

On the benefits and harms of screening for breast cancer

Peter C Gøtzsche

In their qualitative review,¹ Freedman, Petitti and Robins (FPR) claim that our critique of the randomized screening trials has little merit; that there is no reason to believe that the Canadian study was of better quality than the New York Health Insurance Plan (HIP) study or the Two-County study; and that the prior consensus on mammography was correct. However, their review suffers from erroneous assumptions and biased statistical analyses, and their quotations are often selective and misleading. In my discussion of the issues, I will follow when possible the sequence of arguments used by FPR.

Overdiagnosis and overtreatment

FPR claim in their abstract that early detection leads to less invasive therapy. This could have been true, if the only effect of screening had been to detect the same tumours earlier that are detected later if women are not screened. FPR naively believe that screening does just that, i.e. does not lead to overdiagnosis. They note, for example, that the incidence of breast cancers in the New York HIP study is the same in the control group as in the screening group 5–7 years after screening started (their table 2), as they expected. I will explain under the HIP study below why this argument is faulty.

The level of overdiagnosis can be studied reliably in the trials from Canada and Malmö which did not differentially exclude women with prior breast cancer after randomization and did not introduce early, systematic screening of the whole control group.^{2–4} There was an overdiagnosis of 30%^{5,6} which corresponds closely to the 31% excess surgery we have previously described^{2,3} (Table 1). A similar result was seen for the trials that screened the whole control group when only cancers before this screen were included^{7–9} (Table 2).

The excess surgery rate was 20% for mastectomies.^{2,3} We have discussed in detail why it is likely that even today, there would be about 20% more mastectomies when women are screened than if they are not screened.³ In Southeast Netherlands, for example, when screening was introduced from 1990 to 1998, the number of women who underwent breast-conserving surgery increased by 71%, and the number of women who underwent mastectomy increased by 84%.¹⁰ If the study had included carcinoma *in situ* there would have been even more mastectomies.^{3,11} A study from Italy claimed that screening had not led to an increase in mastectomies,¹² but this study had no control group and the premises for the study were also faulty¹³ (see also other letters at www.bmj.com).

If carcinoma *in situ* was the usual precursor for invasive cancer, the incidence of early-stage invasive breast cancer should decrease as the incidence of carcinoma *in situ* increased, but the opposite has happened.^{14,15} Before screening was introduced in US, the age-adjusted incidence of breast cancer was rather constant (ref. 15, Table IV-1). When screening spread in the 1980s, as evidenced by a sharp rise in cases of carcinoma *in situ* (Figure 1), cases of invasive breast cancer increased by 26% in only 7 years, and has remained elevated ever since.¹⁵ If carcinoma *in situ* cases are added, the increase becomes 35%, in good agreement with our data.

It can be discussed whether a tumourectomy is always preferable to, or less aggressive than, a mastectomy because the subsequent radiotherapy is very unpleasant, can be disfiguring, and can lead to increased mortality because of damage to the heart and vessels.¹⁶

Does screening save lives?

FPR claim in their abstract that we should have concluded that mammography does not save lives. We have not, and it is not possible to prove the negative. In our first *Lancet* paper, we concluded that 'there is no reliable evidence that screening decreases breast cancer mortality'.¹⁷ In our second *Lancet* paper and in our *Cochrane Review*, we concluded that 'The currently available reliable evidence does not show a survival benefit of mass screening for breast cancer (and the evidence is inconclusive for breast cancer mortality)'.^{2,4}

Methodology of our systematic review

A third misleading statement in FPR's abstract is that our method was to simply discard positive studies as being of poor quality. Our decisions were not 'justified in turn by a literature review'.¹ We did the opposite^{2–4,17} as we based our quality assessment of the studies on commonly accepted criteria for systematic reviews.¹⁸ We had no opinion on the effect of screening when we started to review the literature. But found it worrying that when we had divided the studies in two quality groups, the effect estimate for the better studies was significantly different from, and less impressive than, that of the poorer studies.

Biased misclassification of cause of death

Bias in classification of cause of death cannot be avoided, not even with the use of blinded end-point committees.^{2–4,19–21} The interesting question is therefore not whether it exists, but how large it is.

FPR argue that 'If bias in classification of deaths exists for HIP or the Two-County trial, it is not large'.¹ However, the lead

The Nordic Cochrane Centre, Rigshospitalet, Dept 7112, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark. E-mail: p.c.gotzsche@cochrane.dk

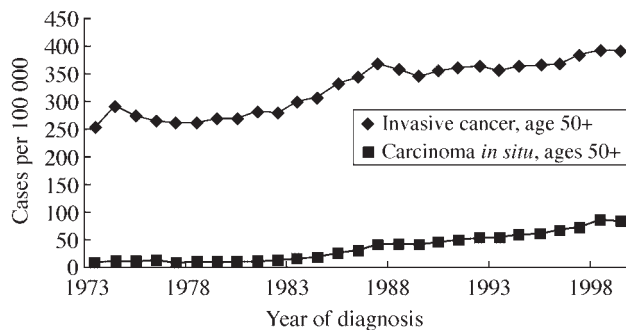
Table 1 Numbers of cancers, and tumourectomies and mastectomies in the screening trials from Canada (after 7 years) and Malmö (after 8.8 years)

	No. of women		No. of cancers		Relative risk (95% CI)	No. of tumourectomies and mastectomies		Relative risk (95% CI)
	Study group	Control group	Study group	Control group		Study group	Control group	
Malmö ⁵	21 088	21 195	588	447	1.32 (1.17, 1.49)	561	419	1.34 (1.18, 1.52)
Canada, 40–49 years ⁶	25 214	25 216	426	327	1.30 (1.13, 1.50)	415	313	1.33 (1.15, 1.53)
Canada, 50–59 years ⁶	19 711	19 694	460	365	1.26 (1.10, 1.44)	448	351	1.28 (1.11, 1.46)
Overall					1.30 (1.20, 1.40)			1.31 (1.22, 1.42)

Table 2 Numbers of cancers, and tumourectomies and mastectomies in the screening trials from Göteborg, Stockholm, and Two-County before the control group was screened. (Numbers for the control group in Stockholm were published as adjusted for different sample sizes in study and control groups; I recalculated the correct numbers)

	No. of women		No. of cancers		Relative risk (95% CI)	No. of tumourectomies and mastectomies		Relative risk (95% CI)
	Study group	Control group	Study group	Control group		Study group	Control group	
Göteborg, 39–49 years ⁷	11 724	14 217	144	151	1.13 (0.90, 1.41)	NA	NA	NA
Stockholm, 40–64 years ⁸	40 318	19 943	428	142	1.49 (1.23, 1.80)	360	120	1.48 (1.21, 1.82)
Two-County, 40–74 years ⁹	77 080	55 985	1378	752	1.33 (1.22, 1.45)	NA	NA	NA
Overall					1.33 (1.24, 1.44)			

NA: not available.

**Figure 1** Incidence rates of breast cancer in USA (SEER data, nine areas)¹⁵

investigator of the Two-County study has provided the most clearcut example that it can be very large.²² FPR fail to mention this although it is described in the papers they quote.^{22–24} The 2002 overview of the Swedish trials²³ reported only a 10% reduction in breast cancer mortality for the Östergötland part of the Two-County study (data were not made available for the Kopparberg part), whereas the lead investigator of the Two-County study in 2000 reported a 24% reduction (Table 3), using the same approach.²² Despite the fact that the follow-up was slightly longer, the trial authors reported 10 fewer deaths from breast cancer in the study group than in the overview and 23 more in the control group.

The likely explanation for this remarkable discrepancy is that the cause of death was assessed openly in the Two-County

study while the Swedish cause of death register was used for the overview. The trial authors' attempts at an explanation of the discrepancy have not been helpful,^{25,26} e.g. they say that they 'prefer to use the original primary research material where available, rather than figures reported in secondary research',²⁶ but the lead investigator has previously co-authored publications with the other overview trialists that included data from the Two-County study that originated from the same register as in the 2002 overview,^{27–29} i.e. 'secondary research', or from a blinded endpoint committee³⁰ which gave very similar results as when the cause of death register was used.²⁸

The New York HIP trial

FPR report in their table 1 that after 5 years, there were 39 breast cancer deaths in the study group and 63 in the control group, i.e. a 38% reduction in breast cancer mortality.¹ Of the 39 deaths in the study group, 16 were among those who were offered screening but declined. FPR 'estimate that the control group includes 16 women who would have refused screening and who died of breast cancer' and subtract 16 deaths, both from the study group and the control group, and then get a 51% reduction in breast cancer mortality $((39-16)/(63-16) = 0.49)$. However, when the relative risk is not exactly one (i.e. no effect), exclusion of the same number of deaths from both groups leads to biased estimates. At its extreme, one could remove 39 deaths from both groups, and the estimate would then be $(39-39)/(63-39) = 0$, corresponding to a reduction in breast cancer mortality of 100%. Death estimates should relate to all deaths.

Table 3 Breast-cancer mortality for the Östergötland part of the Two-County study as reported in a recent overview of the Swedish mammography screening trials and a recent update of the Two-County study (age group 40–74 years, evaluation model)

	Breast cancer deaths		Person-years of follow-up (in 000s)	Relative risk
	Study group	Control group		
Swedish overview ²³	177	190	1161	0.90
Two-County study ²²	167	213	1304	0.76

It is also worth noting that women who refuse to be screened have a worse prognosis than those who accept,^{5,31,32} presumably because some of them are afraid of having a suspicion of breast cancer confirmed.^{31,33} They also have a much higher death rate from all causes, excluding breast cancer, e.g. 93 per 10 000 person years compared with only 69 in the control group in the HIP study (ref. 34, p. 37).

As FPR note, numbers reported for this trial are not consistent, not even numbers of randomized women. We calculated that 517 more women with prior breast cancer were excluded after randomization from the study group than from the control group, FPR found 434. We believe our numbers are the correct ones since the study consists of matched pairs and the largest published exact number of women invited is 31 092^{2,4} (FPR refer to only 30 131). Other numbers have also been reported, e.g. after 10 years the difference was 526.³⁵

Retrospective exclusion of women is a bias-prone process. The PDQ panel of the National Cancer Institute in USA notes about the HIP study:³⁶

...exclusions were determined differently within the two groups... By design, controls did not have regular clinic visits and so the prestudy cancer status of control patients was not determined. When a control patient died and her cause of death was determined to be breast cancer, a retrospective examination was made to determine the date of diagnosis of her disease. If this was prior to the study period then she was excluded from the analysis. This difference in methodology has the potential for a substantial bias in comparing breast cancer mortality between the two groups, and this bias is likely to favor screening.

There are two reasons why the HIP trial failed to document the inevitable overdiagnosis. First, many more women with breast cancer prior to randomization were excluded from the study group than from the control group,^{2,4} and the lead investigator 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.³⁷ This speaks against the assertion by FPR, based on another paper by the lead investigator, that 'ascertainment was nearly perfect'.^{1,34} Second, the mammographic technique used at the time was very poor. FPR note that in the 1960s, ductal carcinoma *in situ* accounted for only 5% of breast cancers in the study group. A pathological review of the HIP study found that only 15% of 299 cancers in the study group were detected solely by mammography and that mammography did not identify a single

case of minimal breast cancer (<1 cm).³⁸ FPR note that the lead time in the HIP study was about one year and explains this as 'screening picks up a cancer roughly a year before it would become clinically manifest'.¹ However, very few cancers were detected at screening, and in the reference they quote³⁴ there is an estimate of the much more relevant 'program lead time' that represents the average advancement of diagnosis in the entire study group compared with the entire control group; this estimate is only 3–4 months. (ref. 34, p. 46) These facts agree very poorly with the large reported reduction in breast cancer mortality, 35% after 7 years.^{2,4}

FPR are aware of the problems but fail to draw the logical consequence of their observations—that it is highly unlikely that the New York trial found a true reduction in breast cancer mortality.

Baseline imbalances

FPR say that we misunderstood the sample sizes upon which our calculations of baseline imbalances were based. The papers are very confusing, however. In one, that FPR also quote, the text describes all randomized women and refers to a table that shows baseline differences as percentages but does not give the numbers upon which the percentages are based. The table has footnotes that say that some of the data are based on 10% and 20% samples.³⁹ 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. If the table header is correct, the data presented are subgroups of a subgroup, in which case FPR are correct that the resulting samples are too small to study possible baseline differences.

FPR make an error when they calculate that 11.7% of the examined women had a lump, compared with 11.8% in the control group and interpret this as indicating that the difference we described was due to chance. The 11.7% comes from women who attended screening,³⁹ whereas 9.5% is the percentage for the whole study group^{34,39} which is the correct fraction to use if one wants to check for possible baseline imbalance in a comparison with the control group. Since none of the many papers on the HIP trial give baseline characteristics in absolute numbers for the study and the control groups, it is not possible to check for baseline imbalances.

The randomization method is complicated, there was no concealment of the allocation, and it has not been documented that it went well. Because of this and because of the differential exclusions, the data published for the HIP study are not reliable. As explained above, they are not plausible either.

Assessment of cause of death

FPR say we are wrong as we had not included the HIP trial among those that use masked assessment of cause of death in our first *Lancet* paper.¹⁷ However, we clarified in our *Cochrane Review*^{2,4} that cause of death assessments were unblinded for 72% of the women with breast cancer.³⁴

For women with no known breast cancer history, subsamples were selected. For women with known breast cancer, a subsample was collected of women whose death certificates unequivocally stated that breast cancer was the underlying cause of death. Since a blinded review of medical and hospital

records confirmed the death certificates, no more women were selected for blinded review. In none of the cases did the investigators give any numbers, or note whether the sampling was random, and it is therefore not clear whether the procedures were adequate.

According to established rules, a subsample (28%) of dubious cases of death among breast cancer cases were selected for blinded review with two observers.³⁴ If there had been a true effect of screening, with a 24% reduction of breast cancer mortality after 18 years as reported among the cases potentially eligible for the review,⁴⁰ one would have expected 24% fewer dubious cases to be selected for review from the study group than from the control group. This was not so as 71 versus 73 cases were selected.³⁴ We therefore wrote that the review appears to be biased since more deaths were classified as caused by breast cancer in the control group than in the screened group.^{2,4} Based on the classification in the control group, there were 21 fewer cases than expected which were classified as breast cancer deaths in the study group ($P = 0.0003$).^{2,4} FPR find our calculation 'a little hazy' although they used the same numbers as we did, 13 of 71 versus 35 of 73, which gives $P = 0.0003$ with Fisher's exact test.

Thus, FPR are wrong when they say we have overlooked the possibility that the reason for this discrepancy could be that ambiguous deaths in the study group were less likely to be breast cancers because screening helps prevent death from this disease.

Length bias and lead-time bias

FPR compare total mortality *among breast cancer cases* in the study group and the control group and find a significant difference. They consider their comparison fair since numbers of incident cases in the two arms have equalized.

This is a gross error⁴¹ that, regrettably, is very common in the screening literature. First, as explained above, numbers should not have equalized. Second, and more important, screening predominantly identifies slow-growing tumours (length bias), including carcinoma *in situ* which most often does not progress, and which is not detected clinically. Breast cancers diagnosed in a group invited for screening are therefore different from those identified in a control group and they have a much better prognosis. This bias is so large that it can explain breast cancer mortality reductions of 50% or even more.⁴¹ Lead-time bias also invalidates such comparisons, although probably to a lesser extent.⁴¹ Total mortality can only be compared reliably if all randomized women are included in the analysis.

All-cause mortality

FPR dismiss our finding that all-cause mortality is the only reliable mortality endpoint for screening trials with the argument that it is impractical.

It need not be impractical. We calculated that to detect a decrease in breast cancer mortality of 30%, with type I and type II errors of 5% and 20%, respectively, a trial with 1.2 million women in each arm would suffice.^{2,4} Such a trial would be feasible in countries that have not introduced screening, and a trial in women at higher risk could be smaller.

There are good reasons for using all-cause mortality.²⁻⁴ Because of overdiagnosis, for example, screening leads to increased use of

radiotherapy,³ and radiotherapy can be predicted to increase all-cause mortality in women at low risk, such as those identified by screening¹⁶ because of an excess of cardiovascular deaths. It has been claimed that modern radiotherapy does not have these adverse effects,⁴² but the study had too little power and too short follow-up to exclude this possibility (11 versus 11 vascular deaths).¹⁶

Two-County study

Much of the information that has been published about this study is contradictory,^{2,4} and it is therefore not possible to know what really happened. FPR note, as they did for the HIP study, that the incidence rate for breast cancer is the same in the study group as in the control group, but that is only because the whole control group was screened before the trial closed. Before the control group screen, the incidence was 33% higher in the study group⁹ (Table 2).

FPR erroneously note that the reduction in breast cancer death rates is 60% in Kopparberg and 75% in Östergötland,¹ but what they calculated are the relative death rates. The reduction in breast cancer death rates were 40% and 25%, respectively.

Baseline imbalances

FPR accuse us of having said that 'cluster randomisation is biased'¹ in our first *Lancet* paper.¹⁷ This is wrong. We wrote about the Edinburgh trial that:

the screening and control groups were very different at baseline; only 26% of the women in the control group belonged to the highest socio-economic group, compared with 53% in the screening group

and concluded that 'the randomisation method was grossly inadequate, even for a cluster analysis'.¹⁷ We wrote about the Two-County study that:

it is unlikely that the meta-analysis took the clustering into account, as we arrived at the same point estimate and the same narrow confidence interval for breast cancer mortality as in the meta-analysis when we based our analysis on individual women. We therefore used women as the statistical unit

(for calculation of possible baseline imbalances). We were aware, of course, that this was not ideal, but that was the only option we had.

Since there were gross imbalances at baseline in the Edinburgh trial, that had 87 clusters, and there were only 45 clusters in the Two-County study, one would also expect imbalances in this study. FPR note that there was near-equality of breast cancer incidence rates before the study began. But they fail to note that in one of their own references, I addressed this question: 'the power of their test was very low, and cannot compare socioeconomic factors'.⁴³

FPR also note that 'death rates from other causes [among all randomised women] are nearly equal'.¹ In fact, there was a 2% higher mortality from other causes in the study group than in the control group in the first reference FPR quote. (Ref. 44, fig. 4)

Second, since breast cancer deaths constitute a small fraction of all deaths, such a result cannot prove anything about possible baseline imbalances. Third, mortality estimates several years after randomization could be influenced by the interventions under study (that is the very idea of doing a study!) and they therefore cannot be used for judgements of baseline imbalances. Thus, the reasoning of FPR is faulty. What matters is that—apart from age where we found baseline imbalances with our, admittedly, primitive analysis method—baseline data have not been published.

Numbers of women and deaths

FPR are correct that one reason for the changing numbers of randomized women is that women with a prior diagnosis of breast cancer were excluded after record linkage to the Swedish Cancer Registry in 1989. This can explain the two examples of discrepant numbers we gave in our first *Lancet* paper,¹⁷ but there are many other discrepancies^{23,30,44–53} that have not been—and cannot be—explained this way, since several of the papers are newer than 1989 (Table 4).

FPR's account of the discrepancies we found in number of breast cancer deaths^{2,4} is selective and misleading. They say that one of the differences in numbers can be explained by longer follow-up. We were aware of this but our point was that *fewer* deaths were reported with *longer* follow-up (see Table 5 for discrepancies for the Kopparberg part of the Two-County study).^{48,51,53–56} The lead investigator of the Two-County study co-authored all of these papers, including the overview³⁰ on which one of the discrepancies was based.⁵¹ Others have also commented on the discrepancies,⁵⁷ but FPR ignore the problems by ignoring the other discrepancies we reported.^{2,4}

Assessment of cause of death

Contrary to what FPR assert, we did make a case for differential bias in death classification,^{2,4} and, as shown above, the lead investigator of the Two-County study, Tabar, has provided the most convincing evidence of such bias. Tabar, Smith, and Duffy wrote in response to the updated overview of the Swedish trials that:

It is asserted in the overview report that the endpoint committees in the Two-County trial were aware of patients' study groups. No evidence is presented for this assertion.²⁵

This statement is astonishing. Nowhere in the protocol for the study⁵⁸ or in the many papers on the Two-County study has a blinded review of cause of death been described. And it is widely known among the Swedish screening trialists that cause of death assessment was not blinded. One of the key investigators involved with the Two-County study, Gunnar Fagerberg, confirmed in an interview that the investigators knew whether or not a woman had been screened when they assessed cause of death.⁵⁹

FPR repeat their erroneous claim that since death rates from other causes (among all those randomized) were similar, bias in death classification seems unlikely. As we have explained,^{2,4} one should look at other deaths among breast cancer cases, in particular deaths from other cancers, since the main problem of assessing cause of death concerns deaths from other cancers the women have in addition to their breast cancer.^{60–62} FPR were unable to replicate our calculation that showed significant bias,

based on deaths from other cancers among breast cancer cases, which is curious as we presented a meta-analysis with the actual data^{2,4} and as FPR mention the very same data (25 versus 6 cancer deaths).¹ FPR note that adjustment for time on risk would increase the *P*-value, but because of the overdiagnosis and the 'healthy screener effect' (see below) it is not clear whether such adjustment gives a less biased estimate than when the raw numbers are used.

If we assume that the results reported for the Two-County study are reliable, we may calculate what the expected all-cancer mortality (including breast cancer mortality) would be. In this study,⁴⁴ 243 out of 1993 cancer deaths were breast cancer deaths. If screening had had no effect, we would have expected 296 breast cancer deaths out of 2046 cancer deaths, i.e. 14.5%. Since relative risk for breast cancer mortality was 0.71 in the Two-County study as originally reported,⁴⁶ the expected reduction in all-cancer mortality becomes $14.5\% \times 29\% = 4.2\%$, i.e. a relative risk of 0.96. A weighted average that includes the other two trials that reported all-cancer mortality^{2–4} (those from Canada and Malmö), gives an expected relative risk of 0.95. However, relative risk for all-cancer mortality was 1.00 (95% CI: 0.91, 1.10) in the Two-County study and 1.02 (95% CI: 0.95, 1.10) in the other two trials combined.^{2–4} I wonder why FPR do not comment at all on our findings for all-cancer mortality,^{2–4} since these are important when discussing bias in assessment of cause of death.

FPR note that additional follow-up led to non-significant differences in deaths from other causes among breast cancer cases. We have not disputed this and we also found a non-significant difference.^{2,4} FPR also note that recent data have showed a significant reduction in total mortality among breast cancer cases. However, this comparison is severely flawed because of overdiagnosis. It has been shown, for example, that women with carcinoma *in situ* have the same or a better survival (100–104%) than women in the general population.⁶³ This could be called the 'healthy screener effect', i.e. those who accept the invitation to screening are healthier than other women, and carcinoma *in situ*—and some very slow-growing and therefore innocent tumours—are not detected without mammography. It is because of this inherent bias that it is worrying that the point estimates for other causes of death than breast cancer among breast cancer cases showed a worse, and not a better, survival.

Total mortality

FPR claim that screening has an impact on total mortality and quote the latest Swedish overview.²³ They provide a reference to my criticism of that overview⁴³ but do not discuss anything from it.

First, Nyström and colleagues did not claim that the 2% reduction in total mortality they found was statistically significant.²³ Second, they claimed that this reduction was what they would have expected (a 2.3% reduction), based on their findings for breast cancer mortality.²³ I showed that the reduction was bigger than expected as one would only have expected a 0.9% decrease.⁴³ Nyström and colleagues initially seemed to dispute my calculation,²⁴ but that was an error and an erratum acknowledged that my calculation is the correct one.⁶⁴ The inflated estimate of 2% was driven by the Östergötland part of the Two-County trial, which contributed half of the deaths.

Table 4 Discrepancies in reported numbers of women randomized in the Two-County study

Publication year	Age range	Study group	Control group
Kopparberg			
1985 ⁴⁵	40+	47 389	22 658
1985 ⁴⁶	40–74	39 051	18 846
1989 ⁴⁷	40–74	38 589	18 582
1993 ³⁰	40–74	38 562	18 478
1995 ⁴⁸	40–74	38 589	18 582
2000 ⁴⁹	40–74	38 568	18 479
2000 ⁵⁰	40–74	38 588	18 582
2002 ²³	40–74	NA	NA
1988 ⁴⁴	40–49	9625	5053
1993 ⁵¹	40–49	NA	NA
1995 ⁴⁸	40–49	9582	5031
1997 ^{52,53}	40–49	9650	5009
Östergötland			
1985 ⁴⁵	total	47 001	45 933
1985 ⁴⁶	40–74	39 034	37 936
1989 ⁴⁷	40–74	38 491	37 403
1993 ³⁰	40–74	38 405	37 145
1995 ⁴⁸	40–74	38 491	37 403
2000 ⁴⁹	40–74	38 942	37 675
2000 ⁵⁰	40–74	39 105	37 858
2002 ²³	40–74	38 942	37 675
1988 ⁴⁴	40–49	10 312	10 625
1993 ⁵¹	40–49	NA	NA
1995 ⁴⁸	40–49	10 262	10 573
1997 ^{52,53}	40–49	10 240	10 411

NA: data not available.

Table 5 Discrepancies in reported numbers of breast cancer deaths in the Kopparberg part of the Two-County study, age group 40–49 years

Publication year	Follow-up (years)	Follow-up (persons-years, in 000s)	Breast cancer deaths	
			Study group	Control group
1993 ^{51,54}	10.4	163	26	18
1995 ⁴⁸	12.5	183 ^a	22	16
1996 ⁵⁵	13.5 ^b	197	22	16
1997 ⁵³	14.9	219	23	18
1999 ⁵⁶	17.0 ^b	248	26	18

^a Data calculated based on the proportion of women aged 40–49.

^b Number of women from ref. 48.

It was surprising that the unadjusted and age-adjusted estimates for total mortality were the same, with relative risk 0.98 since these were 1.00 and 1.06 (95% CI: 1.04, 1.08), respectively, in the 1993 Swedish overview.³⁰ The number of person-years of follow-up was 2.6 million in 1993 and 3.5 million in 2002; the Kopparberg part of the Two-County trial was not available for the 2002 overview, but extended data of dubious quality⁴³ for the Malmö trial were. Nyström and colleagues have not explained how these changes could result in such a large difference from the 1993 overview, and they seem to have analysed the trials as if they were cohort studies, and have not explained how they dealt with those that were not fully randomized. Finally, there are notable, unexplained discrepancies in numbers⁴³ (Table 4).

We therefore do not know whether screening has an effect on total mortality.

Canadian trial

The only meta-analysis which used blinded assessment of the quality of the design of the trials⁶⁵ found that the trials from

Canada and Malmö scored highest (as we did), whereas the Two-County study scored lowest.

FPR seem to have needed considerable help from the investigators¹ to understand the Two-County study. This is not necessary for the Canadian trial as the authors have responded publicly and consistently to all the criticisms that have been raised.

As FPR mention, the trial authors have documented a high degree of observer variation for reading mammograms. The trial should not be blamed for this well-known phenomenon⁶⁶ just because it seems to have been the only one that provided data on it.⁶⁷ What is more important is that the cancers detected were considerably smaller than in the Two-County study⁶⁸ and that the level of overdiagnosis of cancers was equally large in this trial as in other trials (Tables 1 and 2). These facts do not suggest poor performance.

Baseline imbalances

As FPR note, the randomization method was not optimal, but we judged that the trial appears to have been adequately randomized. First, an independent review of ways in which the randomization could have been subverted uncovered no evidence of it.⁶⁹ Second, the validity of the randomization procedure is supported by the nearly identical size of the comparison groups (Table 1) and the comparability at baseline in age and nine other factors of potential prognostic importance.^{2,4}

There were more small cancers with four or more nodes involved in the study group than in the control group among women aged 40–49. FPR misrepresent our use of the expression 'post hoc', claiming that it cannot be *post hoc* when it was mentioned in the original trial report. But with *post hoc* we clearly meant post-randomization^{2,4} and explained why:

The finding of more small node-positive cancers in the screened group than in the control group in the Canadian trial cannot be used to judge the reliability of the randomisation process, because variables that become apparent as a result of the screening process are biased. Some women with four or more positive nodes were probably unrecognized in the control group, and some were not recognized when their cancers spread, and they were more likely to be treated in centres where careful extensive nodal dissection was not the norm.⁷⁰

FPR have ignored this explanation by the trial's lead investigator, and the additional explanation given in the report they quote, (ref. 71, p. 1473) when they say that the finding cannot be a result of the intervention since both groups were subjected to physical examination.

Concomitant physical examination

FPR misrepresent our work when they say that we have confused breast examination by a practitioner and self-examination. It is clear from their own text that we did not.¹ We explicitly noted that a study of self-examination had not found an effect on breast cancer mortality and added that 'any effect of physical examination is likely to be small'. FPR argue that physical examination is effective at cancer detection but as there are no randomized trials of physical detection,⁷² we do not know whether physical detection can lower breast cancer mortality, or even whether it helps detect cancers before they have spread.

Results of Canadian trial

The updated results for the young age group were 105 breast cancer deaths in the study group and 108 in the control group. FPR suggest that the effect will be about 20% after they have excluded some advanced cases from the study group. Such methods are not among accepted standards for randomized trials.

As just explained, the finding of more node-positive cancers in the study group than in the control group does not support the idea that 'something went awry in the randomization'.¹ And the same was seen in the HIP trial which FPR praise (57% versus 46% of cancers had positive nodes); (ref. 34, p. 5) surprisingly, this occurred *despite* the fact that many more women with breast cancer were excluded from this trial after randomization in the study group than in the control group.

Thus, a much more reasonable interpretation is that some of the invasive cancers grow so slowly that they were only detected in the mammography group and not in the control group,⁷³ i.e. what we are talking about is overdiagnosis of invasive cancers, as reported for the Canadian trial⁷³ and as also strongly suggested by the US data discussed above¹⁵ (Figure 1). The existence of considerable overdiagnosis of invasive cancers also agrees well with a study that showed that the majority of invasive cancers grow slowly, with a relative annual mortality rate of only 2.5%.⁷⁴ Thus, in the absence of screening, many of these invasive cancers would never manifest themselves and become disease, before the women died from other causes.⁷⁵

Conclusion: What are the effects of screening?

Like many screening advocates have done, FPR have consistently ignored evidence that goes against their beliefs. The critique by FPR of our systematic review of the screening trials is based on erroneous assumptions, faulty logic, biased statistical analyses, and selective and misleading quotations of our work and that of others. FPR are correct on only two points which are not important for the validity of our review.

The most positive result for the screening trials comes from a Swedish overview that in 1993 reported a 29% reduction in breast cancer mortality among women aged 50–69 years.³⁰ If we, for the sake of the example, accept this finding, it follows that after 1000 women have been invited to regular screening throughout 10 years, one of them (0.1%) may avoid dying from breast cancer.³⁰ Since overall mortality was 10%,²⁹ it means that invitation to screening—most optimistically—may increase survival from 90.2% to 90.3%. Thus, for every 1000 women invited for screening throughout 10 years, at most one—and possibly none, since an effect on overall survival has not been demonstrated^{2–4,43}—is saved; five additional women will be diagnosed with cancer who would not have got a cancer diagnosis had they not been screened; two additional women will get a mastectomy and three a tumourectomy; and more than 100 women will experience important psychological distress for many months because of false positive findings.^{3,4}

Therefore, even under the most optimistic survival estimate, it not clear whether screening does more good than harm. This dilemma deserves further discussion and honest information should be given to women about the uncertainties and harms of screening.⁷⁶

The views expressed in this article represent those of the author and are not necessarily the views or the official policy of the Cochrane Collaboration.

References

- Freedman DA, Petitti DB, Robins JM. On the efficacy of screening for breast cancer. *Int J Epidemiol* 2004;**33**:43–55.
- Olsen O, Gøtzsche PC. Cochrane review on screening for breast cancer with mammography. *Lancet* 2001;**358**:1340–42.
- Olsen O, Gøtzsche PC. Systematic review of screening for breast cancer with mammography. <http://image.thelancet.com/lancet/extra/fullreport.pdf>, 2001.
- Olsen O, Gøtzsche PC. Screening for breast cancer with mammography (Cochrane Review). In: *The Cochrane Library*, Issue 4. Oxford: Update Software, 2001.
- Andersson I, Aspegren K, Janzon L *et al.* Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. *BMJ* 1988;**297**:943–48.
- Miller AB. The costs and benefits of breast cancer screening. *Am J Prev Med* 1993;**9**:175–80.
- Bjurstam N, Bjorneld L, Duffy SW *et al.* The Gothenburg breast screening trial: first results on mortality, incidence, and mode of detection for women ages 39–49 years at randomization. *Cancer* 1997;**80**:2091–99.
- Frisell J. *Mammographic Screening for Breast Cancer* [thesis]. Stockholm: Södersjukhuset, 1989.
- Tabar L, Fagerberg CJG, South MC, Day NE, Duffy SW. The Swedish Two-county Trial of mammographic screening for breast cancer: recent results on mortality and tumour characteristics. In: Miller AB, Chamberlain J, Day NE *et al.* *Cancer Screening*. Cambridge: Cambridge University Press, 1991, pp. 23–36.
- Gøtzsche PC. Trends in breast-conserving surgery in the Southeast Netherlands: Comment on article by Ernst and colleagues *Eur J Cancer* 2001;**37**:2435–40. *Eur J Cancer* 2002;**38**:1288.
- An audit of screen detected breast cancers for the year of screening April 1999 to March 2000. West Midlands NHS Breast and Cervical Screening Quality Assurance Reference Centre, 2001. (BASO Breast Audit 1999/2000, also available from <http://www.cancerscreening.nhs.uk/breastscreen/publications.html>).
- Paci E, Duffy SW, Giorgi D *et al.* Are breast cancer screening programmes increasing rates of mastectomy? Observational study. *BMJ* 2002;**325**:418.
- Gøtzsche PC. Misleading paper on mastectomy rates in a screening programme. <http://bmj.com/cgi/eletters/325/7361/418#24972>, 26 Aug 2002.
- Fletcher SW, Elmore JG. Clinical practice. Mammographic screening for breast cancer. *N Engl J Med* 2003;**348**:1672–80.
- 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 (Accessed 26 June 2003).
- Early Breast Cancer Trialists' Collaborative Group. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: An overview of the randomised trials. *Lancet* 2000;**355**:1757–70.
- Gøtzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000;**355**:129–34.
- Clarke M, Oxman AD (eds). *Cochrane Reviewers' Handbook 4.2.0* [updated March 2003]. <http://www.cochrane.dk/cochrane/handbook/handbook.htm> (Accessed 24 June 2003).
- Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and surgery in early breast cancer: An overview of the randomized trials. *N Engl J Med* 1995;**333**:1444–55.

- 20 Gøtzsche PC. Screening for breast cancer with mammography (author's reply). *Lancet* 2001;**358**:2167–68.
- 21 Brown BW, Brauner C, Minnotte MC. Noncancer deaths in white adult cancer patients. *J Natl Cancer Inst* 1993;**85**:979–87.
- 22 Tabar L, Vitak B, Chen HH *et al*. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am* 2000;**38**:625–51.
- 23 Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist L. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. *Lancet* 2002;**359**: 909–19.
- 24 Nyström L, Andersson I, Bjurstam N, Frisell J, Rutqvist LE. Update on effects of screening mammography (authors' reply). *Lancet* 2002;**360**:339–40.
- 25 Tabár L, Smith RA, Duffy SW. Update on effects of screening mammography. *Lancet* 2002;**360**:337.
- 26 Duffy SW, Tabár L, Smith RA. The mammographic screening trials: commentary on the recent work by Olsen and Gøtzsche (authors' reply). *J Surg Oncol* 2002;**81**:164–66.
- 27 Larsson LG, Andersson I, Bjurstam N *et al*. Updated overview of the Swedish randomized trials on breast cancer screening with mammography: age group 40–49 at randomization. *J Natl Cancer Inst Monogr* 1997;**22**:57–61.
- 28 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 Screen* 1996;**3**:129–32.
- 29 Nyström L, Larsson LG, Wall S *et al*. An overview of the Swedish randomised mammography trials: total mortality pattern and the representivity of the study cohorts. *J Med Screen* 1996;**3**:85–87.
- 30 Nyström L, Rutqvist LE, Wall S *et al*. Breast cancer screening with mammography: overview of Swedish randomised trials. *Lancet* 1993;**341**:973–78.
- 31 First results on mortality reduction in the UK Trial of Early Detection of Breast Cancer. UK Trial of Early Detection of Breast Cancer Group. *Lancet* 1988;**ii**:411–16.
- 32 Tabar L, Fagerberg G, Chen HH, Duffy SW, Gad A. Screening for breast cancer in women aged under 50: mode of detection, incidence, fatality, and histology. *J Med Screen* 1995;**2**:94–98.
- 33 Andersson I, Sigfusson BF. Screening for breast cancer in Malmö: a randomized trial. *Recent Results Cancer Res* 1987;**105**:62–66.
- 34 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 University Press, 1988.
- 35 Aron J, Prorok PC. An analysis of the mortality effect in a breast cancer screening study. *Int J Epidemiol* 1986;**15**:36–43.
- 36 PDQ Panel of the National Cancer Institute. Screening for breast cancer: mammography. <http://www.cancer.gov/cancerinfo/pdq/screening/breast/healthprofessional/#17> (Accessed 16 June, 2003).
- 37 Shapiro S. Discussion II. *Natl Cancer Inst Monogr* 1985;**67**:75.
- 38 Thomas LB, Ackerman LV, McDivitt RW, Hanson TAS, Hankey BF, Prorok PC. Report of NCI ad hoc pathology working group to review the gross and microscopic findings of breast cancer cases in the HIP study. *J Natl Cancer Inst* 1977;**59**:496–541.
- 39 Shapiro S, Strax P, Venet L, Venet W. Changes in 5-year breast cancer mortality in a breast cancer screening program. *Proc Natl Cancer Conf* 1972;**7**:663–78.
- 40 Shapiro S. Periodic screening for breast cancer: the HIP randomized controlled trial. *J Natl Cancer Inst Monogr* 1997;**22**:27–30.
- 41 Berry DA. The Utility of Mammography for Women 40 to 50 Years of Age (Con). In: DeVita VT, Hellman S, Rosenberg SA (eds). *Progress in Oncology*. Sudbury: Jones and Bartlett, 2002, pp. 346–72.
- 42 Hojris I, Overgaard M, Christensen JJ, Overgaard J. Morbidity and mortality of ischaemic heart disease in high-risk breast-cancer patients after adjuvant postmastectomy systemic treatment with or without radiotherapy: analysis of DBCG 82b and 82c randomised trials. Radiotherapy Committee of the Danish Breast Cancer Cooperative Group. *Lancet* 1999;**354**:1425–30.
- 43 Gøtzsche PC. Update on effects of screening mammography. *Lancet* 2002;**360**:338.
- 44 Tabar L, Fagerberg CJG, Day NE. The results of periodic one-view mammographic screening in Sweden. Part 2: Evaluation of the results. In: Day NE, Miller AB (eds). *Screening for Breast Cancer*. Toronto: Hans Huber, 1988, pp. 39–44.
- 45 Socialstyrelsens beredningsgrupp för WE-projektet. Minskad mortalitet i bröstcancer genom hälsokontroll med mammografi. *Nord Med* 1985;**100**:175–78.
- 46 Tabar 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.
- 47 Tabar 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.
- 48 Tabar 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.
- 49 Nyström L. *Assessment of Population Screening: the Case of Mammography* [thesis]. Umeå: Umeå University Medical Dissertations, 2000.
- 50 Nixon R, Prevost TC, Duffy SW, Tabar 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.
- 51 Nyström L, Larsson L-G. Breast cancer screening with mammography [reply]. *Lancet* 1993;**341**:1531–32.
- 52 Nyström L, Wall S, Rutqvist LE *et al*. Update of the overview of the Swedish randomized trials on breast cancer screening with mammography. NIH Consensus Development Conference on Breast Cancer Screening for Women Ages 40–49. National Institutes of Health 1997, pp. 65–69.
- 53 Larsson LG, Andersson I, Bjurstam N *et al*. Updated overview of the Swedish Randomized Trials on Breast Cancer Screening with Mammography: age group 40–49 at randomization. *J Natl Cancer Inst Monogr* 1997;**22**:57–61.
- 54 Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. *J Natl Cancer Inst* 1993;**85**:1644–56.
- 55 Tabar L, Duffy SW, Chen HH. Quantitative interpretation of age-specific mortality reductions from the Swedish Breast Cancer-Screening Trials [letter]. *J Natl Cancer Inst* 1996;**88**:52–55.
- 56 Tabar 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.
- 57 de Koning HJ, Warmerdam PG, Beemsterboer PMM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish Breast Cancer-Screening Trials [reply]. *J Natl Cancer Inst* 1996;**88**:54–55.
- 58 Rapport över mammografiscreening i Kopparbergs och Östergötlands läns landsting (WE-projektet)—Resultat efter första screeningsomgången. Stockholm: Socialstyrelsen, 1982.
- 59 Crewdson J. Swedes doubt mammography trial: disparities found in landmark study. Chicago Tribune 2002;March 15 <http://www.chicagotribune.com/news/chi-0203150264mar15.story> (Accessed 15 March 2002).
- 60 Miller AB. Screening for breast cancer with mammography. *Lancet* 2001;**358**:2164.

- ⁶¹ Andersson I, Janzon L. Mammografi för screening—kritisk inställning stöds av nya fynd [Screening with mammography—a critical attitude is supported by new findings]. *Läkartidningen* 1988;**85**: 3666–69.
- ⁶² Janzon L, Andersson I. The Malmö mammographic screening trial. In: Miller AB, Chamberlain J, Day NE *et al.* (eds). *Cancer Screening*. Cambridge: Cambridge University Press, 1991, pp. 37–44.
- ⁶³ Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma *in situ* of the breast. *JAMA* 1996;**275**:913–18.
- ⁶⁴ Department of error: update on screening mammography. *Lancet* 2002;**360**:1178.
- ⁶⁵ Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50. A quality assessment and meta-analysis. *Med J Aust* 1995;**162**:625–29.
- ⁶⁶ Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Arch Intern Med* 1996;**156**:209–13.
- ⁶⁷ Baines CJ, McFarlane DV, Miller AB. The role of the reference radiologist. Estimates of inter-observer agreement and potential delay in cancer detection in the national breast screening study. *Invest Radiol* 1990;**25**:971–76.
- ⁶⁸ Narod SA. On being the right size: A reappraisal of mammography trials in Canada and Sweden. *Lancet* 1997;**349**:1849.
- ⁶⁹ 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.
- ⁷⁰ Miller AB. The Canadian National Breast Screening Study: update on breast cancer mortality. In: *NIH Consensus Development Conference on Breast Cancer Screening for Women ages 40–49*. 1997, pp. 51–53.
- ⁷¹ 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.
- ⁷² Kösters JP, Gøtzsche PC. Regular self-examination or clinical examination for early detection of breast cancer (Cochrane Review). In: *The Cochrane Library*, Issue 2, 2003. Oxford: Update Software.
- ⁷³ 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.
- ⁷⁴ Fox MS. On the diagnosis and treatment of breast cancer. *JAMA* 1979;**241**:489–94.
- ⁷⁵ Spratt JS, Meyer JS, Spratt JA. Rates of growth of human neoplasms: Part II. *J Surg Oncol* 1996;**61**:68–83.
- ⁷⁶ Jørgensen KJ, Gøtzsche PC. Presentation on websites of possible benefits and harms from screening for breast cancer: cross sectional study. *BMJ* 2004;**328**:148–51.

Commentary: A defence of the Health Insurance Plan (HIP) study and the Canadian National Breast Screening Study (CNBSS)

Anthony B Miller

The commentary of Freedman *et al.*¹ on the reviews by Gøtzsche and Olsen^{2,3} focuses largely on three of the screening trials, and they conclude, like the International Agency for Research on Cancer (IARC) working group that reviewed all the trials,⁴ that mammography screening does save lives.

I agree with their comments on the Health Insurance Plan (HIP) trial. I drew very similar conclusions when the first review of Gøtzsche and Olsen was published.⁵

Having been a participant in the IARC working group that reached similar conclusions to Freedman *et al.* on the Two County trial, and having found the analysis of Nixon *et al.*⁶ particularly compelling in largely dealing with the cluster randomization issue, I also agree with most of their comments

Department of Public Health Sciences, University of Toronto, Canada.

on that trial, though I still have some caveats on its application at the present time. However, Freedman *et al.* cite the analysis of Nystrom *et al.*⁷ as demonstrating equivalence in breast cancer incidence prior to randomization. They neglect to mention that Nystrom *et al.*⁷ were only able to assess this in regard to Östergötland, as Tabar declined to produce the data for the Kopparberg component of the trial for this overview analysis. Thus we still do not have absolute certainty that the clusters in Kopparberg were balanced.

More important, it is not clear that either the HIP or the Two County trials are relevant to the present time, when women with stage 2 breast cancer invariably receive adjuvant chemotherapy or hormone therapy, not available at the time of HIP, and apparently not given in the Two Counties in Sweden when that trial was conducted.^{8,9} The availability of such therapy