COMMENTARY

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

Donald A. Berry

Few medical issues have generated as much controversy as the question of whether or not to recommend regular mammography to women aged 40–49 years *(1,2)*. In January 1997, a Consensus Development Panel[1] at the National Institutes of Health Consensus Conference on Breast Cancer Screening for Women Ages 40–49 recommended that women be informed of the benefits and risks and decide for themselves whether or not screening is appropriate *(3)*. The National Cancer Institute (NCI) took a different view, altering its previous stance *(4)* and joining the American Cancer Society (ACS) in recommending regular screening for women in their forties. The NCI recommends mammograms every 1–2 years, and the ACS recommends annual mammograms.

Recent meta-analyses of mortality data from randomized trials are available *(5,6)*. However, existing analyses do not adequately address some important issues. This commentary provides new analyses and evaluations of the evidence. In addition, it addresses the question of communicating the potential benefits and risks of screening to women and their caregivers.

Deciding whether or not to be screened for a disease is complicated. If the disease is curable at every clinical stage, then screening has no survival benefit. At the opposite extreme, if available therapy is ineffective, then again screening has no benefit—although it may be appropriate for people who ''just want to know'' at the earliest possible time. Depending on the disease, its treatment, and methods of detection, there may be a window of opportunity between these extremes. Screening is beneficial when early detection and subsequent treatment during a disease's preclinical phase delays its progress beyond that achievable in its clinical phase.

The presence of a number of well-known biases [including ''lead-time bias'' and ''length bias'' *(5)*] make it difficult to ascertain the benefits of screening. Observations made from databases and certain types of observations from randomized trials are susceptible to such biases. For example, a number of researchers at the January 1997 consensus development conference presented data demonstrating that cancers detected by mammography have much better prognoses than do cancers detected otherwise. This observation is expected and is obviously correct, but it is irrelevant in assessing mammography's benefits. Mammography is effective in finding breast cancer, and it finds it earlier than is possible using other methods. Indeed, it probably finds some lesions that would never manifest themselves clinically. The relevant question is whether finding and treating breast cancer at an earlier time point extends women's lives or improves the quality of their lives. Observational databases cannot address this question (except by mathematical modeling of the disease's course, which requires making unverified assumptions). Randomized controlled trials that have breast cancer mortality as the end point enable direct assessments of mammography's benefits, and evidence from such trials is the only evidence of effectiveness that I will consider in this commentary.

I have two principal objectives. One is to present a meta-analysis of the randomized trials that allows for differences between the trials and that addresses an individual woman's decision more appropriately than do the existing analyses. A related matter is the trial differences themselves and some questionable aspects of the trial results and the analyses used by the investigators. My other principal objective is to provide ways of presenting the benefits and risks of mammography that will help women and their caregivers in making screening decisions. My focus is the decision faced by an individual woman. Establishing a national health policy concerning screening mammography is an important matter, and it is related to an individual's decision-making process, but I do not address it here.

Randomized Trials of Screening Mammography

The eight randomized trials listed in Table 1 have addressed the benefits of screening mammography. (Some researchers count the Kopparberg and the Östergötland trials as a single ''two-county'' trial, in which case there would be only seven trials.) In these trials, women were assigned in a random fashion to either a regular screening program involving mammography or to a control program not involving mammography (i.e., ''usual care'' in most trials). All participants were informed about breast self-examination, although the methods of informing them varied. Even though some women in the control groups experienced a variety of screening measures, I use ''screening group'' to refer to those women assigned to a regular schedule of mammograms.

The only North American trials were the Health Insurance Plan (HIP) of New York *(6)* and the Canadian National Breast Screening Study I *(7)*. Five trials were conducted in Sweden *(8–11)*, and the other took place in Edinburgh, U.K. *(12)*. The Canadian trial restricted accrual to women who were aged 40–49 years at entry, while women in this age group constituted a subset of a larger cohort in the other trials. The Canadian trial

Correspondence to: D. A. Berry, Ph.D., Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251.

See ''Notes'' following ''References.''

Table 1. Characteristics of the eight randomized trials of screening mammography *(6–12)*: characteristics of the trials

Trial	Entry age, y*	Accrual period	Follow-up, y	Type of randomization	No. of mammograms	Interval between mammograms, mo	No. of views	Analysis type†
HIP‡	40–49	1963–1966	18.0	Individual	4	12	2	Follow-up
Kopparberg	40–49	1977–1989	15.2	Cluster	4	24	1	Evaluation
Östergötland	40–49	1978–1989	14.2	Cluster	4	24	1	Evaluation
Malmö	45–49	1977–1990	12.7	Individual	5	18, 24	1, 2	Follow-up
Edinburgh	45–49	1978–1985	12.6	Cluster	4	24	1, 2	Follow-up
Gothenburg	39–49	1982–1984	12.0	Cluster/individual§	5	18	1, 2	Evaluation
Stockholm	40–49	1981–1986	11.4	Cluster	2	28	1	Evaluation
Canada NBSSI	40–49	1980–1985	10.5	Individual	4, 5	12	2	Follow-up

*Only Canada NBSSI (National Breast Screening Study I) considered the indicated age range prospectively. In the other trails, women in this age group constituted a post-hoc subset.

†A follow-up analysis considers all deaths from breast cancer detected at any time after study entry. An evaluation analysis does not consider deaths from breast cancer only if the cancers are detected after the study screening period.

‡HIP = Health Insurance Plan of New York.

§There was shift in the randomization procedure part way through the trial; 82% of the women were randomly assigned individually.

included volunteers, and the other trials were population based. Table 1 indicates that the Malmö and Edinburgh trials did not include women in their early forties. The most mature trial is HIP, with average reported follow-up of 18 years. Depending on the trial, the unit of randomization was either individual or cluster. In the former, women were assigned to screening one at a time, for example, according to whether their date of birth was even or odd. In cluster randomization, all women who visited a particular clinic or practice were assigned to the same group, either screening or control. The proposed number of mammographic screens, the target interval between screens, and the number of mammographic views are shown in Table 1. The last column of Table 1 indicates the type of analysis used. These analyses and related biases will be described later in this commentary.

A major issue in randomized studies—including screening trials—is lack of compliance with the study protocol. While the rates of noncompliance with group assignment varied greatly across the trials listed in Table 1, 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. Women at lower risk of breast cancer may have skipped mammograms, and women at higher risk may have sought them out. In any case, reliable information concerning which participants sought mammograms and for what reasons is not available.

Table 2 shows the numbers of deaths attributed to breast cancer and the total numbers of life-years by group *(13)*. Life-years accumulate from the time of randomization into a trial. Fig. 1, A, is a plot of the breast cancer mortality rates for the control and screening groups. Control group mortality rates tended to be higher than screening group mortality rates. Control group rates were more variable than screening group rates *(14)*. The penultimate column in Table 2 shows the estimated breast cancer mortality reduction resulting from assignment to screening. [Adjusted mortality reduction, shown in the last column of Table 2, will be explained presently.] A negative reduction indicates that the breast cancer mortality rate in the screening group was greater than the breast cancer mortality rate in the control group. Fig. 1, B, is a plot of the estimated reductions. Five trials showed a positive reduction with screening, and three

Table 2. Results of the eight randomized trials of screening mammography *(6–12)*: breast cancer mortality rates, restricted to women aged 40–49 years at time of accrual*

	Screening group			Control group			Mortality reduction, %	
Trial	Breast cancer deaths	Life-years	Mortality rate†	Breast cancer deaths	Life-years	Mortality rate†	Observed‡	Adjusted§
HIP	49	248 000	19.8	65	253 000	25.7	23	22
Kopparberg	23	144 000	16.0	18	75 000	24.0	33	22
Östergötland	27	143 000	18.9	27	147 000	18.4	−3	8
Malmö	57	166 000	34.3	78	144 000	54.2	38	28
Edinburgh	46	146 000	31.5	52	135 000	38.5	18	18
Gothenburg	18	138 000	13.0	39‖	168 000	23.2	44	29
Stockholm	24	174 000	13.8	12	88 000	13.6	−1	5
Canada	82	283 000	29.0	72	283 000	25.4	−14	0
Total	326	1 442 000	22.6	363	1 293 000	28.1	18	18

*Numbers of deaths and life-years are from Hendrick et al. *(13)*.

†Breast cancer mortality rate = deaths per 100 000 life-years.

‡Observed percent breast cancer mortality reduction is 100 (1-relative risk), where relative risk is the ratio of mortality rates, screening to control. For ''Total,'' relative risk is calculated using the Mantel–Haenszel statistic.

§Adjusted breast cancer mortality reduction, as described in the text.

‖Bjurstam et al. *(10)* reported 40 deaths in the control group, with a mortality reduction of 45%. Using 39 deaths facilitates comparison with the conclusions of Hendrick et al. *(13)*, assuming 40 versus 39 has little impact on the conclusions.

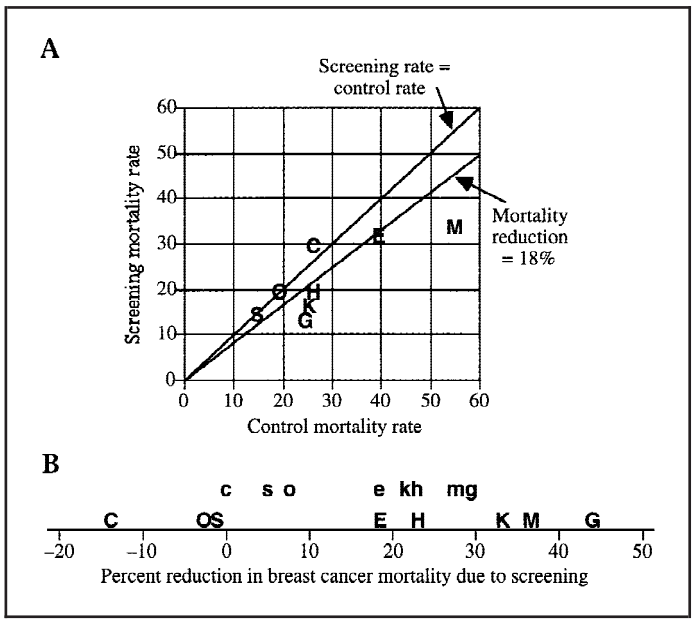

Fig. 1. Plots of results of the eight randomized trials of screening mammography *(6–12).* Data points are labeled by use of trial symbols as follows: H = Health Insurance Plan of New York; K = Kopparberg; O = Östergötland; M = Malmö; E = Edinburgh; G = Gothenburg; S = Stockholm; and C = Canada National Breast Screening Study I. **A**) Breast cancer mortality rates per 100 000 life-years by group. The upper line would be obtained if mortality rates in screening groups and control groups were the same. The lower line reflects the 18% reduction in mortality observed in the eight trials combined. **B**) Estimated reduction in breast cancer mortality—screening over control. Positive reductions indicate an observed benefit for the screening group. Upper-case symbols show the estimated reductions in the trials [*see* penultimate column of Table 2], and lower-case symbols show the adjusted or regressed estimates as described in the text and as indicated in the last column of Table 2. Trial results further from the overall mean estimate of 18% are regressed more than are those closer to the mean, and smaller trials are regressed more than larger trials.

trials showed an increase in breast cancer mortality in the screening group.

Statistical Analysis of the Randomized Trials

A standard statistical analysis of multiple screening trials involves the Mantel–Haenszel test and its associated confidence intervals *(15).* On the basis of data shown in Table 2, the Mantel–Haenszel 95% confidence interval about the estimated 18% reduction in breast cancer mortality attributed to screening is 5%–29% *(13).* Since 0 is not included in this interval, the results are regarded as statistically significant at the 5% level.

Some participants in the consensus development conference took statistical significance at the 5% level to be definitive and final. Such a view of significance testing is restricted. First, a statistically significant conclusion does not mean that screening is beneficial; furthermore, it does not imply that screening is probably beneficial *(16).* Second, decisions should not be based on arbitrary cutoffs of statistical significance—such as 5%—but should weigh the consequences of the decisions. Third, the confidence intervals should account for the many significance tests that have been undertaken (and will be undertaken in the future) on the data from these trials *(17,18);* exact adjustments are not possible because none of the analyses were planned in advance, but it is clear that the results would lose statistical significance if multiple testing were considered. Fourth, the Mantel–Haenszel test addresses the mean reduction in the population of trials, which is not directly relevant for a woman's decision regarding screening. A related issue is that the Mantel–Haenszel test assumes trial homogeneity—that the relative risk is the same in all eight trials. This assumption is questionable at best, as is evident in the plotted data (Fig. 1, B).

There are many obvious trial differences that make assuming homogeneity questionable. There are differences in populations, in quality of the mammographic techniques, in screening intervals, in numbers of screens, in types of control, in sizes of control tumors *(14),* in use of clinical breast examinations, in lengths of follow-up, in types of analysis, etc.

Assuming homogeneity leads to a fixed-effects model that, approximately, treats the results as having arisen from a single large trial. Allowing for heterogeneity—but not assuming it—leads to a random-effects model. A class of random-effects models is hierarchical *(16,19).* In the application at hand, there are two levels of experimental units, one is ''trial'' and the other is ''participant within trial.'' The eight trials are viewed as having been selected from a larger population. There are two parameters of interest. One is the average mortality reduction in the population of trials. The second is the mortality reduction in the next trial sampled from this population. (In a fixed-effects model, these two parameters are the same.) In a hierarchical Bayesian analysis *(16,19),* reduction in the next trial has a probability distribution, called a predictive distribution. This predictive distribution is directly relevant for an individual woman's decision problem. In particular, a candidate ''next trial'' is a woman in this age group who is faced with a screening decision. The benefit she will receive can be regarded as the next observation from the population of trial benefits.

The results of a Bayesian analysis using hierarchical modeling are presented below. This analysis is described in the Appendix. It treats the eight trials symmetrically, except that larger trials weigh somewhat more heavily than smaller ones. More detailed hierarchical analyses are possible. For example, to address the type of analysis employed in the trial (evaluation versus follow-up), a third level could be introduced into the hierarchy: ''participant within trial,'' ''trial within type of analysis,'' and ''type of analysis.''

Fig. 2 shows the distributions of reduction in breast cancer mortality under varying assumptions. The Mantel–Haenszel curve results from assuming that the trials are homogeneous. The other two curves are probability densities based on the trial results and are derived from a Bayesian hierarchical analysis. The three curves have essentially the same center, 17% or 18%. But variability differs. The (Bayesian) 95% probability interval for the mean reduction in the population of trials is from −1% to 36%, wider than the Mantel–Haenszel confidence interval (5%–29%), and it includes 0. The 95% predictive (next trial) interval is wider yet, −17% to 51%. This greater width reflects the substantial uncertainty in the benefits of screening in the next trial. The probability that the next cohort of women who choose screening will actually benefit from the screening—that is, experience a reduction in breast cancer mortality—is the area under the curve to the right of 0, which is about 85%.

The Bayesian hierarchical approach allows for estimating the reduction in mortality in a particular trial, taking the results of the other trials into consideration. The last column of Table 2 gives Bayesian estimates of mortality reduction. The trials ''bor-

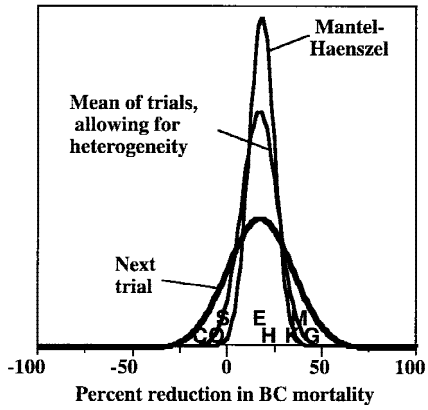

Fig. 2. Distribution of reduction in breast cancer (BC) mortality. The Mantel–Haenszel likelihood function assumes homogeneity in the eight trials of screening mammography [H = Health Insurance Plan of New York; K = Kopparberg; O = Östergötland; M = Malmö; E = Edinburgh; G = Gothenburg; S = Stockholm; and C = Canada National Breast Screening Study I] *(6–12).* The other two curves are from a Bayesian analysis, which allows for trial heterogeneity. The areas under the curves are the same. Allowing for heterogeneity, the posterior distribution of the population mean exhibits greater variability than that observed with the Mantel–Haenszel likelihood function. The predictive distribution of the next observation (trial) from the population of trials—including that of a particular woman choosing screening—has much greater variability than that observed with the Mantel–Haenszel function.

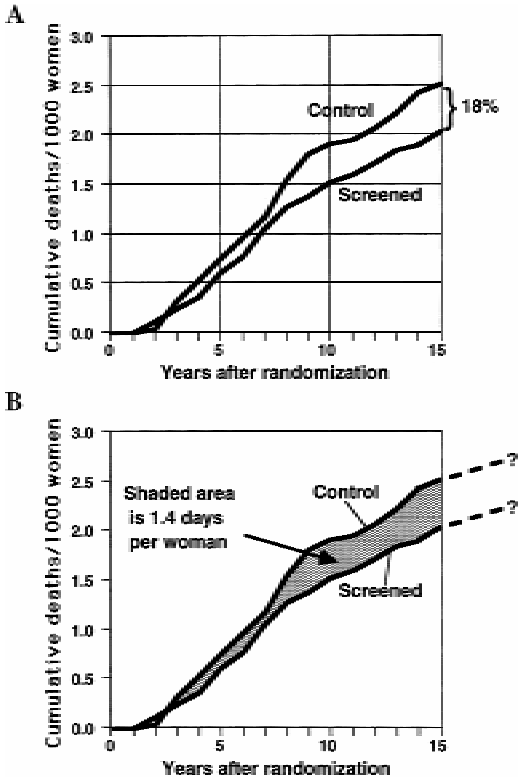

Fig. 3. Breast cancer mortality per 1000 women, all eight trials of screening mammography *(6–12).* **A**) Estimated from Larsson et al. *(20),* Miller et al. *(7),* Alexander et al. *(12),* and Shapiro et al. *(6),* with updated overall mortality as reported by Hendrick et al. *(13).* Larsson et al. contains an older version of the Swedish trial results than does Hendrick et al., and, in particular, Larsson et al. includes only Malmö Mammography Screening Trial I. Therefore, the mortality curves in this figure incorporate updated results of the Swedish trials to include Malmö Mammography Screening Trial II *(11).* For trials with less than 15 years of follow-up, mortality rates from the end of follow-up out to 15 years were estimated using parallel straight lines. The observed 15-year reduction in mortality for screened women compared with control women is 18%. **B**) The shaded area is equivalent to 3.76 years/1000 women or 1.37 days per woman. This is the estimated increase in life expectancy out to 15 years after the initiation of screening. The text gives estimates of increases in overall life expectancy by assuming that the reduction of 18% applies at 15 years and that it is maintained indefinitely.

row strength'' from each other, and so these estimates are shrunk or regressed toward the common mean of 18%. The amount of shrinkage depends on the distance from the mean and the sample size. The Canadian and Gothenburg trials are at opposite extremes, and so they receive the greatest adjustments—the former upwards and the latter downwards, but both toward the mean. The adjusted estimates for the HIP and the Kopparberg trials coincide, but Kopparberg has moved further because of its smaller sample size.

Screening is an investment, and benefits do not accrue immediately. The estimated 18% mortality reduction applies at the latest available follow-up time, which averages about 15 years after randomization. Fig. 3, A, shows the estimated mortality rates in the two groups from randomization out to 15 years. These curves represent the best and most recent estimates based on the combined information from the trials *(7,12,13,20).* However, mortality results for the Swedish trials as reported by Hendrick et al. *(13)* are more recent than those reported by Larsson et al. *(20),* and the former show a greater benefit for screening. For example, Larsson et al. include only the first phase of the Malmö trial (Malmö Mammography Screening Trial I). The mortality curves in Fig. 3, A, incorporate updated data from the Swedish trials, including Malmö Mammography Screening Trial II *(11),* by adjusting the curves of Larsson et al. *(20)* to agree with Hendrick et al. *(13).* For trials with less than 15 years of follow-up, mortality rates out to 15 years were estimated using straight parallel lines.

The area between the two mortality curves, shown in Fig. 3, B, is the estimated additional life expectancy for screening over control. This is 3.8 years per 1000 women, or 1.4 days per woman. To estimate the total life expectancy benefit requires an assumption about the discrepancy between the two curves in subsequent years. The observed benefit will likely continue beyond 15 years, but there is little information about its magnitude. Consider a range of possibilities, from assuming no incremental benefit beyond 15 years after initiating screening to assuming that the 18% reduction observed at 15 years continues indefinitely thereafter. Consider a 40-year-old woman and account for actuarial survival out to age 85 years. For no incremental benefit after age 55 years, her life expectancy is increased by 1.4 days, and for an 18% lifelong benefit from the early screening, her life expectancy is increased by 5.3 days. [Saltzmann et al. *(21)* assume a 16% reduction in breast cancer mortality and calculate an added life expectancy to be 6.5 days, which drops to 2.5 days when assuming a 3% annual discount rate.]

Evaluation Bias and Other Concerns

I have many concerns about the randomized trials. Some have to do with their conduct and analysis. Several relate to marked

imbalances in the characteristics of the two groups, most favoring the screened group. Some concerns relate to data quality. Reports of results have changed in unusual ways. For example, at the 1996 Falun meeting *(22)*, the Gothenburg trial reported 37 deaths in 129 000 life-years in the control group and 19 deaths in 106 000 life-years in the screened group. The numbers of life-years are about 30% greater in the most recent report (Table 2) *(10,23)*, but the number of deaths in the screening group is now one fewer. The number of deaths in the screened group in the Stockholm trial also decreased by one over the same period, with the numbers of life-years and the number of deaths in the control group unchanged *(9,22)*.

There is insufficient space to address all of my concerns fully. Therefore, I will focus primarily on two issues. One issue relates to trial differences in the type of analysis used (Table 1). Four of the Swedish trials used ''evaluation'' *(24)*, while the other trials used ''follow-up.'' A follow-up analysis counts all deaths attributed to breast cancer, regardless of the age at diagnosis or whether the diagnosis occurred during the screening period. Therefore, a woman who died of breast cancer diagnosed at any age (subsequent to randomization) would be counted as a breast cancer death in a follow-up analysis. A follow-up analysis is unbiased and legitimate, but it is inefficient because cancers diagnosed after the period of screening introduce noise that may dilute differences between the two groups. However, it would be wrong to eliminate from the control group any cancer detected after the screening period because this same cancer may have been detected earlier had the woman been assigned to screening. On the other hand, her cancer may not have been detected had she been screened, and, therefore, women in the screened group whose cancers are detected after screening must also be included.

An evaluation analysis eliminates this dilution by requiring a single mammographic examination in the control group, timed to correspond with the last mammographic examination in the screened group. This design was used in the four indicated Swedish trials. For example, women born in 1944 entered the Gothenburg trial in April 1984 when they were age 39 or 40 years. A woman assigned to screening was scheduled to have five mammograms 18 months apart *(10)*. A woman assigned to control would get a single mammographic examination at the fifth and last screen in May 1990. The only deaths that count in an evaluation analysis are those resulting from breast cancers detected (after randomization and) before or at the May 1990 screen. Thus, cancers detected after the end of the schedule of screening that could dilute group comparisons are eliminated.

An essential requirement of an evaluation analysis is for the control group screen to occur when the last mammographic examination would have occurred had the woman been in the screened group. Unfortunately, the control mammographic examination of women in their forties was delayed in all four trials that used evaluation analyses, and the delay was substantial, about 1 year *(24)*—whether and by how much the screening group's mammograms were also delayed is not clear, except in the Gothenburg trial where the delay was about 3 months *(10)*. As a result, many cancers were detected in the control group that would not have been found in the screened group. This bias is substantial. A consequence of this bias was that the control group in the Gothenburg trial had 12% more cancers detected per participant than the screened group *(10)*.

Not adhering to the planned schedule of mammograms in the control group is a major flaw in the conduct and ensuing analysis of these four Swedish trials. Apparently concerned about this flaw, the Gothenburg investigators carried out a second analysis in which the four deaths (10%) arising from breast cancers detected at the control screen were eliminated *(10)*. On the basis of modeling, they increased the number of life-years in the screened group by 8%, and, therefore, the net change in mortality reduction was only 2% (or one percentage point)—from 45% to 44%. However, the assumptions and logic behind the modeling as described in their publication *(10)* are not clear. These modeling assumptions led the study investigators to the conclusion that a clearly biased evaluation analysis turned out not to be biased after all.

The great virtue of randomized trials is they obviate the need for modeling. The investigators of the four trials using an evaluation analysis have the means to present their trial results unencumbered by evaluation bias and by modeling assumptions. They should provide a follow-up analysis as well as an evaluation analysis, counting all deaths from breast cancer in both groups, regardless of when they were diagnosed. Not only would this eliminate evaluation bias, but using the same analysis as the other four trials would also allow for better comparisons of the results of the eight trials. In addition, should it develop that radiation from mammography induces breast cancer in younger women, then these cancers would likely show up after the screening period and would be ignored in an evaluation analysis.

It is not clear what effect adopting a uniform follow-up analysis would have on the estimated mortality reduction of 18%. Since it could greatly decrease this estimate, no organization should endorse screening until these follow-up analyses are forthcoming. My recommendations for communicating the potential benefits of screening to women are predicated on the available evaluation analyses and would likely change depending on what the follow-up analyses show.

The second issue on which I would like to focus relates both to concerns about the conduct of the Canadian trial and criticisms of it. This trial has been subjected to much criticism *(7,25–29)*. Some authors [e.g., *(30)*] have suggested that the Canadian trial not be included in meta-analyses. All trials have flaws, but, in many ways, the Canadian trial is of the highest quality among the screening trials *(31,32)*. A legitimate concern is that the randomization process employed in the trial was not blinded to center coordinators, and, therefore, it could have been subverted *(27,30)*. Circumstantial evidence exists that the screened group had substantially more locally advanced cancers at the initial visit. Bailar and MacMahon *(28)* addressed some of the possible types of subversion and found little evidence for them. However, the trial's randomization procedures left room for types of subversion that could not be detected retrospectively. If subversion occurred, it cannot be discerned from a comparison of risk factors in the two groups *(7)*. Moreover, according to the trial's principal investigator (Miller AB: data presented at the consensus development conference), removing from consideration all volunteers who had locally advanced breast cancers detected at the initial visit would not change the trial's overall negative conclusion—although it would lessen its magnitude.

Among the reasons suggested for excluding the Canadian trial is that the control group received clinical breast examinations *(30)*. Comparing two screening strategies is legitimate, and it would be important to know whether or not clinical breast examination followed by mammography in abnormal cases compares favorably with mammography in every case. (The existence of differences in control strategy is one of the reasons that an analysis allowing for trial heterogeneity is appropriate.) The Canadian trial is an outlier *(30)*, but so is the Gothenburg trial—it makes no more sense to exclude one than it does to exclude the other.

Some criticisms of the Canadian study are statistical. One such criticism is that the trial was under-powered *(26)*, even though it was the largest of all eight trials. After a study is completed, power calculations are irrelevant. We know the trial sample size and the trial results. Power is calculated assuming potential but fictitious benefits. Actual trial results may contradict assumptions made at the planning stage. If the results are inconclusive, then we will know that. Whatever the results, they contribute to our knowledge. An openly and honestly conducted trial should not be ignored because it was too small to achieve an artificial goal or because its results are unexpected.

The designs of all the trials make their relevance for women in their forties questionable. Women in the trials were randomly assigned at any age from 40 to 49 years. Some women who were randomly assigned in their forties got most of their scheduled mammograms in their fifties. For example, for the scheme used in the Kopparberg and Östergötland trials, 30% of the mammograms—and an unknown proportion of any benefit derived from screening—occurred when the women were more than 49 years of age. In all the trials, for women ages 40–49 years at study entry, the average age of the women at the time of their mammograms was in the late forties (48 years in the Kopparberg and Östergötland trials) or, in the Malmö and Edinburgh trials, the early fifties. Assuming that the benefits of screening are greater in one's fifties, the results in Table 2 overestimate the benefits of screening in the forties. (It would be important to address the benefits of screening depending on age at the initial screen—for example, ages 40–44 years versus 45–49 years. This type of information is available only sporadically from the trials.)

Risks of Mammography Should Not Be Doubly Counted

To decide whether or not to be screened, a woman might make separate lists of the benefits and risks and compare them. On the benefit side of the ledger is a possible delay in death, as discussed above. The risks have been discussed extensively *(3,33–39)* and include false-positive findings, inconvenience, pain, and anxiety. All of these risks should count on the negative side of the ledger, but not all commonly cited risks should be counted there. Among these latter risks are the following: 1) radiation-induced breast cancer that develops during the screening period, 2) false-negative mammograms, 3) false reassurance from false-negative mammograms (which may lead one to eschew breast self-examinations and clinical breast examinations), and 4) finding aggressive tumors when they are smaller (which may lead to less than optimal chemotherapy). These risks have already been included—as negatives—on the benefit side of the ledger and should not be counted twice. For example, false-negative mammograms may lead to increased cancer mortality in the screening groups, and this would lower the reduction in breast cancer mortality. If screening mammography reduces breast cancer mortality overall, then the amount of the reduction would not be as great as it might otherwise have been if the technique were more sensitive. Once potential mortality benefits have been considered, risks 1–4 can be ignored in decision making.

This is not to say that such risks and the understanding of such risks are unimportant. Knowing how a woman's characteristics can influence mammography's false-negative rate, for example, may allow for improvements in mortality reductions through a lowering of this rate.

In Situ Cancers—Benefit or Risk?

Screening mammography is effective in detecting ductal carcinoma *in situ* (DCIS), and, in fact, its widespread use has created an epidemic of *in situ* cancers *(38)*. Finding DCIS may be both a benefit and a risk. Unfortunately, it is not possible to discriminate very well between those *in situ* cancers that will develop into invasive disease and those that would not progress even if left undetected. Finding the former may extend some women's lives, but finding the latter serves only to increase the number of women who are told they have breast cancer. Any benefit translates into a mortality reduction for screening, and, therefore, it is already considered on the positive side of the ledger. But once the estimated mortality reduction due to screening has been considered, detecting DCIS is properly viewed as a risk, and it should be listed only on the negative side of the ledger. Not counting the detection of DCIS as a benefit is analogous to not counting false negatives as a risk.

Communicating Benefits and Risks to Women and Their Caregivers

Risk is inherently quantitative, and yet risks conveyed quantitatively are difficult for many people to understand *(40)*. Therefore, it may not be appropriate to ask women to decide for themselves whether or not to be screened, as did the Consensus Development Panel *(3)*. However, some women do understand risks, and it is paternalistic to suggest that all women should be screened (or not screened) just because some of them do not. Clinicians should communicate the benefits and risks to their patients as effectively as possible. If a woman balks and says ''just tell me whether or not I should get mammograms,'' then a recommendation is obviously appropriate—but even then it should be presented with qualifications about the benefits and risks and an indication that no one knows if there is a benefit or how large it is.

A variety of measures might be used to communicate benefits. The focus in this commentary has been on relative benefit, i.e., percent reduction in breast cancer mortality. A disadvantage of this measure is that it does not account for incidence, and so it is of little value in making decisions. One measure of absolute benefit is the number of deaths prevented for each 10 000 women screened regularly during their forties—or, what is closely related, the number of mammograms required to save one life. Unfortunately, these measures cannot be calculated

from the trial results. Even if the results shown in Fig. 3 were known to apply to a population of women, it is not possible to say whether a few deaths are prevented or that no deaths are prevented but a moderate number are delayed. If the number of women who benefit is small, then the benefit that accrues to each of them is large; if the number who benefit is large, then the benefit that accrues to each is small. In the former case, and assuming an 18% reduction in breast cancer mortality at 15 years and beyond, about three women per 10 000 screened have their deaths prevented, and about 15 000 (biannual) mammograms are required to prevent one death. On the other hand, if all women who are diagnosed with breast cancer have their lives lengthened by the same amount, then no deaths are prevented but about 150 women per 10 000 screened have their deaths delayed for about 10 months each. Screening is effective in both cases, but the distribution of the benefit is different. On the basis of the randomized trials, there is no way to distinguish between these two extremes or any of the intermediate possibilities.

Communicating benefit is easier if the units are understandable. It is easy to relate to numerical gains in life expectancy. The drawback of using such units is that they are averages (just as reported results from trials are averages). Thus, it is unlikely that any 40-year-old woman benefits by exactly 1.4 days over 15 years or 5.3 days over her lifetime. Screening is a lottery. Any winnings are shared by the minority of women—about one in 60 or 70—who are diagnosed with breast cancer in their forties. The overwhelming majority of women experience no benefit. In the lottery analogy, they lose their entry fee; in screening, they pay with the time involved and with the risks associated with screening.

Among the risks of screening mammography are false positives *(39)* and the possibility of finding DCIS or cancers that are relatively indolent and, unbeknownst to anyone, not life-threatening. Another risk is the flip side of finding cancer early, i.e., a woman knows she has cancer for a longer period of time. This period is typically 18 months *(22)*, which is about 9 days per cancer detected mammographically. A woman should evaluate the time spent getting mammograms over this decade and add it to time analogs of the negative aspects of the associated risks, comparing the total to the best estimate of 5.3 days of life saved.

An additional advantage of communicating benefit in terms of increased life expectancy is that the estimates are rather stable. If the mortality reduction beyond 15 years after initiating screening is rather different from 18%, then the estimated incremental life expectancy changes by only a few days.

A way to understand risks is to relate them with risks that are familiar. For example, the estimated average of 5 days of life lost if a woman in her early forties delays mammography for 10 years is similar to that for not wearing a seat belt over 20 years of typical automobile travel, of riding a bicycle for 15 hours without a helmet (or 50 hours if wearing a helmet), and of gaining two ounces of body weight (and keeping them on) *(41)*.

Advice for Women in Their Forties

Although some people feel passionately, no one knows for certain whether, or by how much, getting mammograms before the age of 50 years increases life expectancy. Some women die from breast cancer, even if their cancers are small when detected by mammography. In the eight randomized trials, women were assigned to receive regular mammograms or not by the equivalent of a coin toss. For the first 8 years after enrolling in the trials, there was little difference in the numbers of deaths in the two groups. However, after about 15 years, there were 29 deaths due to breast cancer per 10 000 women assigned to receive mammograms and 36 breast cancer deaths per 10 000 women in the control groups. The reduction of seven deaths per 10 000 women after 15 years is a rough estimate and it may not apply to women outside of the trials. In particular, there might be no mortality reduction in any particular set of women aged 40 to 49 years who undergo regular screening. Assuming that the observed reduction of seven deaths per 10 000 at 15 years applies to women generally, such a reduction would make having mammograms attractive to some women, but it does not translate into a large increase in life expectancy. Assuming that this reduction applies at 15 years and beyond, regular mammography adds about 5 days to the life expectancy of each woman screened.

The risks of mammography are better understood than the benefits. All women who get mammograms are subject to these risks. One risk is that of false positives. A false positive occurs when a woman who does not have cancer has an abnormal mammographic examination and requires follow-up procedures. About 30 of 100 women who get five mammograms between the ages of 40 and 49 years experience at least one false positive. Between five and 10 of these 100 women will undergo at least one biopsy—the surgical removal of a small amount of breast tissue that can be checked for cancer by a pathologist. Women usually experience anxiety as a result of a false-positive mammogram, and some women's anxiety continues even after they find they do not have cancer. In addition, about 15 of 100 women experience severe pain from mammograms. Finally, since mammography is good at finding cancer, it can find cancers that would never threaten a woman's life, and, unfortunately, we are not able to tell which cancers these are.

The screening groups in some of the controlled trials had annual mammograms, whereas, in other trials, the time between mammograms was longer. If having five mammograms between the ages of 40 and 49 years is good, then having 10 mammograms is probably better. However, there is no evidence from the trials that a yearly schedule is beneficial. Most of any benefit from mammography is found when the procedure is performed every other year, and the incremental benefit of yearly mammography is smaller. Of course, having twice as many mammograms approximately doubles the chances of the risks cited above.

Regular mammography between the ages of 40 and 49 years probably lengthens some women's lives. But the benefits are not great. You should weigh the possible benefits and risks and decide whether screening is appropriate for you.

Appendix: Hierarchical Bayesian Model and Sensitivity Analyses

This appendix gives the mathematical specifics of the hierarchical model whose results are presented in this commentary, for example, in the rightmost column of Table 2 and the two ''heterogeneity curves'' in Fig. 2. In addition, it contains sensi-

tivity analyses with respect to the prior distributions of the model parameters.

Suppose the rates of death due to breast cancer in the control groups of the eight trials are $c_1, c_2, \ldots, c_8$ and the corresponding rates in the screening groups are $s_1, s_2, \ldots, s_8$. Then $r_i = s_i/c_i$ is the hazard ratio of screening to control in trial i, and the corresponding breast cancer mortality reduction is $1 - r_i$. Suppose the numbers of life-years in the control and screening groups are, respectively, $n_1, n_2, \ldots, n_8$ and $m_1, m_2, \ldots, m_8$. Assume that the number of observed deaths in the control group of trial i is a Poisson variable with parameter $n_i c_i$ and the corresponding number in the screening group is a Poisson variable with parameter $m_i s_i$.

In a Bayesian analysis, unknown parameters of sample distributions have probability distributions *(42)*. Assume the control rates c_i have a gamma distribution with parameters a and b. In a hierarchical Bayesian analysis, parameters a and b are also unknown and so they themselves have probability distributions, taken to be unit exponentials. The distribution of the hazard ratios r_i is lognormal with parameters μ and σ^2, and so the conditional mean of r_i is $\exp(\mu + \sigma^2/2)$.The prior distributions of μ and σ^2 are, respectively, standard normal and inverse gamma with parameters α and β [using the parameterization in which the mean of σ^2 is $1/(\beta(\alpha - 1))$]. The results in this commentary take α and β equal to 5, in which case the mean of σ^2 is 0.05 and the mean of σ is 0.22. Computation involves Markov chain Monte Carlo methods *(43)*.

Bayesian conclusions depend on the prior distribution assumed. Reports of Bayesian analyses should address the effect of the prior distribution on the conclusions—a sensitivity analysis. The conclusions reported here are robust with respect to the prior distributions indicated here, with one exception: the parameters α and β of the prior distribution of σ^2. The reason conclusions are sensitive to these parameters is that σ reflects the trial heterogeneity and there is only a moderate amount of information concerning heterogeneity in a sample of only eight trials. If σ is large, then there is substantial heterogeneity; if σ is small, then the eight hazard ratios r_i are likely to be similar—homogeneity. Taking α and β large means σ is likely to be small and corresponds to assuming that the trials are homogeneous. Taking α and β small means that σ is likely to be large and corresponds to assuming heterogeneity—that the trials have different hazard ratios. Assuming homogeneity in the Bayesian approach gives posterior probabilities that are similar to the frequentist analogs (confidence intervals, for example) using the Mantel–Haenszel test. Assuming heterogeneity in the Bayesian approach means that the individual hazard ratios r_i are taken to be different *a priori* and there is no ''borrowing strength'' from one trial to another. Neither extreme is appropriate; there are basic differences in the trials that may well affect the screening benefit, and yet, on the other hand, all eight trials address the benefits of mammography. Assuming that α and β are both equal to 5 is intermediate in that it includes the possibility that the trials are heterogeneous without assuming them to be heterogeneous.

To be specific and to quantify the above statements, take α and β to be equal and compare conclusions when the common value is either 10 or 3 as opposed to the assumed value of 5. Consider the posterior distribution of σ and of the mean of the r_i's, that is, $\exp(\mu + \sigma^2/2)$. (The predictive distribution of mortality reduction in the next trial is always more variable than is the mean of the r_i's, just as for the case $\alpha = \beta = 5$; cf. Fig. 2.) When $\alpha = \beta = 5$, the prior mean and standard deviation of σ are 0.22 and 0.055. The corresponding posterior mean and standard deviation of σ are 0.21 and 0.046, and the posterior mean and standard deviation of the mean hazard ratio are 0.83 and 0.096. For $\alpha = \beta = 10$, the corresponding values are 0.10, 0.017, 0.11, 0.018, 0.81, and 0.074. For $\alpha = \beta = 3$, the corresponding values are 0.38, 0.14, 0.31, 0.074, 0.85, and 0.12. (Since the posterior distribution of the mean hazard ratio is close to normal, the respective 95% posterior probability intervals are 0.64 to 1.02, 0.67 to 0.96, and 0.61 to 1.09.) Assuming $\alpha = \beta = 10$ allows for very little heterogeneity and gives conclusions that are roughly consistent with those of a Mantel–Haenszel test. Assuming $\alpha = \beta = 3$ allows for a substantial amount of heterogeneity—even more than that present in the data, as evinced by the decrease in the posterior mean of σ (0.31) from the prior mean (0.38).

Conclusions of a Bayesian analysis depend on the prior distribution selected, and, in this sense, a Bayesian approach mirrors the controversy regarding screening mammography for women in their forties. People have different opinions about study heterogeneity, and so they will draw different conclusions, even though they observe the same data. One prior distribution is not inherently better than another. Examining the characteristics of the trials (but not their results) may lead one person to regard the benefits as very likely similar over the eight trials, and such a person would take α and β to be large—for example, both equal to 10. Another person may view the trials as being quite different and choose α and β equal to 3. The prior distribution used in this commentary is intermediate between these two extremes in that it allows for a moderate degree of heterogeneity.

REFERENCES

(1) Taubes G. The breast-screening brawl [news] published erratum appears in Science 1997;267:1485]. Science 1997;275:1056–9.

(2) Ransohoff DF, Harris RP. Lessons from the mammography screening controversy: can we improve the debate? Ann Intern Med 1997;127:1029–34.

(3) National Institutes of Health Consensus Development Conference Statement: breast cancer screening for women ages 40–49, January 21–23, 1997. J Natl Cancer Inst 1997;89:1015–26.

(4) Statement from the National Cancer Institute on the National Cancer Advisory Board Recommendations on Mammography. National Institutes of Health. Office of Cancer Communications. March 27, 1997.

(5) Kerlikowske K. Efficacy of screening mammography among women aged 40 to 49 years and 50 to 69 years: comparison of relative and absolute benefit. Monogr Natl Cancer Inst 1997;22:79–86.

(6) Shapiro S. Periodic screening for breast cancer: the HIP randomized controlled trial. Monogr Natl Cancer Inst 1997;22:27–30.

(7) Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study: update on breast cancer mortality. Monogr Natl Cancer Inst 1997;22:37–41.

(8) Tabar L, Chen HH, Fagerberg G, Duffy SW, Smith TC. Recent results from the Swedish two-county trial: the effects of age, histologic type, and mode of detection on the efficacy of breast cancer screening. Monogr Natl Cancer Inst 1997;22:43–7.

(9) Frisell J, Lidbrink E. The Stockholm Mammographic Screening Trial: risks and benefits in age group 40–49 years. Monogr Natl Cancer Inst 1997;22:49–51.

(10) Bjurstam N, Bjorneld L, Duffy SW, Smith TC, Cahlin E, Eriksson O, 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;80:2091–9.

(11) Andersson I, Janzon L. Reduced breast cancer mortality in women under age 50: Updated results from the Malmo Mammographic Screening Program. Monogr Natl Cancer Inst 1997;22:63–7.
(12) Alexander FE. The Edinburgh Randomized Trial of Breast Cancer Screening. Monogr Natl Cancer Inst 1997;22:31–5.
(13) Hendrick RE, Smith RA, Rutledge JH III, Smart CR. Benefit of screening mammography in women aged 40-49: a new meta-analysis of randomized controlled trials. Monogr Natl Cancer Inst 1997;22:87–92.
(14) Narod SA. On being the right size: a reappraisal of mammography trials in Canada and Sweden [letter]. Lancet 1997;349:1846.
(15) Agresti A. Categorical data analysis. New York: Wiley; 1990.
(16) Berry DA. A case for Bayesianism in clinical trials. Stat Med 1993;12: 1377–404.
(17) Berry DA. Interim analysis in clinical trials: classical vs. Bayesian approaches. Stat Med 1985;4:521–6.
(18) O'Brien PC, Fleming TR. A multiple testing procedures for clinical trials. Biometrics 1979;35:549–56.
(19) Berry DA, Stangl DK. Bayesian methods in health-related research. In: Berry DA, Stangl DK, editors. Bayesian biostatistics. New York: Marcel Dekker; 1996. p.1–66.
(20) Larsson LG, Andersson I, Bjurstam N, Fagerberg G, Frisell J, Tabar L, Nyström L. Updated overview of the Swedish Randomized Trials on Breast Cancer Screening with Mammography: age group 40–49 at randomization. Monogr Natl Cancer Inst 1997;22:57–61.
(21) Saltzmann P, Kerlikowske K, Phillips K. Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age [published erratum appears in Ann Intern Med 1998;128:878]. Ann Intern Med 1997;127:955–65.
(22) Tabar L, Larsson LG, Andersson I, Duffy SW, Nystrom L, Rutqvist LE, et al. Breast cancer screening with mammography in women aged 40–49 years: report of the organizing committee and collaborators, Falun Meeting, Sweden, March 21–22, 1996. Int J Cancer 1996;68:693–9.
(23) Bjurstam N, Bjorneld L, Duffy SW, Smith TC, Cahlin E, Erikson O, et al. The Gothenburg Breast Cancer Screening Trial: preliminary results on breast cancer mortality for women aged 39–49. Monogr Natl Cancer Inst 1997;22:53–5.
(24) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.
(25) Kopans DB, Feig SA. The Canadian National Breast Screening Study: a critical review. AJR Am J Roentgenol 1993;161:755–60.
(26) Kopans DB, Halpern E, Hulka C. Statistical power in breast cancer screening trials and mortality reduction among women 40–49 years of age with particular emphasis on the National Breast Screening Study of Canada. Cancer 1994;74:1196–203.
(27) Tarone RE. The excess of patients with advanced breast cancer in young women screened with mammography in the Canadian National Breast Screening Study. Cancer 1995;75:997–1003.
(28) Bailar JC, MacMahon B. Randomization in the Canadian National Breast Screening Study. Report of a review team appointed by the National Cancer Institute of Canada. Can Med Assoc J 1997;156:213–5.
(29) Boyd NF. The review of randomization in the Canadian National Breast Screening Study. Is the debate over? Can Med Assoc J 1977:156:207–9.
(30) Kopans DB. An overview of the breast cancer screening controversy. Monogr Natl Cancer Inst 1997;22:1–3.
(31) Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50: a quality assessment and meta-analysis. Med J Aust 1995;162:625–9.
(32) Glasziou PP, Irwig L. The quality and interpretation of mammographic screening trials for women ages 40–49. Monogr Natl Cancer Inst 1997;22: 73–7.
(33) Fletcher SW. Breast cancer screening among women in their forties: an overview of the issues. Monogr Natl Cancer Inst 1997;22:5–9.
(34) Feig SA, Hendrick RE. Radiation risk from screening mammography of women aged 40–49 years. Monogr Natl Cancer Inst 1997;22:119–24.
(35) Rimer BK, Bluman LG. The psychosocial consequences of mammography. Monogr Natl Cancer Inst 1997;22:131–8.
(36) Harris R. Variation of benefits and harms of breast cancer screening with age. Monogr Natl Cancer Inst 1997;22:139–43.
(37) Kerlikowske K, Barclay J. Outcomes of modern screening mammography. Monogr Natl Cancer Inst 1997;22:105–11.
(38) Ernster VL, Barclay J. Increases in ductal carcinoma *in situ* (DCIS) of the breast in relation to mammography: a dilemma. Monogr Natl Cancer Inst 1997;22:151–6.
(39) Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med 1998;338:1089–96.
(40) Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. Ann Intern Med 1997;127:966–72.
(41) Laudan L. The book of risks. New York: John Wiley & Sons; 1994.
(42) Berry DA. Statistics: a Bayesian perspective. Belmont (CA): Duxbury Press; 1996.
(43) Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman & Hall; 1995.

NOTES

[1]*Author's note:* I was a member of this Consensus Development Panel. My assessments are consistent with the recommendations of the Panel but they represent my own views and do not necessarily reflect the views of other Panel members.

I thank Barbara Rimer and my fellow Consensus Development Panel members (especially Daniel Sullivan) for their helpful discussions and Scott Berry for his help with modeling and for carrying out the Markov chain Monte Carlo calculations described in the Appendix.

Manuscript received February 27, 1998; revised April 30, 1998; accepted April 30, 1998.