837	27
<b>Control</b>					
	31 000	63	2.0	879	28

Data from p. 20 in Freedman *et al.*<sup>68</sup> also see Tables 4.3, 6.2, and 6.3 in Shapiro *et al.*<sup>6</sup> The numbers used in Freedman *et al.* were provided by Sam Shapiro (personal communication, around 1988). Counts differ slightly from those in earlier publications, like Table 2 in Shapiro,<sup>4</sup> presumably due to data editing. Screened means, accepted  $\geq 1$  screens.

group, accounting for some of the difference between the study and control groups.

In both arms of the trial, women with a diagnosis of breast cancer prior to randomization were excluded from counts of breast cancer cases or deaths. In the screening group, exclusions were mainly done at first screen; date of diagnosis was determined from medical records. For women in the study group who refused screening, and for women in the control group, exclusions were made when there was a recurrence of breast cancer or death; again, date of diagnosis was determined from medical records. (With paper records, exclusion prior to randomization would have been expensive because it would have been necessary to go through all 62 000 medical files.) This is not the most elegant of designs but it does not introduce bias in the counts—if follow-up is good and exclusions are done correctly. There is a small upward bias in determining person years at risk in the control arm and refused-screening group, which on balance works against mammography.

The design implies unequal group sizes after exclusions unless follow-up continues until the last breast cancer case has died. By way of illustration, suppose there were 1000 women with prior diagnosis of breast cancer in each group at baseline. If 80% of those in the study group accept screening and there is no detection in the control group, the initial imbalance would be 800. Over the next 18 years, perhaps half of the  $1000 - 800 = 200$  cases in the refused-screening group and a similar percentage of the 1000 cases in the control group would have a recurrence of cancer or die, and then be excluded. On this basis, the difference between the study and control groups would drop to something like 400. Although this result depends on parameters subject to considerable uncertainty, the calculation shows that the difference of 434 (cited above) between the groups at year 18 of follow-up is not evidence of bias.<sup>75</sup>

Screening does not prevent breast cancer but only speeds up detection. The study group should start with a higher incidence of breast cancer ('lead time bias'), but the control group will catch up a year or so after screening stops. That is what happened (Table 2). Screening was finished (or nearly so) in 4 years, and lead time is on the order of 1 year: in other words, screening picks up a cancer roughly a year before it would become clinically manifest (pp. 43–44, 105–106 in Shapiro *et al.*<sup>6</sup>). Thus, the incidence of breast cancers in the two arms should equalize between years 5 and 7, and it does.<sup>76</sup> The differences in the Table, 304 – 295, 426 – 439, and 767 – 740, are well within the range of chance variation. If high-risk women were differentially excluded from the study group, the count of incident cases in the study group would be lower than the control count, which is not the case.

The number of differential exclusions changes from one HIP paper to another, because there are women in the control group

**Table 2** Health Insurance Plan data. Incident cases during 5, 7, and 12 years of follow-up

	Study	Control
5 years	304	295
7 years	426	439
12 years	767	740

Data from Table 5.1 in Shapiro *et al.*,<sup>6</sup> which has results for years 1–12. With histological confirmation.

and the refused-screening group who had a diagnosis of breast cancer prior to randomization. As follow-up goes on, some of these women suffer a recurrence or die, and are then excluded from the counts (see above). Thus, differentials depend on length of follow-up period and, perhaps, on data editing. GO seem to be aware of these facts.<sup>35</sup> [We abbreviate 'Gøtzsche and Olsen' to 'GO'; the quote is from p. 131; our comments are in square brackets.]

Deaths from breast cancer diagnosed before entry to the trial were generally excluded from analysis. [Done in HIP and Two-County.] Such exclusions can lead to bias when the first round of screening identifies cancer in women who have already noted a tumour in their breast if these women are subsequently excluded. [But exclusions were not made that way.] The New York trial excluded more cancers in the screening group than in the control group. [This is true at first screen but false when tables were compiled, because exclusions depend only on events prior to first screen, as discussed above—therefore irrelevant.]

Design issues have been explored<sup>75</sup> and Gøtzsche has responded: 'We furthermore doubt that retrospective exclusion of women after 18 years of follow-up, as in the New York study, is reliable.' (ref. 77, p. 2168)

Here, Gøtzsche implicitly withdraws previous arguments<sup>35–38</sup> about differential exclusions, claiming instead that procedures were retrospective and therefore unreliable. However, the exclusions were done as the study progressed. They were not based on participants' memories but on diagnosis of breast cancer prior to baseline, as documented in the medical records. HIP surveillance of vital records, hospitals, and health insurance reports was designed to pick up all incident breast cancer cases, recurrences, and deaths, in the study group and in the control group. The evidence in Shapiro *et al.* (ref. 6, pp. 3–4, 17–18, 22–24) shows that exclusions were made in a balanced, comprehensive manner. Ascertainment was nearly perfect through year 10, (ref. 6, p. 24) and not much worse in years 11–18. Although bias due to differential exclusion of breast cancer cases is possible, significant bias seems unlikely; Table 2 supports the data in Shapiro *et al.*<sup>6</sup> On the other side, no tangible evidence has been produced to support claims of differential exclusions.

Of course, another interpretation may be offered for Table 2. Some large number of invasive breast cancers were excluded from the study group, balanced by the inclusion of a very similar number of DCIS (ductal carcinoma *in situ*) cases detected by screening; the corresponding cases on the control side remaining occult. This scenario seems far-fetched, for two reasons (apart from the nicety of the requisite balancing): (i) in the 1960s, DCIS might have accounted for 5% of breast cancers in the study group (compare Table 4 in Tabár *et al.*<sup>15</sup> with Table 1 in Rosner *et al.*<sup>78</sup>) and (ii) roughly half of untreated DCIS cases become clinically manifest. (Health Council of The Netherlands, ref. 48, pp. 45–47) The rate of screen-detected DCIS increased rapidly in the 1980s,<sup>79,80</sup> but this is some 20 years after HIP.

### Imbalance of baseline characteristics

The next point in the critique is the alleged imbalance of baseline characteristics between the study and control groups in HIP. The discussion is based on Table 4.1 in Shapiro *et al.*<sup>6</sup> Part of the

table—with header and footnotes—is shown here as Table 3. ‘College’ means the respondent had some college education; ‘menopause’ means the respondent had or was now having menopause; ‘lump’ means the respondent had a lump in the breast at some point. According to GO,

... in the table of seven selected characteristics ... we calculated imbalances for previous lump in the breast ( $p < 0.0001$ ), menopause ( $p < 0.0001$ ), and education ( $p = 0.05$ ); there were no differences for age, religion, marital status, or pregnancies. These findings are incompatible with an adequate randomisation. (ref. 35, p. 130)

We found no details of the calculation in any published paper, and believe GO misunderstood the sample sizes in the table header and footnotes. The table reports not on the whole cohort, but only on a sample of those recruited during 1964. (Samples were taken to reduce costs.) With allowance for non-response, the sample sizes for the examined, refused, and control groups are about 700, 600, and 1800 respectively. (Raymond Fink has contemporaneous documentation for planned sample sizes, personal communication.) On this basis, differences in education and menopause are insignificant. For lumps—the worst of the comparisons—we compute  $z = 2$ ,  $P = 0.05$  (two-sided), using the correct sample sizes. Table 2 in Shapiro *et al.*<sup>2</sup> has more complete data for the examined group; the percentage with lumps may be computed from those data as 11.7%, which is much closer to the control figure of 11.8% in Table 3, confirming that the discrepancy noted by GO is just due to chance; also see Table 8 in Fink *et al.*<sup>3</sup> We conclude there was no imbalance at baseline in HIP.

**Bias in death certification**

We turn now to the third element of the critique, namely, differential bias in death certification; the idea here is that breast cancer is less likely to be recognized as the cause of death in the screening group, due perhaps to a belief in the efficacy of screening. GO make the following claims.

- (1) ‘Knowledge of screening status may affect the judgment of cause of death. Masked assessment of cause of death was used only in the trials from Canada and Malmö ...’ (ref. 35, p. 131)

**Table 3** Characteristics of women entering Health Insurance Plan project during 1964 (per cent)

Characteristic <sup>a</sup>	Study group			Control <sup>c</sup>
	Total <sup>b</sup>	Examined	Refused	
Age 40–44	24.2	25.3	22.3	24.5
College	30.9	33.7	25.8	32.9
Menopause	70.9	66.6	78.8	74.1
Lump	9.5	10.9	7.0	11.8

<sup>a</sup> Not stated categories (non-response) range from <1% to 4% of total and are distributed in the same manner as knowns.

<sup>b</sup> Data for age are based on totals. For all other characteristics, data are based on a 10% sample of the examined group and a 20% sample of the non-examined group.

<sup>c</sup> Based on a 20% sample of the control group.

- (2) ‘The review provided evidence that assessment of cause of death is unreliable and biased in favour of screening. Even when endpoint committees masked to group assignment were used, uncertain causes of death were significantly more commonly ascribed to breast cancer than to other causes in the control group.’ (ref. 36, p. 1340)
- (3) ‘Cause of death assessments [for 71 women from the HIP study group and 73 from the control group] were considered dubious and were reviewed blindly ... This review appears to be biased with two to seven times ... more deaths classified as caused by breast cancer in the control group than in the screened group ... ( $p = 0.0003$ ).’ (ref. 37, p. 6)

Determining cause of death can be difficult, and bias is possible even in a blind review. (For instance, women in the screening group will often have been treated earlier in the course of the disease, which may influence judgements as to whether cancer at another site is primary or secondary.) In short, the accusation is plausible. However, the evidence presented by GO is flimsy, and there is good evidence against their position. If bias in classification of deaths exists for HIP or the Two-County trial, it is not large.

With respect to (1), HIP used blind review on all breast cancer cases assigned to another cause of death on the death certificate (ref. 6, pp. 29–33). Moreover, there was extensive blind review on the Two-County data.<sup>59,60</sup> The second sentence in (1) is simply wrong. With respect to (2) and (3), GO’s statistics ignore the possibility that ambiguous deaths in the screening group are indeed less likely to be breast cancers, because screening and early therapy help to prevent death from that disease.<sup>50,75</sup> By our calculations, the HIP review process moved 13 out of 71 deaths from other causes to breast cancer in the study group, and 35 out of 73 in the control group (last column of Table 4). Details of the GO calculation in (3) remain a little hazy, but they seem to be looking at an odds ratio like

$$\frac{(73 - 38)/38}{(71 - 58)/58} = 4.1,$$

**Table 4** Deaths through end of follow-up (year 18), among breast cancer cases diagnosed in years 1–7

Classification by Health Insurance Plan			
	Cases	Breast cancer deaths	Deaths from other causes
Study	431	180	58
Control	448	236	38
Death certification (inferred)			
	Cases	Breast cancer deaths	Deaths from other causes
Study	431	167	71
Control	448	201	73

Data from Tables 3.5, 5.1, 6.3, 6.5 Shapiro *et al.*<sup>6</sup> We take Table 3.5 as reporting a review of all breast cancer cases assigned to another cause of death on the death certificate, with results shown in Tables 6.3 and 6.5. Cases and deaths include diagnosis without histological confirmation (5 in study group, 9 in control). Diagnoses reported in Table 2 above had histological confirmation. Health Insurance Plan review data were published for year 7, but other years were similar (ref. 6, p. 32).

which is significantly different from 1.0. Their procedure tests a composite null hypothesis that (i) screening has no effect, and (ii) there is no bias in the death certificates, and (iii) there is no bias in the HIP review. Rejecting the composite null provides little evidence about the particular hypothesis of concern—bias in the HIP review—unless hypotheses (i) and (ii) can be taken as true. Hypothesis (i) is the chief point at issue. Assuming that mammography has no effect in order to prove bias in the HIP review seems perverse.

Another way to handle the possibility of errors in determining the cause of death is to use death itself as the endpoint. The total mortality rate among incident breast cancer cases has less statistical power than mortality from breast cancer, but is unaffected by cause-of-death classifications. That endpoint too favours mammography (Table 5), although the difference is only borderline significant:  $\chi^2 = 3.2$  on 1 degree of freedom,  $P = 0.07$ . The test is two-sided, and does not consider time of death. All screening was done during the first 5 years after entry, so there can be no benefit for women with cancers detected in years 6 and 7. Moreover, as time goes on, the number of deaths from causes other than breast cancer will increase, further reducing power. If we shorten the follow-up time to 10 years (Table 6), power will be better:  $\chi^2 = 5.8$ ,  $P = 0.02$ . As the examples show, significance levels will be different for different time periods; Tables 5 and 6 seem representative of the HIP data. At year 7, the numbers of incident cases in the two arms have equalized, so the comparison of mortality rates is fair.<sup>76</sup> This idea has been developed on data from the Swedish trials, where sample sizes are much larger and results are highly significant.<sup>81</sup>

## All-cause mortality

We have discussed two endpoints so far: (i) breast cancer mortality, and (ii) total mortality amongst breast cancers. However, GO propose all-cause mortality as the definitive endpoint in trials of screening:

The main outcome measure in the screening trials was breast-cancer mortality. This choice seems rational, since larger trials would be needed to show an effect on overall mortality. However, we showed that the assumption that a demonstrated effect on breast-cancer mortality can be translated into a reduction in overall mortality rests on suppositions

**Table 5** Number of breast cancer cases diagnosed in years 1–7. Percentage that died through end of follow-up (year 18)

	Cases	Deaths
Study	431	55%
Control	448	61%

Source of data is Table 4; cause of death is not material; includes diagnosis without histological confirmation.

**Table 6** Number of breast cancer cases diagnosed in years 1–7. Percentage that died in 10 years of follow-up

	Cases	Deaths
Study	431	34%
Control	448	42%

Data sources as in Table 4; includes diagnosis without histological confirmation.

that are not correct.... The only reliable mortality estimates are therefore those for overall mortality.... Thus, although the trials were underpowered for all-cause mortality, the reliable evidence does not indicate any survival benefit of mass screening for breast cancer. (ref. 36, p. 1341)

To clarify the power issue, we sketch a hypothetical clinical trial. Randomize 200 000 women, half to mammography and half to control. Follow the women for 10 years. Assume 100% compliance in both arms, no loss to follow-up, 50% reduction in risk of death from breast cancer in the screening arm, and no other effects. Assume baseline mortality rates as in HIP. A trial of this size would be extraordinarily difficult to implement, and the power to detect a significant reduction in total mortality (at the 0.05 significance level) is barely 0.80. With more realistic compliance parameters and contemporary death rates, power would be even lower. Power is limited because death from breast cancer is a rare event. The conclusion is equally obvious—all-cause mortality is impractical as the defining endpoint for any single trial of mammography.<sup>51</sup> (Pooling the trials is discussed later.)

## The Two-County trial: Sweden

### Study design

The trial was done in two Swedish counties, Kopparberg and Östergötland. (Kopparberg is later called Dalarna, and in some publications, Kopparberg is labelled 'W' while Östergötland is labelled 'E'.) The study population consisted of all women age 40+ in the two counties. Each county was divided into 'blocks', and the blocks further subdivided into 'clusters', these being small geographical areas with administrative identities of their own. The objective was to make the clusters similar within blocks, in terms of demographics and socioeconomics. Different blocks, however, were allowed to be quite different. Kopparberg had seven blocks, each subdivided into three clusters. Some of the blocks were small cities, the clusters being parishes or tax districts; other blocks were more rural, although clusters were called 'municipalities'. Within each block, two clusters were chosen at random for the ASP (Active Study Population, intervention); the remaining cluster went into the PSP (Passive Study Population, control). In Kopparberg, the ASP is therefore about double the size of the PSP. Östergötland had 12 blocks, each subdivided into 2 clusters. Within each block, one cluster was chosen at random for the ASP, and the other went into the PSP: here, the ASP and PSP are nearly equal in size. In this fashion, the entire population of both counties is randomized.

Randomization began block by block in 1977 in Kopparberg and in 1978 in Östergötland. After randomization, the ASP in a block was invited to screening (mediolateral oblique-view mammography only). There were two to four (occasionally five) rounds of screening, with more for the younger women and less for the older. In 1984–1986, the PSP was invited to screening, and then the trial was closed. Subsequently, all women in the two counties were invited to screening on a 'service' basis (as part of their routine health care). Compliance among women age 75+ was poor, so this group was dropped in the analysis phase. Compliance for women age 70–74 was also not so good: those in the ASP were therefore invited to two rounds of screening only.

Incident cases are counted for the period 1977/78–1986, in both arms of the trial, based on Swedish cancer registry data. More specifically, the incidence period is from the randomization of a block until closure of the trial, that is, completion of the first PSP screen in the block. Women with a diagnosis of breast cancer prior to randomization are excluded, using registry data. There are 2468 incident cases. Follow-up of these cases to determine mortality continues indefinitely, and the bulk of the reports on the Two-County trial focus on the experience of these cases. (This design is called ‘the evaluation model’ by Nyström *et al.*,<sup>53,59–61</sup> although no model is involved.) No one source fully describes the Two-County study. Details are in various publications by the investigators.<sup>8–13,17</sup> Further clarification was provided by Duffy and Tabár (personal communication).

## Results

Tables 7 and 8 show virtual equality of breast cancer incidence rates in the ASP and PSP over the period of the trial. The tables also show that death rates from breast cancer are significantly lower in the ASP. The reduction in death rates due to screening (more precisely, due to assignment to ASP) is  $3.9/6.5 = 60\%$  in Kopparberg, and  $4.3/5.7 = 75\%$  in Östergötland (see Discussion). This intention-to-treat analysis suffers from dilution effects. For one thing, there was a 10–20% crossover rate in the treatment arm; there was a 10–15% crossover in control, although some of this may reflect diagnostics rather than screening.<sup>9,11</sup> And, if we were to extend the incidence period beyond the period of the trial, as in the ‘follow-up model’ of Nyström *et al.*,<sup>59</sup> service screening would play a major role.

## The critique

The major points are as follows (Gøtzsche and Olsen, ref. 35, p. 130).

- (1) Cluster randomization is biased.

**Table 7** Two-County data. Counts

	Kopparberg			Östergötland		
	Cases	Deaths	N	Cases	Deaths	N
ASP <sup>a</sup>	694	152	38 589	732	167	38 491
PSP <sup>b</sup>	359	121	18 582	683	213	37 403

<sup>a</sup> Active study population.

<sup>b</sup> Passive study population.

N's from Table 2 in Tabár *et al.*<sup>12</sup> Breast cancer cases and deaths from Table 1 in Tabár *et al.*,<sup>17</sup> with about 18 years of follow-up (to 1998). Incident cases during the period of the trial, approximately 1977–1986, as explained in the text. Deaths during the trial or afterward, among incident cases (evaluation model).

**Table 8** Two-County data. Incidence rates of breast cancer. Death rates from breast cancer. Per 1000 women randomized

	Kopparberg		Östergötland	
	Cases	Deaths	Cases	Deaths
ASP <sup>a</sup>	18.0	3.9	19.0	4.3
PSP <sup>b</sup>	19.3	6.5	18.3	5.7

<sup>a</sup> Active study population.

<sup>b</sup> Passive study population.

Data sources as in Table 7.

This is false, with a minor exception—ratio estimator bias,<sup>82</sup> which affects all rates whose denominators are random (a typical denominator being person-years at risk). Numerators and denominators in the Two-County study are unbiased, because sample averages are unbiased. With a study of this size, ratio estimator bias is likely to be negligible. Of course, variances may be larger with cluster randomization, and this needs to be taken into account when analysing the data.

- (2) The ASP is older than the PSP, ‘ $p < 0.0001$ ’, demonstrating the failure of the randomization.

The difference was discussed by Tabár *et al.*<sup>12,83</sup> It amounts to a few months, and (if anything) dilutes the effect of screening.<sup>84</sup> GO exaggerated the statistical significance of this difference by ignoring the cluster randomization when computing  $P$ .<sup>85–90</sup> Furthermore, there is good evidence to show that randomization was successful, producing comparable groups of women in the ASP and PSP along several important dimensions. For instance, there is near-equality of breast cancer incidence rates before the study began (Figure 1 in Nyström *et al.*<sup>53</sup>). Likewise, death rates from other causes are nearly equal.<sup>11,61</sup>

- (3) There is inconsistent reporting of population size: for instance, 134 867 in 1985<sup>9</sup> and 133 065 in 1989.<sup>12</sup>

GO cite the 1989 paper<sup>12</sup> but miss some crucial details. The Two-County investigators linked their database to the Swedish cancer registry and cleaned the data by eliminating women with diagnosis of breast cancer prior to randomization. Before linkage, such women were excluded at recurrence of disease or death. The 1985 paper<sup>9</sup> was written before linkage; the 1989 paper,<sup>12</sup> after linkage—explaining the difference in reported population size.<sup>49,83</sup>

- (4) There is inconsistent reporting of deaths from breast cancer. For example, take women age 40–49 in Kopparberg. Are the ASP:PSP counts 22:16 or 26:18? (Olsen and Gøtzsche, ref. 37, p. 16).

Table 2 in Tabár *et al.*<sup>15</sup> has 22:16 at 12.5 years of average follow-up, whereas Table II in Tabár *et al.*<sup>16</sup> has 26:18 at 15.5 years (average follow-up for all subjects, our calculation). The difference in counts is due to longer follow-up, which GO ignore.<sup>90</sup> We resolved other ‘discrepancies’ in a similar fashion.

- (5) There is bias in assigning cause of death.

This is the most disturbing of the arguments, and we take it up in some detail.

- (5.1) ‘The decrease in breast-cancer mortality with screening in the Two-County study when the endpoint committee did not know status was similar to that when cause of death was assessed openly (and where we found bias in the classification process). Therefore, our findings that masked endpoint committees make biased assessments are supported.’ (Gøtzsche, ref. 77, p. 2167)

- (5.2) ‘We found data from the Two-County trial (Tabár *et al.*<sup>11</sup>) that could illustrate this possible misclassification

directly ... Among women with a diagnosis of breast cancer, mortality for other cancers was significantly higher (RR = 2.42, [CI] 1.00–5.85) ( $p = 0.05$ ); mortality from all other causes was also higher, although not significantly (RR = 1.37, [CI] 0.93–2.04) ( $p = 0.11$ ).’ (Olsen and Gøtzsche, ref. 37, p. 16)

With respect to (5.1), the agreement between endpoint committees shows if anything that bias is unlikely. GO would have an argument only if they had evidence to demonstrate bias in open reading; but they do not, as will be seen by examining (5.2). Most of the data in Tabár *et al.*<sup>11</sup> show near-equality of death rates from other causes among the ASP and PSP, which makes bias in death classification seem unlikely. Tables 5 and 6 in Tabár *et al.*<sup>11</sup> report deaths by cause among the breast cancer cases, with 25/1295 deaths from other cancers in the ASP and 6/768 in the PSP. That is the probable source for the claimed  $P = 0.05$ , although we cannot quite replicate the calculation. What should we make of this finding? Adjustment for time on risk would increase  $P$ , since the ASP cases live longer than the PSP cases. Adjustment for multiple comparisons—and GO have clearly made many comparisons—would also have a substantial effect on  $P$ . This is not good evidence.

Longer follow-up confirms the view that GO have capitalized on an artifact. For instance, with 8 years of follow-up, Table 11 in Tabár *et al.*<sup>12</sup> shows that risk of death from other causes among breast cancer cases is similar in the ASP and PSP ( $P = 0.7$ ). Using data with 11 years of follow-up in Table 9 of Tabár *et al.*<sup>13</sup> we consider deaths from other causes among the breast cancer cases, comparing the observed number of events in the ASP to an expected number computed from the PSP: observed – expected =  $7 \pm 22$ ,  $P = 0.8$ , taking into account age and county. (Deaths from other causes were allocated proportional to time on risk, which is conservative because hazard rates increase with age.) Tabár *et al.*<sup>51</sup> and Duffy *et al.*<sup>91</sup> analyse more recent data with similar conclusions; in particular, there is significant reduction in all-cause mortality among breast cancer cases, which cannot be explained by errors in death classification.<sup>81</sup>

We turn now to total mortality in the whole ASP and PSP. Of the endpoints under consideration, this one has the least statistical power, but is the most robust against classification error. Nyström *et al.*<sup>53</sup> report on a pooled analysis of the Swedish trials: in combination, nearly 250 000 women were randomized and followed for 16 years. There is a 2% reduction in overall mortality among those invited to screening (RR = 0.98; 95% CI: 0.96, 1.00). The standard error computed by Nyström *et al.*<sup>53</sup> does not take clustering into account; however, sample design has relatively little effect on standard errors for death rates.<sup>18,53,91</sup> By contrast, the effect of clustering on standard errors for average ages remains to be studied. The results of Nyström *et al.*<sup>53</sup> support the following conclusions: (i) screening has an impact on total mortality, and (ii) bias in death certification plays little role in explaining the results of the clinical trials.

In summary, the Swedish data demonstrate a protective effect from screening if breast cancer deaths are determined from death certificate data or by either of two independent endpoint committees; screening has an effect whether breast cancer is the underlying cause of death, or present at death.<sup>53,59–61</sup> Further support for the Two-County data comes from the statistical analyses described above, using deaths among breast cancer

cases as the endpoint,<sup>12,51,91</sup> or deaths in the whole study population.<sup>53</sup> GO have not made a case for differential bias in death classification. Their other points are even less convincing. Reviewers have concurred with this assessment: see §§ 4.2, 5 in Health Council of The Netherlands.<sup>48</sup>

## Other work

Statistician Donald Berry has the following points.

- (1) ‘The presence of a number of well-known biases (including ‘lead-time bias’ and ‘length bias’) make it difficult to ascertain the benefits of screening.’<sup>46</sup>

‘Lead-time bias’ means that screening speeds up detection, so incidence is higher in the study group at the beginning of follow-up. ‘Length bias’ means that screening is likely to pick up slower rather than faster growing tumours. However, clinical trials like HIP or Two-County are skewed by neither bias, because (i) they use death from breast cancer as the endpoint not detection rates or lifespan after detection, and (ii) they use intention-to-treat analyses. Indeed, that is why attention is restricted to evidence from clinical trials.<sup>46</sup>

- (2) ‘A major issue in randomized studies—including screening trials—is lack of compliance with the study protocols.... participants skipped their assigned mammograms about 20% of the time. In addition some participants assigned to be control subjects opted to have screening mammograms. The extent of the bias caused by lack of compliance is not known.’<sup>46</sup>

Generally, however, crossover dilutes the effect of screening: intention-to-treat measures the effect of assignment not the effect of treatment.

- (3) ‘Women with pre-existing breast cancer were preferentially excluded from the screening group. The problem was most severe in the New York trial....’<sup>47</sup>

As shown above, the evidence for preferential exclusions in HIP or Two-County is speculative at best. This point gets a different twist in an interview (*New York Science Times* 9 April 2002, p. D4).

“Only the screening group had mammograms,” Dr Berry said. “On second look at a woman’s first mammogram, one might find that breast cancer was present at the time but it had been missed,” he said. “So more women might have been excluded from the mammography group after they developed breast cancer.” [But women were excluded if their breast cancer was diagnosed before randomization; what was found or missed on mammography is irrelevant to exclusion.]

- (4) ‘... the scheduled control mammogram slipped in all three [Swedish] trials, allowing for more time to detect cancers in the control group [after last ASP screen until completion of first PSP screen].’<sup>47</sup>

The Two-County trial followed its timetable as well as could be expected, and obtained near-equality of incidence rates in the

ASP and PSP due to its design. In theory, however, time elapsed between the last screen of ASP and first screen of PSP could create a bias in favour of screening ('time-lapse bias'). The argument is not entirely straightforward. The evaluation model<sup>59</sup> counts only cases incident during the period of the trial; however, all deaths among these cases are counted, including those occurring after the trial closes. Screening the PSP when closing the trial advances the time of diagnosis for some breast cancers; subsequent deaths are then counted against the PSP. That is what creates a possible bias in favour of mammography. However, Tabár *et al.*<sup>8</sup> showed a significant effect from screening the ASP, at a time when only 5% of the PSP had been screened.<sup>91</sup> These results cannot be affected by 'time-lapse bias'.

The follow-up model<sup>59</sup> looks at all cancers incident after the trial starts, including cancers incident after the PSP is screened, and is also immune to time-lapse bias. The protective effect of mammography is significant according to the follow-up model: Table II in Nyström *et al.*,<sup>59</sup> after pooling the two counties. Here, dilution is the problem: the effect of mammography is understated, because the PSP was screened at the end of trial, and continues to receive service screening after the trial is over. The theoretical bias created by the time lapse has no practical relevance.

## Canadian National Breast Screening Study

This negative study is judged to have acceptable quality by GO and Berry. The study covered two age groups, 40–49 and 50–59. Women were recruited in the period 1980 to 1985, and followed to 1988; average follow-up was 8.5 years.<sup>40,41</sup> There was late follow-up for the younger women, to 1993; and for the older women, to 1996.<sup>42–45</sup> Two-view mammography was used, initially craniocaudal and mediolateral, but the latter was changed to mediolateral oblique in 1985.<sup>92</sup>

To describe the designs, the following abbreviations will be useful:

MA = mammography, PE = physical exam,  
BSE = breast self exam, UC = usual care.

In both age groups, subjects were volunteers who turned up at a screening centre and signed the consent form. PE was done mainly by highly trained nurses. In both age groups and both arms of the trial, participants were shown how to do BSE. All participants randomized to treatment were invited to four screens, and 62% of them were invited to a fifth screen. The treatment screens in both age groups comprised MA and PE. The control condition, however, was different in the two age groups. In the group aged 40–49, the control was PE at first screen then UC only, that is, only one round of screening; 50 430 women were randomized. In the group aged 50–59, control women were offered four or five rounds of screening by PE; 39 405 women were randomized.

CNBSS differed from HIP (Table 9). The HIP trial measured the impact of screening by MA and PE compared with usual care, whereas CNBSS2 measured the impact of screening by MA and PE relative to screening by PE only. (More precisely, the trials are measuring the impact of invitations to be screened.) Furthermore, Two-County differs from HIP and CNBSS. The different trials are measuring different things.

**Table 9** Comparing Canadian National Breast Screening Study (CNBSS), Two-County study, and Health Insurance Plan (HIP)

	Screening		Control
	Modality	No.	
CNBSS1: 40–49	MA <sup>a</sup> + PE <sup>b</sup>	4–5	PE at first screen then UC <sup>c</sup>
CNBSS2: 50–59	MA + PE	4–5	4–5 rounds of PE
Two-County	MA	2–5	UC then 1 round of MA
HIP	MA + PE	4	UC

<sup>a</sup> Mammography.

<sup>b</sup> Physical exam.

<sup>c</sup> Usual care.

Power is an issue. CNBSS had low power because there were few deaths from breast cancer—66 in CNBSS1 (Table 9, Miller *et al.*<sup>40</sup>) and 77 in CNBSS2 (Table 9, Miller *et al.*<sup>41</sup>). Any effect, or lack of effect, can be only be demonstrated with poor precision. Moreover, CNBSS has been dogged from the beginning by accusations of (i) poor radiology, and (ii) 'steering' high-risk women to MA; the trialists and others have responded to the accusations.<sup>92–103</sup> We consider the points in turn.

According to Baines, McFarlane, and Miller,<sup>93</sup> centre radiologists only agreed with the reference radiologist 30–50% of the time. 'Observer error and technical problems' led to delayed detection in 22–35% of cancers. Suggestions—for instance, don't mark up the film with a grease pencil—'were sometimes resisted by center radiologists'. Baines and Miller were the two lead investigators on CNBSS, and McFarlane was the reference radiologist. Their report is not reassuring about the quality of the radiology.

The evidence on steering is generally anecdotal but should not be dismissed—or accepted—for that reason. There is one statistical analysis to report.<sup>94,100</sup> In CNBSS1 (the 40–49 age group), 22 advanced breast cancers (4+ nodes involved) were detected by PE at first screen: 17 in the treatment arm and 5 in the control arm,  $P = 0.017$ . In the treatment group, there were two additional cancers detected by MA only, which are irrelevant for present purposes.

## Responses

Bailar and MacMahon<sup>102</sup> (with commentary ref. 103) assessed the randomization and found it acceptable. Bailar and MacMahon did not consider the radiology or follow-up procedures. They did not look at CNBSS2 (women age 50–59). They did not follow their own plan for the review: among other things, they did not interview field staff. They acknowledge the 17:5 imbalance.<sup>94,100</sup> They acknowledge that the protocol for the trial was not followed, with the result that (i) steering would have been easy to do, and (ii) there could have been some motivation for steering. Indeed, assignment to treatment or control was generally done locally, *after results of physical examination were known*. Nurses may have wanted high-risk subjects to get what seemed to be the better treatment. The CNBSS log books were altered, but 'document experts found no evidence of a deliberate attempt to conceal the alterations'. That seems weak: among other things, randomization could have been subverted simply by changing the order in which names were entered into the log books.

Baines (ref. 99, p. 329) notes that a comparison of advanced cancers detected by MA + PE in treatment to those detected by PE in control (19 to 5) is biased. Such a comparison might indeed be biased, but it is not the comparison that was made.<sup>94,100</sup> Baines also addresses questions about follow-up, radiology, and steering, as do other papers.<sup>42-45,101</sup> CNBSS remains controversial in some quarters<sup>20,23,103-105</sup> but approved in others.<sup>28,29</sup>

**Gøtzsche and Olsen on the Canadian National Breast Screening Study**

GO have not addressed the radiology, except to say that CNBSS was the only trial to have assessed the mammograms (ref. 37, p. 5) and CNBSS found small tumours.<sup>35</sup> But that does not address questions about which tumours were missed, which were found, and when they were found. Baines, McFarlane, and Miller<sup>93</sup> are not reassuring about study quality. GO cite this paper without discussion.

With respect to the statistical evidence on steering, say GO, the 17:5 imbalance in advanced cancers detected by PE at first screen of the younger women

‘is a post-hoc subgroup finding which is probably a result of the intervention, and exclusion of the deaths caused by these cancers does not change the result.... (ref. 37, p. 10)

(i) GO are not in the strongest of positions to complain about ‘post hoc subgroup findings’, since most of their analysis is *post hoc*. (ii) This particular finding is hardly *post hoc*, being mentioned in the original report of the trial (Miller *et al.*, ref. 40, pp. 1470, 1473). (iii) The 22 cancers detected by PE cannot be ‘a result of the intervention’, since PE was done at first screen in both arms of the trial. (iv) These 22 cases may only be the tip of the iceberg. We cannot know how many other high-risk women were steered to treatment. The answer may be 0. But this is the number in question, and until this number can be estimated, adjustment for steering is impossible. GO also say:

A persistent criticism has been that an effect would be difficult to find because the breasts of all women in the age-group 50–59 years were physically examined regularly. This criticism is unwarranted because mammography will identify many tumours that are too small to be detected on physical examination alone. Furthermore, any effect of physical examination is likely to be small. A study of 122 471 women found no effect of regular self-examination of the breast on breast-cancer mortality after 9 years of follow-up, even though twice as many of the intervention group consulted an oncologist. (ref. 35, p. 132)

This response to criticism of CNBSS is irrelevant, because breast self examination is not the same as breast physical examination. Breast self examination is done by the woman herself: breast physical examination is done to the woman by a trained professional. Breast self examination may be of little value,<sup>106</sup> whereas breast physical examination is effective at cancer detection. That seems to be the case in CNBSS, especially among the younger women (Table 10). PE even detects many cancers missed by MA.

**CNBSS1: Invasive breast cancer**

Table 11 summarizes the incidence of invasive breast cancer and deaths after 8.5 years of follow-up in CNBSS1:<sup>40</sup> there is no saving in lives from mammography. On the other hand, Figure 3 in Miller *et al.*<sup>45</sup> suggests a beneficial effect at the end of late follow-up: the cumulative mortality curves cross. If advanced breast cancers detected by PE at baseline are excluded from the study population, the effect will be about 20%, although statistical significance is not achieved. Table 11 and results from late follow-up support the idea that something went awry in the randomization: indeed, at 8.5 years, there is a highly significant excess of invasive cancers in the screening group. The excess persists through late follow-up,<sup>45</sup> and there is a similar (less significant) excess in CNBSS2.<sup>44</sup> These differentials are hard to explain on the basis of lead time, unless many cancers classified as ‘invasive’ do not progress to clinical significance, raising other questions about CNBSS. The differential in CNBSS1 is not significant at the end of late follow-up, and strikingly good balance is achieved between MA and UC groups in terms of risk factors and referral for surgical evaluation, observations which reduce the force of Table 11. Considering all these factors, we see no reason to judge CNBSS as higher in quality than HIP or Two-County.

**Conclusion**

In their meta-analysis, GO<sup>35-38</sup> excluded positive studies on mammography, like HIP and Two-County. Consequently, there was no benefit from screening. Were the exclusions justified? To answer this question, we had to examine in detail the various studies and the criticisms raised by GO. We focused on HIP, Two-County, and CNBSS. The chief criticisms levelled at HIP were (i) differential exclusion of high-risk women, and (ii) imbalance of risk factors at baseline. The chief criticisms of the Two-County trial were (i) use of cluster randomization, (ii) imbalance at

**Table 10** Canadian National Breast Screening Study data. Treatment arm. Cancers detected by various modalities

	Cancers detected by		
	MA <sup>a</sup> only	PE <sup>b</sup> only	Both
Age 40–49	105	81	73
Age 50–59	180	64	89

<sup>a</sup> Mammography.

<sup>b</sup> Physical exam.

Computed by us from Table 5 in Miller *et al.*<sup>40</sup> and Table 5 in Miller *et al.*<sup>41</sup>

**Table 11** Cumulative number of invasive breast cancer to end of follow-up (mean 8.5 years). Deaths from breast cancer through 7th year after entry. Canadian National Breast Screening Study age 40–49

	No.	Deaths
MA <sup>a</sup>	331	38
UC <sup>b</sup>	272	28
Difference	59 ± 25	10 ± 8
z	2.38	1.17

<sup>a</sup> Mammography.

<sup>b</sup> Usual care.

Computed by us from Tables 7 and 9 in Miller *et al.*<sup>40</sup>

baseline, and (iii) inconsistent reporting of data. Both trials were criticized for bias in determining cause of death.

With respect to HIP, point (i) reflects a misunderstanding of the design. Point (ii) reflects a misunderstanding of the table that was analysed. This is bothersome, because the design features relevant to (i) are discussed in the reports GO cite, within a few pages of the numbers they use. Similarly, the table cited for (ii) contains most of the relevant information in headnotes and footnotes. For Two-County, points (i) and (ii) show some misunderstanding of basic statistical concepts like bias, variance, and clustering. Point (iii) depends on lack of care in reading tables, or lack of attention to explanatory material presented within a few pages of the tables used. Bias in determining cause of death remains a possibility. However, evidence cited to demonstrate this bias evaporates when examined, and there is compelling evidence on the other side, including comparability of death rates in treatment and control from causes other than breast cancer, reduction of total mortality among breast cancer cases, and a reduction in total mortality in the whole intervention group when the Swedish trials are pooled.

CNBSS has been criticized for (i) failures in randomization, and (ii) poor mammography. It has also been observed that (iii) CNBSS compared mammography to physical breast examination by trained personnel, rather than comparing mammography to

usual care. GO's defence of the randomization mischaracterizes the evidence.<sup>94,100</sup> The comparison was not post hoc; nor could the finding possibly have resulted from the intervention, because the comparison was of tumours discovered by physical examination at baseline in each arm of the trial. With respect to (ii), Baines, McFarlane, and Miller<sup>93</sup>—the two principal investigators and the reference radiologist—are not reassuring about study quality, and GO have not discussed this paper. GO's response to point (iii) involves a confusion between breast examination by a practitioner and self examination; it also ignores CNBSS data on the efficacy of breast examination by a practitioner (Table 10).

GO's critique of the positive studies (HIP and Two-County), like their defence of CNBSS, is careless at best. Rather than clarifying the issues, their papers have instead generated much confusion. Clinical trials of mammography have led to substantial advances in understanding breast cancer, and a substantial reduction in mortality from this disease. It is time to move on,<sup>107–109</sup> although some questions may remain.<sup>110,111</sup>

## Acknowledgements

We thank László Tabár and Stephen Duffy for their patience in answering questions. Raymond Fink, Judith Goldberg, Jon McAuliffe, and Barbara Monsees were also very helpful.

### KEY MESSAGES

- There is good evidence from clinical trials that mammographic screening reduces the death rate from breast cancer.
- The critique by Gøtzsche and Olsen has little merit and has generated much confusion.

## References

- <sup>1</sup> Shapiro S, Strax P, Venet L. Periodic breast cancer screening in reducing mortality from breast cancer. *JAMA* 1971;**215**:1777–85.
- <sup>2</sup> Shapiro S, Strax P, Venet L, Venet W. Changes in 5-year breast cancer mortality in a breast cancer screening program. In: *Proc Seventh Natl Cancer Conference*. Philadelphia: Lippincott, 1972, pp. 663–78.
- <sup>3</sup> Fink R, Shapiro S, Roeser R. Impact of efforts to increase participation in repetitive screenings. *Am J Public Health* 1972;**62**:328–36.
- <sup>4</sup> Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* 1977;**39**(Suppl.):2772–82.
- <sup>5</sup> 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. *Natl Cancer Inst Monogr* 1985;**67**: 65–79. With discussion.
- <sup>6</sup> 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.
- <sup>7</sup> Shapiro S, Venet W, Strax P, Venet L. Current results of the breast cancer screening randomized trial: The Health Insurance Plan (HIP) of greater New York study. In: Day N, Miller A (eds). *Screening for Breast Cancer*. Toronto: Hans Huber, 1988, pp. 3–15.
- <sup>8</sup> Tabár L, Gad A. Screening for breast cancer: The Swedish trial. *Radiology* 1981;**138**:219–22.
- <sup>9</sup> Tabár L, Fagerberg CJ, Gad A *et al*. Reduction in mortality from breast cancer after mass screening with mammography: Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet* 1985;**i**:829–32.
- <sup>10</sup> Fagerberg CJG, Tabár L. The results of periodic one-view mammography screening in a randomized controlled trial in Sweden. Part 1: Background, organization, screening program, tumor findings. In: Day N, Miller A (eds). *Screening for Breast Cancer*. Toronto: Hans Huber, 1988, pp. 33–38.
- <sup>11</sup> Tabár L, Fagerberg CJG, Day NE. The results of periodic one-view mammography screening in a randomized controlled trial in Sweden. Part 2: Evaluation of the results. In: Day N, Miller A (eds). *Screening for Breast Cancer*. Toronto: Hans Huber, 1988, pp. 39–44.
- <sup>12</sup> Tabár L, Fagerberg G, Duffy SW, Day NE. The Swedish two county trial of mammographic screening for breast cancer: Recent results and calculation of benefit. *J Epidemiol Community Health* 1989;**43**:107–14.
- <sup>13</sup> Tabár L, Fagerberg G, Duffy SW, Day NE, Gad A, Grönroft O. Update of the Swedish two-county program of mammographic screening for breast cancer. *Radiol Clin North Am* 1992;**30**:187–210.
- <sup>14</sup> Tabár L, Fagerberg G, Day NE, Duffy SW, Kitchin RM. Breast cancer treatment and natural history: New insights from results of screening. *Lancet* 1992;**339**:412–14.

- <sup>15</sup> 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.
- <sup>16</sup> Tabár L, Vitak B, Chen HH, Prevost TC, Duffy SW. Update of the Swedish two-county trial of breast cancer screening: Histologic grade-specific and age-specific results. *Swiss Surg* 1999;**5**:199–204.
- <sup>17</sup> Tabár L, Vitak B, Chen HH *et al*. The Swedish two-county trial twenty years later: Updated mortality results and new insights from long-term followup. *Radiol Clin North Am* 2000;**38**:625–51.
- <sup>18</sup> Nixon R, Prevost TC, Duffy SW, Tabár L, Vitak B, Chen HH. Some random-effects models for the analysis of matched-cluster randomised trials: application to the Swedish two-county trial of breast-cancer screening. *J Epidemiol Biostat* 2000;**5**:349–58.
- <sup>19</sup> Tabár L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001;**91**:1724–31.
- <sup>20</sup> Feig SA. Effect of service screening mammography on population mortality from breast carcinoma. *Cancer* 2002;**95**:451–57.
- <sup>21</sup> Duffy SW, Tabár L, Chen HH *et al*. The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties: A collaborative evaluation. *Cancer* 2002;**95**:458–69.
- <sup>22</sup> Tabár L, Yen M-F, Vitak B, Chen H-HT, Smith RA, Duffy SW. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *Lancet* 2003;**361**:1405–10.
- <sup>23</sup> Cady B, Michaelson JS. The life-sparing potential of mammographic screening. *Cancer* 2001;**91**:1699–703.
- <sup>24</sup> Margolese RG, Fisher B, Hortobagyi GN, Bloomer WD. Neoplasms of the Breast. In: Bast OC Jr, Kufe DW, Pollock RE *et al*. (eds). *Cancer Medicine*. Hamilton, Ontario, Canada: BC Decker, 2000, Ch. 118, pp. 1735–822. Available on-line at www.cancer.org.
- <sup>25</sup> US Preventive Services Task Force. *Guide to Clinical Preventive Services. 1st Edn*. Washington, DC: US Department of Health and Human Services, Office of Public Health and Science, Office of Disease Prevention and Health Promotion, 1989.
- <sup>26</sup> Nass SJ, Henderson IC, Lashof J (eds). *Mammography and Beyond: Developing Technologies for the Early Detection of Breast Cancer*. Washington, DC: Institute of Medicine, National Research Council, 2001.
- <sup>27</sup> International Agency for Research on Cancer. *Breast Cancer Screening*. Volume 7 of the IARC Handbooks of Cancer Prevention. Lyon: IARC, 2002.
- <sup>28</sup> US Preventive Services Task Force. Screening for breast cancer: Recommendations and rationale. *Ann Intern Med* 2002;**137**:344–46.
- <sup>29</sup> US Preventive Services Task Force. Breast cancer screening: A summary of the evidence. *Ann Intern Med* 2002;**137**:347–67.
- <sup>30</sup> Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. *JAMA* 1995;**273**:149–54. Discussion, 1995;**274**:380–83.
- <sup>31</sup> Kerlikowske K. Efficacy of screening mammography among women aged 40 to 49 years and 50 to 69 years: Comparison of relative and absolute benefit. *J Natl Cancer Inst Monogr* 1997;**22**:79–86.
- <sup>32</sup> Gohagan JK (ed.). *National Institutes of Health Consensus Conference on Breast Cancer Screening for Women Ages 40–49*. *J Natl Cancer Inst Monogr* 1997;**22**.
- <sup>33</sup> Kattlove H, Liberati A, Keeler E, Brook RH. Benefits and costs of screening and treatment for early breast cancer. *JAMA* 1995;**273**:142–48. Discussion, 1995;**274**:380–83.
- <sup>34</sup> Wright CJ and Mueller CB. Screening mammography and public health policy—the need for perspective. *Lancet* 1995;**346**:29–32. Discussion, 1995;**346**:852.
- <sup>35</sup> Gøtzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000;**355**:129–34.
- <sup>36</sup> Olsen O, Gøtzsche PC. Cochrane review on screening for breast cancer with mammography. *Lancet* 2001;**358**:1340–42.
- <sup>37</sup> Olsen O, Gøtzsche PC. *Screening for Breast Cancer with Mammography (Cochrane Review)*. Oxford: Update Software. The Cochrane Library, Issue 4, 2001.
- <sup>38</sup> Olsen O, Gøtzsche PC. *Systematic Screening for Breast Cancer with Mammography*. 2001. <http://image.thelancet.com/lancet/extra/fullreport.pdf>
- <sup>39</sup> Miller AB, Howe GR, Wall C. The National Study of Breast Cancer Screening. *Clin Invest Med* 1981;**4**:227–58.
- <sup>40</sup> 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.
- <sup>41</sup> Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Can Med Assoc J* 1992;**147**:1477–88.
- <sup>42</sup> Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study: Update on breast cancer mortality. *J Natl Cancer Inst Monogr* 1997;**22**:37–41.
- <sup>43</sup> Baines CJ, Miller AB. Mammography versus clinical examination of the breasts. *J Natl Cancer Inst Monogr* 1997;**22**:125–29.
- <sup>44</sup> 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**:1490–99.
- <sup>45</sup> 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.
- <sup>46</sup> Berry DA. Benefits and risks of screening mammography for women in their forties: A statistical appraisal. *J Natl Cancer Inst* 1998;**90**:1431–39.
- <sup>47</sup> Berry DA. Testimony, Senate hearing, 28 February 2002.
- <sup>48</sup> Health Council of The Netherlands. *The Benefit of Population Screening for Breast Cancer with Mammography*. The Hague, 2002.
- <sup>49</sup> Duffy SW. Interpretation of the breast screening trials: A commentary on the recent paper by Gøtzsche and Olsen. *Breast* 2001;**10**:209–12.
- <sup>50</sup> Duffy SW, Tabár L, Smith RA. The mammographic screening trials: Commentary on the recent work by Olsen and Gøtzsche. *CA Cancer J Clin* 2002;**52**:68–71.
- <sup>51</sup> Tabár L, Duffy SW, Yen M-F *et al*. All-cause mortality among breast cancer patients in a screening trial: Support for breast cancer mortality as an end point. *J Med Screen* 2002;**9**:159–62.
- <sup>52</sup> Tabár L, Smith RA, Vitak B *et al*. Mammographic screening: A key factor in the control of breast cancer. Randomisation and endpoint evaluation. *Cancer J* 2003;**9**:15–27.
- <sup>53</sup> 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.
- <sup>54</sup> Baum M. Screening mammography re-evaluated. Letter. *Lancet* 2000;**355**:751.
- <sup>55</sup> Rozenberg S, Liebens F, Ham H. Screening mammography re-evaluated. Letter. *Lancet* 2000;**355**:751–52.
- <sup>56</sup> Thornton H. Screening for breast cancer with mammography. Letter. *Lancet* 2001;**358**:2165.
- <sup>57</sup> Vaidya JS. Screening for breast cancer with mammography. Letter. *Lancet* 2001;**358**:2166.
- <sup>58</sup> Dixon-Woods M, Baum M, Kurinczuk JJ. Screening for breast cancer with mammography. Letter. *Lancet* 2001;**358**:2166–67.

- <sup>59</sup> Nyström L, Rutqvist LW, Wall S *et al.* Breast cancer screening with mammography: Overview of Swedish randomised trials. *Lancet* 1993; **341**:973–78. Discussion, 1993; **341**:1531–32.
- <sup>60</sup> Nyström L, Larsson LG, Rutqvist LE *et al.* Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials: A comparison between official statistics and validation by an endpoint committee. *Acta Oncol* 1995; **34**:145–52.
- <sup>61</sup> 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.
- <sup>62</sup> Tabár L, Smith RA, Duffy SW. Update on effects of screening mammography. Letter. *Lancet* 2002; **360**:337.
- <sup>63</sup> Bonneux L. Update on effects of screening mammography. Letter. *Lancet* 2002; **360**:337–38.
- <sup>64</sup> Gøtzsche PC. Update on effects of screening mammography. Letter. *Lancet* 2002; **360**:338–39.
- <sup>65</sup> Cheng KK. Update on effects of screening mammography. Letter. *Lancet* 2002; **360**:339.
- <sup>66</sup> Gulbrandsen P. Update on effects of screening mammography. Letter. *Lancet* 2002; **360**:339.
- <sup>67</sup> Nyström L, Andersson I, Bjurstam N, Frisell J, Rutqvist LE. Update on effects of screening mammography. Authors' reply. *Lancet* 2002; **360**:339–40.
- <sup>68</sup> Freedman DA, Pisani R, Purves R, Adhikari A. *Statistics*. 2nd Edn. New York: WW Norton & Company, Inc, 1991.
- <sup>69</sup> Angrist J, Imbens G. Identification and estimation of local average treatment effects. *Econometrica* 1994; **62**:467–75.
- <sup>70</sup> Gøtzsche PC. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. Letter. *Cancer* 2002; **94**:578.
- <sup>71</sup> Olesen O. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. Letter. *Cancer* 2002; **94**:578–79.
- <sup>72</sup> Ponzzone R, Baum M. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. Letter. *Cancer* 2002; **94**:580.
- <sup>73</sup> Kopans DB. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. Letter. *Cancer* 2002; **94**:580–81.
- <sup>74</sup> Tabár L, Duffy SW, Smith RA. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. Authors' reply. *Cancer* 2002; **94**:581–83.
- <sup>75</sup> Miller AB. Screening for breast cancer with mammography. Letter. *Lancet* 2001; **358**:2164.
- <sup>76</sup> 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.
- <sup>77</sup> Gøtzsche PC. Screening for breast cancer with mammography. Author's reply. *Lancet* 2001; **358**:2167–68.
- <sup>78</sup> Rosner D, Bedwani RN, Vana J, Baker HW, Murphy GP. Noninvasive breast carcinoma: Results of a national survey by the American College of Surgeons. *Ann Surg* 1980; **192**:139–47.
- <sup>79</sup> Ernster VL, Barclay J. Increases in ductal carcinoma *in situ* (DCIS) of the breast in relation to mammography: A dilemma. *J Natl Cancer Inst Monogr* 1997; **22**:151–56.
- <sup>80</sup> Cady B, Stone MD, Schuler JG, Thakur R, Wanner MA, Lavin PT. The new era in breast cancer: Invasion, size, and nodal involvement dramatically decreasing as a result of mammographic screening. *Arch Surg* 1996; **131**:301–08.
- <sup>81</sup> Larsson LG, Nyström L, Wall S *et al.* The Swedish randomised mammography screening trials: analysis of their effect on the breast cancer related excess mortality. *J Med Screening* 1996; **3**:129–32.
- <sup>82</sup> Cochran WG. *Sampling Techniques*. 3rd Edn. New York: John Wiley & Sons, 1977.
- <sup>83</sup> Duffy SW, Tabár L. Screening mammography re-evaluated. Letter. *Lancet* 2000; **355**:747–48.
- <sup>84</sup> de Koning HJ. Assessment of nationwide cancer-screening programmes. *Lancet* 2000; **355**:80–81.
- <sup>85</sup> Moss S, Blanks R, Quinn MJ. Screening mammography re-evaluated. Letter. *Lancet* 2000; **355**:748.
- <sup>86</sup> Nyström L. Screening mammography re-evaluated. Letter. *Lancet* 2000; **355**:748–49.
- <sup>87</sup> Law M, Hackshaw A, Wald N. Screening mammography re-evaluated. Letter. *Lancet* 2000; **355**:749–50.
- <sup>88</sup> Cates C, Senn S. Screening mammography re-evaluated. Letter. *Lancet* 2000; **355**:750.
- <sup>89</sup> Senn S. Screening for breast cancer with mammography. Letter. *Lancet* 2001; **358**:2165.
- <sup>90</sup> Duffy SW, Tabár L, Smith RA. Screening for breast cancer with mammography. Letter. *Lancet* 2001; **358**:2166.
- <sup>91</sup> Duffy SW, Tabár L, Vitak B *et al.* The Swedish Two-County Trial of mammographic screening: Cluster randomisation and endpoint evaluation. *Ann Oncol* 2003, in press.
- <sup>92</sup> Baines CJ, Miller AB, Kopans DB *et al.* Canadian National Breast Screening Study: Assessment of technical quality by external review. *Am J Roentgenol* 1990; **155**:743–47. Discussion, 1990; **155**:748–49, 1133–34.
- <sup>93</sup> Baines CJ, McFarlane DV, Miller AB. The role of the reference radiologist. Estimates of interobserver agreement and potential delay in cancer detection in the national breast screening study. *Invest Radiol* 1990; **25**:971–76.
- <sup>94</sup> Mettlin CJ, Smart CR. The Canadian National Breast Screening Study: An appraisal and implications for early detection policy. *Cancer* 1993; **72**(Suppl.):1461–65.
- <sup>95</sup> Kopans DB, Feig SA. The Canadian National Breast Screening Study: A critical review. *Am J Roentgenol* 1993; **161**:755–60.
- <sup>96</sup> Burhenne LJ, Burhenne HJ. The Canadian National Breast Screening Study—A Canadian critique. *Am J Roentgenol* 1993; **161**:761–63.
- <sup>97</sup> Boyd NF, Jong RA, Yaffe MJ, Tritchler D, Lockwood G, Zylak CJ. A critical appraisal of the Canadian National Breast Cancer Screening Study. *Radiology* 1993; **189**:661–63.
- <sup>98</sup> Kopans DB, Halpern E, Hulka CA. Statistical power in breast cancer screening trials. *Cancer* 1994; **74**:1196–203. Discussion, 1994; **74**:1204–16.
- <sup>99</sup> Baines CJ. The Canadian National Breast Cancer Screening Study: A perspective on criticism. *Ann Intern Med* 1994; **120**:326–34.
- <sup>100</sup> 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.
- <sup>101</sup> Baines CJ. The Canadian National Breast Cancer Screening Study: Why? What next? And so what? *Cancer* 1995; **76**(Suppl.):2109–12.
- <sup>102</sup> Bailar JC 3rd, MacMahon B. Randomization in the Canadian National Breast Screening Study: A review for evidence of subversion. *Can Med Assoc J* 1997; **156**:193–99.
- <sup>103</sup> Boyd NF. The review of randomization in the Canadian National Breast Screening Study. *Can Med Assoc J* 1997; **156**:207–09. Discussion, 1997; **157**:247–250.
- <sup>104</sup> Cady B. The screening mammography: The continuous dilemma. *Breast J* 2002; **8**:185–86.
- <sup>105</sup> Kopans DB. The most recent breast cancer screening controversy about whether mammographic screening benefits women at any age: Nonsense and nonsense. *Am J Roentgenol* 2003; **180**:21–26.

- <sup>106</sup> Thomas DB, Gao DL, Self SG *et al.* Randomized trial of breast self-examination in Shanghai: Methodology and preliminary results. *J Natl Cancer Inst* 1997;**89**:355–65.
- <sup>107</sup> Gelmon KA, Olivotto I. The mammography screening debate: Time to move on. *Lancet* 2002;**359**:904–05.
- <sup>108</sup> Begg CB. The mammography controversy. *Oncologist* 2002;**7**:174–76. Editorial commentary, 2002;**7**:170–73.
- <sup>109</sup> Fletcher SW, Elmore JG. Mammographic screening. *New Engl J Med* 2003;**348**:1672–80.
- <sup>110</sup> Sox H. Screening mammography for younger women: Back to basics. *Ann Intern Med* 2002;**137**:361–62.
- <sup>111</sup> Goodman SN. The mammography dilemma: A crisis for evidence-based medicine? *Ann Intern Med* 2002;**137**:363–65.