Screening Mammography: A Reassessment

AGENCE D'ÉVALUATION DES TECHNOLOGIES
ET DES MODES D'INTERVENTION EN SANTÉ

Québec

Screening Mammography: A Reassessment

Report prepared for AETMIS
by Wilber Deck
with the contribution of Ritsuko Kakuma

March 2006

Agence d'évaluation
des technologies
et des modes
d'intervention en santé
Québec

The content of this publication was written and produced by the Agence d'évaluation des technologies et des modes d'intervention en santé (AETMIS). Both the original English report and its French version, titled *Mammographie de dépistage : une réévaluation*, are available in PDF format on the Agency's Web site.

Scientific review
Alicia Framarin, MD, MSc, Deputy Scientific Director
Jean-Marie Lance, MSc, Senior Scientific Advisor

Editorial Supervision
Suzie Toutant

Page layout
Jocelyne Guillot

Proofreading
Suzie Toutant
Frédérique Stephan

Bibliography research
Micheline Paquin
Denis Santerre

Coordination
Lise-Ann Davignon

Communications and dissemination
Diane Guilbault

For information about this publication or any other AETMIS activity, please contact:

Agence d'évaluation des technologies et des modes d'intervention en santé
2021, Union avenue, Suite 1050
Montréal (Québec) H3A 2S9

Tel.: (514) 873-2563
Fax: (514) 873-1369
e-mail: aetmis@aetmis.gouv.qc.ca
http://www.aetmis.gouv.qc.ca

How to cite this publication:

Agence d'évaluation des technologies et des modes d'intervention en santé (AETMIS). Screening Mammography: A Reassessment. Report prepared by Wilber Deck with the contribution of Ritsuko Kakuma (AETMIS 05-03). Montréal: AETMIS, 2006, xii-77 p.

Legal Deposit
Bibliothèque nationale du Québec, 2006
Library and Archives Canada, 2006
ISBN 2-550-44486-X (Printed) (French edition ISBN 2-550-44487-6)
ISBN 2-5550-44486-8 (PDF) (French edition ISBN 2-550-44488-4)

MISSION

The mission of the Agence d'évaluation des technologies et des modes d'intervention en santé (AETMIS) is to contribute to improving the Québec health-care system and to participate in the implementation of the Québec government's scientific policy. To accomplish this, the Agency advises and supports the Minister of Health and Social Services as well as the decision-makers in the health care system, in matters concerning the assessment of health services and technologies. The Agency makes recommendations based on scientific reports assessing the introduction, diffusion and use of health technologies, including technical aids for disabled persons, as well as the modes of providing and organizing services. The assessments take into account many factors, such as efficacy, safety and efficiency, as well as ethical, social, organizational and economic implications.

EXECUTIVE

BOARD OF DIRECTORS

FOREWORD

SCREENING MAMMOGRAPHY: A REASSESSMENT

Screening mammography, a technique which is 50 years old, aims to advance the diagnosis of breast cancer in order to offer early treatment, thereby improving the chances of cure. The practice of mammography, in constant evolution, varies widely according to the equipment used, the interpretation of films, and program aspects such as the age of women when they are invited to their first screen, the interval between rounds of screening and participation rates.

In 1990, a report by Québec's Conseil d'évaluation des technologies de la santé (CETS) recommended that this practice be structured as part of a formal program which would include quality standards. A second CETS report published in 1993 underlined the absence of proof in favour of screening women younger than 50. Since 1998, the *Programme québécois de dépistage du cancer du sein* (PQDCS) offers systematic screening every 2 years to all women aged 50 to 69. Younger women can still obtain mammography with a prescription from their physician. Many other countries have also started screening mammography programs, but there remains controversy regarding the age when screening should start, and recent studies have raised doubts about the value of screening mammography at any age.

In this context, the ministère de la Santé et des Services sociaux (MSSS) has asked the Agence d'évaluation des technologies et des modes d'intervention en santé (AETMIS) to re-examine the quality of the scientific evidence on which the PQDCS is based and on the pertinence of extending screening to women less than 50 years old. This report evaluates various aspects of the validity of screening trials and their pertinence with regard to the performance and quality assurance of a modern screening program such as the PQDCS.

The analysis indicates that most trials had serious problems with validity, making it difficult to use them to estimate the potential benefits of mammography. The best trials show a modest reduction in breast cancer mortality; this reduction is greater when the analysis is limited to women 50 to 69 years old. No trial was designed and conducted in such a way that the full potential of screening mammography could be achieved. It is thus plausible that a modern program, conducted under conditions of superior quality, might obtain better results than the trials suggest.

In conclusion, a screening program targeting women 50 to 69 years old remains justified by the available data. This justification does not extend to younger women. However, it is possible that screening of individual women, based on a personalized risk assessment, might be of benefit to some younger women. This conclusion should be reviewed in a few years, when results of the ongoing UK Age Trial become available. In the meantime, a modern mammography screening program like the PQDCS can benefit from measures which aim to maximize the quality of screening and to increase participation rates.

In submitting this report, AETMIS hopes to contribute to the optimal use of screening mammography for the benefit of all women.

Dr. Luc Deschênes
President and Chief Executive Officer

ACKNOWLEDGEMENTS

This report was prepared by Dr. Wilber Deck, MSc (epidemiology and biostatistics), with the collaboration of Ritzuko Kakuma, PhD candidate (epidemiology), at the request of the Agence d'évaluation des technologies et des modes d'intervention en santé (AETMIS).

The Agency would like to thank the external reviewers, who have provided extremely helpful comments:

Dr. Jean-François Boivin
Full Professor, Department of Epidemiology and Biostatistics, McGill University, Montréal, Québec

Dr. Jacques Brisson
Professor, Faculty of Medicine, Université Laval, researcher, Unité de recherche en santé des populations, Hôpital du Saint-Sacrement, and associate expert, Institut national de santé publique du Québec

Dr. Heather Bryant
Research Scientist, Division of Population Health and Information, Tom Baker Cancer Centre, Calgary, Alberta

Julietta Patnick
Director, UK NHS Breast Cancer Programmes, United Kingdom

Dr. André Robidoux
Surgical Oncologist, Director, Clinique des maladies du sein, Hôtel-Dieu de Montréal (CHUM), Montréal, Québec

AETMIS would also like to thank a radiologist external reviewer (anonymous) for his comments.

DISCLOSURE OF CONFLICT OF INTEREST

None declared.

SUMMARY

INTRODUCTION

Eight trials examining the performance of screening mammography have been conducted in the USA, Sweden, the United Kingdom and Canada, beginning in 1963. A first report by the Conseil d'évaluation des technologies de la santé (CETS) published in 1990 concluded that screening mammography trials had shown reductions in mortality from breast cancer of 35%, with 45% in the subgroup of women aged 50 to 69. A second report in 1993 concluded that mammographic screening of women under 50 had not been shown to reduce mortality. By the year 1998, when Québec introduced the *Programme québécois de dépistage du cancer du sein* (PQDCS), all Canadian provinces and many other countries had organized screening programs in place.

A recent Cochrane Collaboration Group review, challenging the belief that mammography screening is an effective tool for reducing breast cancer deaths, has raised concerns about the validity of the published randomized trials. This update addresses three questions:

(1) What is the strength of the scientific evidence on which screening mammography programs are based?

(2) What is the evidence in support of screening for women aged 40 to 49 years?

(3) What are the implications of research studies for maximizing the effectiveness of modern programs such as the *Programme québécois de dépistage du cancer du sein* (PQDCS)?

METHODOLOGIC ANALYSIS

An evaluation of efficacy trials essentially aims to determine whether the conditions under which the trials were performed and the results that were obtained can guide decisions regarding the intervention in question. Scientific evidence must satisfy three prerequisites in order for it to be the basis for such decision-making: relevance, validity, and precision. A study is **relevant** if it is designed to contrast two or more interventions that are options of interest to decision makers. A study is **valid** if it is designed, conducted and analyzed in such a way as to ensure that no important biases affect the measured comparison of the effectiveness of the technologies that are compared. A study is **precise** if it allows for an estimation of efficacy that is not vulnerable to random effects.

Previous analyses of screening mammography trials have tended to emphasize trials' validity, and in particular factors that might bias trials' results in an unknown direction. These analyses have also calculated the precision of the estimates of breast cancer mortality reduction, and have narrowed confidence intervals around these estimates by combining trials' results in meta-analyses. However, since the issues of relevance and of bias in known direction have not been adequately addressed, they will be further developed in this reassessment, which also includes a meta-analysis.

Relevance

To better appreciate the concept of relevance, we introduce here the notion of *contrast*, which represents the opposition or temporary divergence between an experimental intervention, offered to the screening cohort, and a reference strategy offered to the control cohort. Since the question at issue here is the value of mammography screening, relevant trials are those which contrast screening mammography with no screening. We will thus not include in this analysis trials which compare in principal different screening strategies. In practice, the reference strategy (no screening) may include some uncontrollable screening activities, which will weaken the contrast with the screening intervention.

Validity

A valid study must be a fair comparison between screening and no screening. Thus screening and control cohorts should have the same baseline risk of breast cancer mortality, should be treated equally in all regards except concerning the screening or control intervention, and should have the information on their outcome measured in a way that is independent of their assignment to the screening or control group. Validity can be compromised by bias of known direction and by bias of unknown direction.

In this evaluation, to further develop the notion of bias of known direction, we use the concept of *strength of contrast*. It corresponds to the degree to which a trial succeeds in bringing out the divergence between the two strategies compared and in measuring the effects that this divergence produces. Five elements are evaluated in this report which help assess the strength of contrast:

- the technical contrast, or the nature of the difference between screening and control interventions;
- the era in which these techniques are applied;
- the quality of the intervention, including quality control measures;
- rates of participation and contamination measured among screening and control cohorts; and
- the timing of the measurement of the effects of screening on mortality (or timing dilution).

For each trial, a score for the strength of contrast corresponds to the product of individual estimates for each of these elements, as assessed by two researchers in this analysis. For comparison's sake, we have applied this scale to a modern screening program, the PQDCS, in an analogous fashion.

As in other analyses, we also examine trials' biases of unknown direction, in particular concerning randomization, the equivalence of the risk of breast cancer mortality between the screening and control cohorts, the equivalence of criteria for exclusion from the two cohorts, and the equivalence of the follow-up of the two groups. A score is attributed to each of these elements, and the sum of these scores constitutes a global validity score for each of the eight published trials.

Precision

Precision is evaluated by progressively combining the results of screening trials, weighted by their variance, adding in studies in order of their score on our validity scale.

ANALYSIS OF PUBLISHED TRIALS AND COMPARISON TO THE QUÉBEC CONTEXT

The Québec program involves screening mammography every two years offered to about 900,000 Québec women 50 to 69 years old. The program’s structure, process and objectives are based on established breast cancer screening programs in place in Sweden, the United Kingdom and Australia. Based on pre-existing medical facilities, it includes approximately 80 screening centres, which must meet province-wide quality assurance standards. In 2003, participation rates were 46.7%. However, if all mammographic exams are included, including mammography outside the program and diagnostic mammography, mammography rates in women 50 to 69 years old reached 63%, compared to 50% before the onset of the program.

Eight screening mammography trials have published results; a ninth trial, involving younger women only (UK Age Trial), begun in 1991 in the United Kingdom, has not yet reported final mortality results. The eight trials are the following, in order of their initial years:

- Health Insurance Plan Trial (HIP), New York (1963);
- Malmö Mammographic Screening Trial (Malmö), Sweden (1976);
- Two-County Trial (TCS), Sweden (1977);
- Edinburgh Randomised Trial of Screening for Breast Cancer (Edinburgh), Scotland (1979);
- National Breast Screening Study #1: women 40 to 49 years old (NBSS-1), Canada (1980);
- National Breast Screening Study #2: women 50 to 59 years old (NBSS-2), Canada (1980);
- Stockholm Mammographic Screening Trial (Stockholm), Sweden (1981);
- Gothenburg Breast Screening Trial (Gothenburg), Sweden (1982).

Although the results of trials have been used to estimate how effective a screening program might be in reducing mortality, it is important to first compare the screening regimens in the trials with the screening regimens that modern programs such as the PQDCS have put in place. This comparison applies particularly to criteria of relevance and validity related to the notion of contrast. In several regards, trials have obtained elements of contrast which are stronger than those in the Québec program: Some have included clinical breast exams along with mammography (HIP, NBSS), some have used annual mammography (HIP, NBSS-1 and -2) rather than every two years, and some have used expert readers and double reading (Swedish studies).

However, many of the conditions of screening trials resulted in significantly weaker contrasts than those of modern programs. In particular, all studies are from earlier eras when mammographic equipment and techniques were less refined. Many have used a single view of each breast, as opposed to modern standards using double views. Some studies have used intervals longer than two years (28 months in Stockholm, 33 months in Two-County). Participation rates have been as low as 53% (Edinburgh) and 54% (HIP) and as high as 87% (TCS) and 88% (NBSS), but in all studies the effective contrast has been reduced because women in the control group also received mammography, ranging from about 5% (HIP) to as high as 15% (Gothenburg) and 20% (Stockholm). The timing of the relationship between the screening period and mortality results has caused significant dilution in all studies, compared to the steady-state reduction that a program could achieve after a suitable delay. In particular, the durations of all studies have been much shorter than the 20 years of screening that most programs propose (from age 50 to 69); some studies had only two rounds (Stockholm), three rounds (Two-County) or four rounds (HIP, Edinburgh, NBSS) at only 12-months interval.

The Canadian NBSS-2 trial is not included in our meta-analysis, since the interventions contrasted in this trial are not relevant to answering the decisional question of the efficacy of screening mammography in Québec. Indeed, that trial did not compare screening with no screening, but rather the reference screening intervention (regular high-quality clinical breast examination and teaching of breast self-examination) with another screening regimen (the same exams, with added mammography).

SYNTHESIS OF RESULTS

Our review indicates marked differences in the quality of the design and execution of mammography trials, which have tended to fall short of modern quality standards, as indicated by their validity scores. Some trials were not randomized at all, and most studies have been poorly or inconsistently documented. In particular, most have not provided baseline characteristics of women in study and control groups, often because next to nothing was known about the control cohort. Exclusion of previously diagnosed cancers has been inconsistent in six of the eight studies. Blinding has not been attempted in any trial, either of patients or of care providers.

On the other hand, no study has been designed and conducted in such a way that the full potential of mammography screening could be determined, as indicated by their strength of contrast scores. To the extent that published trials have not maximized the potential of mammography screening, there is thus a potential for modern programs to identify earlier lesions and, perhaps, to achieve greater reductions in breast cancer mortality than those that have been reported in the scientific literature.

Our evaluation indicates that three trials (HIP, Edinburgh and Two-County) contain flaws which preclude their use in estimating the effectiveness of screening. These exclusions are consistent with those of other reviewers that have judged the quality of individual trials. An examination of only traditional aspects of study validity that takes no account of relevance or contrast issues would cast strong doubt on the efficacy of mammography screening. This meta-analysis has included successive study results by order of validity: good or medium quality (scores of 3 and over out of 4), poor quality (scores of 1.5 but less than 3), and flawed (scores less than 1.5).

For women of all ages, results show an inverse relation between the quality of the study and reduction in breast cancer mortality. If all studies are included, regardless of validity, mortality reduction is estimated at 23%, but this estimate diminishes to 15% when only studies of medium and poor quality are included, and to 9% if only medium quality studies are included. Thus, the more valid studies tend to show lesser reductions, and confidence intervals sometimes include the null value.[1]

In women under 50, many women were enrolled towards the end of their forties, and many of their cancers were not detected before they were in their fifties, so these results are pertinent to women who start screening in their late 40s, not at age 40. In the medium-quality studies of this age group, the cumulative risk reduction is 2%. A similar inverse relation is observed, since including data from studies with weaker validity increases mortality reduction to 8%. Mortality reduction is thus much smaller in younger women, and confidence intervals include the null value for all combinations of studies.

In women older than 50, on the other hand, results are more favourable. The only study of medium validity gives a risk reduction of 27%. Including data from studies of poor quality, the overall reduction would be 24%, and including all data irrespective of validity gives a reduction of 29%. These mortality reductions are substantially higher than in women of all ages combined. Confidence intervals in this sub-group are naturally wider, but no combination of studies includes the null value.

As for strength of contrast, no study came close to the standard of an ideal program with many years of regular screening using modern equipment, quality assurance, two-view mammography at intervals of two years or less, and with optimal participation. We estimate that, compared to this standard, the eight published mammography screening trials only achieved a strength of contrast of between 12 and 45% of what would be possible. Using this same standard, modern programs are likely to achieve considerably greater strength of contrast, with the Québec program estimated to obtain 63% of full potential.

DISCUSSION AND CONCLUSIONS

Question 1: What is the strength of the scientific evidence on which screening mammography programs are based?

There are serious concerns regarding the validity of most of the trials supporting mammography screening, based on methodological weaknesses in the screening trials. Studies are highly heterogeneous with

[1] The null value indicates absence of efficacy. A confidence interval that includes the null value indicates that the observed results would be likely to be observed by chance even if no true efficacy is present.

regard to the strength of the contrast that they studied, with numerous weaknesses identified in all the major studies, meaning that the potential of screening mammography has perhaps not been thoroughly explored. Using the best available data, one can conclude that there is fair evidence of moderate reduction of breast cancer mortality, of the order of 9 to 15%; data restricted to women over the age of 50 show greater reductions, of the order of 24 to 29%. Furthermore, our analysis has demonstrated that modern mammography, carried out under quality conditions that maximize its performance, has the potential to identify cancerous lesions earlier in their progression, and this may allow for some further reduction in mortality.

Conclusion: *Existing scientific trials, despite their flaws, support mammography screening programs. In addition, there are good reasons to believe that modern, well-conducted screening programs may achieve earlier detection and diagnosis of breast cancer and, perhaps, greater reductions in breast cancer mortality than what has been found in screening trials.*

Question 2: What is the evidence in support of screening mammography for women aged 40 to 49 years?

There is much less data available to answer the question, since most study experience is in women over 50, even though some women in some of the studies started screening several years earlier than their fiftieth birthday. The best data available show no significant reduction in breast cancer mortality in women screened before the age of 50. In the absence of any convincing data that mammography is efficacious in this age group, harmful effects may outweigh any positive effects.

Conclusion: *Trial data published to date do not provide scientific justification to recommend screening for women younger than 50. However, this conclusion does not exclude the possibility that screening of individual women, based on a personalized risk assessment, could be of benefit. These conclusions should be reviewed when results from the UK Trial become available.*

Question 3: What are the implications of research studies for maximizing the effectiveness of modern programs such as the *Programme québécois de dépistage du cancer du sein* (PQDCS)?

Although the PQDCS already includes rigorous control of the quality of films produced, certain aspects of the structure and process of trials examined under the rubric of strength of contrast can be transposed as additional quality norms. Notable among these are double reading of films and an annual reading volume sufficient to allow each radiologist to acquire and maintain the necessary expertise to detect breast cancer in its early stages. These aspects should also allow for a reduction in false positive rates and subsequent unnecessary diagnostic procedures. Moreover, high participation rates at each screening round will contribute to achieving and perhaps exceeding the mortality reductions obtained by screening trials.

Conclusion: *Modern screening programs such as the PQDCS may produce outcomes comparable or even superior to those observed in screening trials if they achieve a standard of quality equal to or better than the standard achieved by trials. Measures that should reduce false positive rates and assure high-quality screening include making sure that high-quality mammographic films are being produced, that readers have the necessary expertise to detect early cancer and avoid false positives, and double reading of a proportion of films. While participation rates should be as high as possible, efforts to increase participation should not overstate the benefits of mammography nor understate the risks and uncertainties which remain.*

TABLE OF CONTENTS

LISTE OF TABLES AND FIGURES

1 INTRODUCTION

1.1 STUDY REQUEST

The Québec Ministry of Health asked the Agence d'évaluation des technologies et des modes d'intervention en santé (AETMIS) to re-examine the pertinence of including in its mammography screening program women aged 40 to 49 years.

The Conseil d'évaluation des technologies de la santé (CETS, predecessor of AETMIS) had already published two reports on screening mammography. The first, published in November 1990, was commissioned with the objective of evaluating the evidence about the health effects of screening for breast cancer with a view to estimating the impact of a possible program in Québec. The report based its estimates on the results of four published randomized controlled trials (RCTs): the Health Insurance Plan Trial (HIP) from New York [Shapiro et al., 1985], the Malmö Mammographic Screening Trial [Andersson et al., 1988], the Swedish Two-County Trial [Tabár et al., 1985a] and the Edinburgh Randomised Trial of Screening for Breast Cancer [Roberts et al., 1984]. At that time, four further trials were already in progress: the two Canadian trials NBSS-1 and NBSS-2 [Miller et al., 1992a; 1992b], and Swedish trials in Stockholm [Frisell et al., 1986], and in Gothenburg [Andersson et al., 1983]. However, these more recent studies had not yet published results that could be used in the 1990 analysis. The first CETS report was based on data accumulating beyond the fifth year of each trial whose results were available at that time. It found that mortality from breast cancer was reduced by 35%, with a 95% confidence interval (CI) of 24% to 47%. When the analysis was restricted to women aged 50 to 69, the reduction was 45% (95% CI: 33% to 58%) [CETS, 1990]. Given the high number of mammograms already being performed in Québec in 1990, recommendations were made with a view to optimizing existing screening activities, particularly by directing activities towards the age group which offered the greatest benefits (women 50–69 years old) and by improving the quality of screening.

A second study on screening for breast cancer was published by the CETS in March 1993. The objective was to evaluate the possible benefits of mammographic screening for women aged 40 to 49. Since 1990, new data had become available from the four studies already examined, and three additional studies, that is, the two Canadian National Breast Screening Studies (NBSS-1 and NBSS-2) [Miller et al., 1992a; 1992b] and the Stockholm study [Frisell et al., 1991], had published preliminary results. Although only the NBSS-1 trial specifically examined the age group 40 to 49, data was available for women under 50 for all studies. Once again, the analysis was restricted to breast cancer deaths occurring after five years of follow-up, except for the Canadian study where this was not possible since only cumulative results were published. Pooling the results of the five most recent studies available at that time, the effect of screening on breast cancer mortality was estimated to be a 1% increase (95% CI: – 26% to +28%). It was concluded that mammographic screening of younger women had not been shown to reduce mortality; however, the evidence was considered to be inadequate to exclude the possibility of a benefit [CETS, 1993].

Since 1993, further follow-up information has accrued for the seven trials referenced in the 1993 report and new results have been published for one additional trial, conducted in Gothenburg, Sweden [Bjurstam et al., 2003]. This more recent evidence has, in general, continued to support the efficacy of mammography screening, particularly in the age group of women 50 to 69 years old. A further United Kingdom trial, undertaken in 1991, has not yet published findings [Moss, 1999].

1.2 SCREENING PROGRAMS

By the year 1998, Québec and the nine other Canadian provinces had all introduced screening programs, justified in large part by the proof of efficacy provided by the screening trials. In all these programs, women aged 50 to 69 are actively recruited by their physicians or by mailed invitations for mammography screening every two years. Younger women have access to screening mammography in the same centres, provided that it is prescribed by their physicians, but in most provinces, including Québec, they are not actively recruited [Health Canada, 2003].

1.3 CONTROVERSY REGARDING THE QUALITY OF MAMMOGRAPHY TRIALS

A recent review has challenged the general belief that mammography screening is an effective tool for reducing breast cancer deaths. Gøtzsche and Olsen, members of the Nordic Cochrane Collaboration Group, published the results of their systematic review of mammography screening in The Lancet in 2000 [Gøtzsche and Olsen, 2000] and their full Cochrane Collaboration report in 2001 [Olsen and Gøtzsche, 2001]. In these reports, they raised a number of concerns about the validity of the published randomized trials and the choice of breast cancer mortality as a measure of efficacy. They judged that of the seven[2] studies, two were flawed, three were of poor quality, and two were of medium quality. Including only the medium quality studies, they concluded that mammography screening had no significant effect on breast cancer mortality, with a risk ratio (RR) of 0.97, and 95% CI: 0.82 to 1.14, or on overall mortality (RR = 1.05; 95% CI: 0.83 to 1.33). For the poor-quality studies, they found a significant effect on breast cancer mortality (RR = 0.68; 95% CI: 0.58 to 0.78) but not on overall mortality (RR = 1.01; 95% CI: 0.98 to 1.04). They did not consider it appropriate to combine the results of medium- and poor-quality studies because of the heterogeneity and the methodological differences between the two groups.

An extensive debate has ensued concerning the merits of the concerns raised by Gøtzsche and Olsen, and a number of researchers have responded with corrections, refutations and further information regarding their studies. This debate has prompted national and international organizations responsible for clinical guidelines and screening programs to re-examine the evidence supporting their recommendations and activities.

1.4 THE NEW STUDY QUESTION

Since the CETS review in 1993, screening programs have been set up in those provinces which did not already have them. In addition, several research studies have published updated results, and there have been active scientific debates about the age at which screening should begin and about the validity of many of the screening studies and the value of screening mammography in general. Given these developments, the revised research questions addressed by this report are:

1) What is the strength of the scientific evidence on which screening mammography programs are based?
2) What is the evidence in support of screening mammography for women aged 40 to 49 years?
3) What are the implications of research studies for maximizing the effectiveness of modern programs such as the *Programme québécois de dépistage du cancer du sein* (PQDCS)?

We will first examine the methodological issues pertaining to mammography screening trials and their relevance to decision making about screening programs. The first two questions will be answered by

[2] These researchers counted the Canadian NBSS as one study. In this report, we refer to NBSS-1 and -2 as separate studies, since the interventions compared in these studies were materially different.

reviewing the published research trials, with an eye to how they apply to the Québec context. Although the experience of established screening programs is not comparative and cannot be used to assess the efficacy of screening, the comparison of their features with those of the trials allows for an assessment of the extent to which their results are likely to be replicated by individual screening programs.

In the sections that follow, we look at all the published mammography screening trials in light of these concerns, and in particular we examine their qualitative strengths and weaknesses. Finally, we discuss which of these research findings are particularly relevant to decision making, and how they might be applied in the Québec context.

2 METHODOLOGIC ANALYSIS

2.1 GENERAL CONCEPTS: RELEVANCE, VALIDITY AND PRECISION

This analysis will focus on three prerequisites that scientific evidence must satisfy in order for it to be the basis for informed decision making: its relevance, its validity and its precision. For a study to be **relevant**, it must be designed to contrast the intervention of interest with alternative interventions that seek to obtain the same results. For a study to be **valid**, it must be designed, conducted and analysed in such a way as to ensure that no important biases affect the measured comparison of the effectiveness of the strategies that are contrasted. Validity may be threatened by two sorts of bias: bias of known direction and bias of unknown direction. Finally, for a study to be **precise**, it must provide an estimate of effectiveness that is not vulnerable to random effects. Precise results have small standard error and are thus associated with narrow confidence intervals.

We believe that reviews of screening mammography trials have tended to focus primarily on the issue of validity, and in particular the issues associated with bias in unknown direction. They have thus closely examined validity issues such as randomization, baseline equivalence of screening and control cohorts, exclusion criteria, and equal follow-up for outcome. A secondary issue in these reviews has been the precision of the estimate of breast cancer mortality, since individual studies have had wide confidence intervals but combining studies has allowed reviewers to report narrower intervals. Amalgamating studies' results, a technique known as meta-analysis [DerSimonian and Laird,1986] has allowed reviewers to obtain narrower confidence intervals. However, the issues of relevance and validity, and in particular bias of known direction, have not been thoroughly addressed in studies and reviews of mammography.

For these reasons, the present analysis will focus on the conditions necessary to insure the relevance and validity in each mammography trial. Emphasis on these aspects allows for insights not only for the interpretation of the scientific literature on mammography screening but also for the design of a screening program. This analysis will be completed by a meta-analysis in order to address the issue of precision.

Study contrast

In order to better discern the conditions allowing for relevance and validity in a study, we introduce the concept of study contrast.

Definition: An experimental study's contrast is the opposition or temporary divergence between the experimental intervention, offered to the screened cohort, and reference strategy offered to the control cohort.

Since the main policy concern at issue is whether screening mammography is worthwhile or not, relevant studies will have a design which contrasts screening mammography and, ideally, no screening. In principle, our interest is not in studies which compare different screening strategies between themselves. In practice, the reference strategy (non-screening) may contain uncontrollable elements of screening, which would tend to weaken the contrast with the experimental intervention.

Strength of contrast

Although validity is always a major preoccupation of studies of screening mammography, reviews have tended to concentrate on the control of bias of unknown direction. This sort of bias may arise when screening and control cohorts do not have the same baseline risk of breast cancer mortality or are not treated equally in all regards except concerning the screening or control intervention. Such biases can also occur when information on their outcome is measured differently for the two groups.

Along with the examination of bias of unknown direction, the present analysis will further examine bias of known direction, with reference to the notion of strength of contrast.

Definition: A study's **strength of contrast** is the degree to which the difference between screening and control strategies is achieved and maintained and its effects adequately captured.

The main features of a strong contrast are the quality of the performance of the study intervention, the thoroughness of its application to the screening population and its absence among the control population, and the choice and analysis of results (measures and indicators) which fully reflect the different impacts of the two interventions.

A weak contrast will tend to bias a study's results in the direction of the null hypothesis, which is the hypothesis that screening and control interventions are of equal value. This concept is further developed in section 2.4, and is discussed in greater detail in Appendix A: Screening contrast and its duration, Appendix B: Timing of the effects of screening on breast cancer mortality, and Appendix C: Strength of contrast scale.

2.2 CHALLENGES FOR STUDIES OF SCREENING MAMMOGRAPHY

Three methodological challenges make it more difficult to study screening mammography than is the case for many other interventions. First, mammography seeks to identify a condition (breast cancer) which, untreated, typically has a long clinical course extending over many years. Whereas many cancers have fatal results within the first few years following their diagnosis (leading to the common use of 5-year survival statistics), breast cancer is often particularly indolent, and deaths often occur beyond 5 or even 10 years after diagnosis. This is all the more true for cases diagnosed early in their course, as is typical of screening-detected cases. The long delay between diagnosis and death means that follow-up over many years is necessary to ascertain the final results of an intervention that purports to reduce mortality.

Second, mammography screening is not a one-time event, but rather involves periodic screening over many years. Hence arise all sorts of variability in study or program design such as the interval between screens, and also complexity in defining concepts such as participation rates, average interval between screens, and total number of screening episodes, given the fact that different individuals can have widely different patterns of participation during the course of a given program or study.

Third, the duration of the contrast in screening studies is very different from the duration of the mammography programs that are proposed or in place in various jurisdictions. For the eight published mammography trials, the length of the screening contrast varied from 3.5 years in the New York study, up to 8.8 years in the Malmö study. On the other hand, for a mammography program, the duration of screening is typically 20 years if screening begins at age 50, and even longer if women are invited before age 50 or if screening continues after the age of 70. Of course, a trial with such a long contrast would require enormous resources, and the results would not be available before 30 to 50 years from study inception. Rather than compare a cohort of young women who are invited to a typical program of 20–30 years of screening, almost all trials have thus elected to study a shorter duration of contrast,[3] which aims to evaluate the hypothesis that screening mammography can reduce breast cancer mortality, and not how much a full 20 or

[3] The UK Age Trial is an interesting exception, and will be discussed in a later section.

30 year program might reduce mortality. This choice is practical, but the extrapolation of its results so that they will be relevant to decisions regarding much longer screening program has implications that will be addressed in the following section.

2.3 DESIGN OF THE CONTRAST

The first issue in assessing the relevance of a particular study of mammographic screening is to verify just what mammographic technique is being studied, and to what it is being compared. Different screening modalities have been proposed, evaluated, and integrated into screening programs. In particular, mammographic techniques, including the type of views (mediolateral, mediolateral oblique [MLO], craniocaudal [CC], etc.) have not been completely standardized. Other technical issues such as breast compression, technique of positioning the breast, timing during the menstrual cycle, dietary and cosmetic preparation, etc. have usually not been described in the published literature, but presumably vary from study to study.

In any trial, mammographic screening must be compared with some alternative strategy. In particular, it would be important to know the level of use of mammography (screening or diagnostic), clinical breast examination (CBE) and breast self-examination (BSE) in both screening and control cohorts, in order to understand what contrast has been achieved. Most studies have not described details of the alternative strategy, often referring only to ‘usual care’ or some variant. Since ‘usual care’ will vary over time and place, one can sometimes only guess what ‘usual care’ might have prevailed in each study, depending on local circumstances at that time.

2.4 STRENGTH OF CONTRAST

In order for a study to provide a valid measure of screening mammography's potential, it must obtain a strong contrast between screening and control interventions. Obviously, failure to do so would tend to bias a trial's results in the direction of the null hypothesis.[4] We discuss here the elements which contribute to the strength of the contrast between the study and control interventions, and which form the basis for developing a scale that will be used to measure each study's strength of contrast.

2.4.1 Technical contrast

The technical contrast is the opposition of two interventions, the study intervention and the control intervention. Studies of screening mammography include three principal technical contrasts: 1) invitation to screening mammography alone versus no invitation (all the Swedish studies); 2) invitation to screening mammography and clinical breast exam versus no invitation (HIP, Edinburgh, NBSS-1); and 3) invitation to mammography, clinical breast exam and breast self-exam versus invitation to clinical breast exam and breast self-exam without mammography (NBSS-2). While some technical contrasts may be considered not relevant to the policy questions at issue, one can also consider the relative strength of different technical contrasts which are deemed to be relevant. In particular, the comparison of mammography and clinical breast examination (CBE) versus no screening is a stronger contrast than the comparison of mammography alone versus no screening.

2.4.2 Era

Mammography screening was first introduced in the 1950s, using images produced by standard X-ray machines. The first dedicated mammography machine was developed in 1966 [GE Medical Systems, 2004], and the technology has continued to evolve since then. The first study (HIP) was

[4] Trials use hypothesis testing to confront two statistical hypotheses: the null hypothesis (there is no difference or association) and an alternative hypothesis (there is a difference or association). Here, the null hypothesis signifies that there is no difference in the observed effect (breast cancer mortality rate) between women in the screening and control cohorts.

initiated in 1963, and studies have begun in every decade since. In particular, some experts believe that major qualitative advances were made at the beginning of the 1980s. Sickles [1997], for example, has shown that prognostic indicators of modern mammography in clinical practice in the United States are superior to indicators that can be derived from the eight screening trials.

2.4.3 Quality of screening

Although the basic mammographic technique has remained largely unchanged since its development 40–50 years ago, refinements have been progressively introduced, so that compression techniques, film-screen contact, film resolution and contrast, exposure to radiation, development of films, display of films, and archiving, have all generally improved over time. Mammography is a technique whose quality depends on a long chain of human activities each of which must be expertly executed in order for its full potential to be realized. The production of technically superior films is a particular challenge, given the high resolution of mammography films compared to general radiological films. Expertise is required to distinguish the subtle differences between benign and malignant radiographic findings, if high rates of cancer detection are to be achieved while keeping the rate of false positive screens as low as possible. Other aspects of quality screening include quality control measures such as double reading of films, and requirements for the expertise of film readers. Norms of mammography practice increasingly favour two views for each breast rather than one. Some studies (TCS, Stockholm) have used one view only, and some (NBSS 1 and 2) have used other views than MLO, at least for some of the exams. Some programs have used one view as an initial screen, with a second view in the case of dense breasts or breast with any positive findings. The more recent protocols tend to use MLO and CC views for each breast. Finally, the interval between scheduled[5] rounds of screening should be short enough to detect most new lesions while they are still small. In published mammography trials, the suggested interval between mammographic exams has varied considerably, from 12 months (HIP, NBSS 1 and 2), to 18 months (Malmö, Gothenburg), 24 months (TCS, Edinburgh), and 28 months (Stockholm). For some studies, a range of intervals has been reported (Malmö, TCS).

2.4.4 Participation and contamination

In addition to these qualitatively different contrasts, studies differ in the extent to which this contrast was maintained quantitatively. Greater contrast is achieved by attaining a high rate of uptake of the recommended procedure by women in the screening group (participation, or compliance), and also by limiting the extent of use of that same technique by women in the control group (contamination).

In order to maintain validity, all studies have analyzed outcome by contrasting the outcome experience of those women allocated to screening to those allocated to the control group, regardless of whether women in the screening group receive screens or not, or whether women in the control group forego screening or not. This approach is generally called an intention-to-treat analysis. It is appropriate for maintaining validity, since it maintains the baseline equality of the two groups as randomized. However, this analysis comes at the cost of reducing the measured contrast between women screened and not screened. Because the ‘screened’ cohort of women has not all been screened, and many of the ‘control’ cohort have had some screening, overall trial results can, at best, only show

[5] Although an interval or a range of intervals are always proposed by the studies, observed intervals can be both shorter and, more commonly, longer, since study subjects have considerable freedom to participate when it is convenient for them. However, in this analysis, this individual variation will be considered under the heading of ‘participation’, leaving the term ‘interval’ to mean the suggested interval planned by the study’s designers.

the ‘average’ effect which can be expected in a population to whom screening is offered. If the purpose of the study is to anticipate the effect of introducing a screening program in a population that may already be receiving some screening, and which will participate in a sub-optimal way, then the intention-to-treat analysis will be not only a valid measure but also the relevant one.

However, for the individual women contemplating mammography screening, the intention-to-treat analysis will normally underestimate that effect. Clearly, a woman who does not participate in a screening program that is offered to her will derive no benefit from it. Just as clearly, a woman who does participate will derive a benefit that is greater than the average for the screening cohort, which includes non-participants.[6] Population efficacy results are a weighted average of the efficacy among participants (unknown) and the efficacy among non-participants (zero).

To achieve a fair comparison, the analysis of screening trials involves measuring outcome events in all women to whom screening was offered (intention-to-treat paradigm). However, a study which does not succeed in maintaining a high degree of contrast can not expect to demonstrate the full effect of screening, and results should be interpreted in this light. In addition, while it may be appropriate from a societal point of view to project the results of a given study in order to predict screening’s effect on a given population, these results do not apply to a given individual who does not participate (and thus receives no benefit from the offer of screening) or who participates in the optimal fashion (and thus should receive more benefit than the average member of the population invited to screening).

Glasziou [1992] has addressed this problem with respect to breast cancer screening and has proposed a formula for adjusting measured results.[7] While we have not made such a quantitative adjustment to results, it is consistent with our inclusion of sub-optimal participation as a concern in interpreting the strength of contrast achieved in a given study.

2.4.5 Timing of mortality effects (timing dilution)

A major issue in the assessment of screening studies is the choice of follow-up time during which the effects of a screening protocol are estimated. To assess the potential of a program to reduce mortality, it is essential to examine the data pertaining to deaths sufficiently distant in time from the onset of screening. This issue was treated at some length in the CETS 1990 report, where the analysis sought to determine the steady-state mortality reduction that a screening program might obtain after an initial period without effect. It is summarized by Miettinen et al. [2002] as follows: “…the reduced case-fatality rate presumed to prevail under screening results in fewer deaths from the cancer among the screened *only after an appropriate delay*, and not on entry into the trial; one needs to focus on deaths in the appropriate segment of follow-up—i.e., not too soon after study entry and not too late after discontinuation of screening.” [emphasis ours]

[6] A well-known example of this phenomenon is in contraception, where it is common to refer to a contraceptive’s ‘method’ efficacy versus ‘use’ efficacy. The former is the average protection afforded to a group of people who use the contraceptive as their method; it includes some who forget to use it or use it improperly. The latter is the protection afforded to those individuals who actually use the contraceptive, and is, not surprisingly, much higher. Mammography screening’s efficacy has always been reported as ‘method’ efficacy, but an individual woman’s ‘use’ efficacy is certain to be higher. This can also be seen as one example of the difference between ideal use (‘efficacy’) and the real life, sub-optimal efficacy (‘effectiveness’) that may apply in a given environment.

[7] Glasziou essentially proposes an approach (attributed to Newcombe) whereby the intention-to-treat estimate can be adjusted by a factor $1/\Delta$, where $\Delta = p_1 + p_2 - 1$, and p_1 and p_2 are compliance rates in the screened and control groups. Since p_2 is the compliance rate in the control group, $1 - p_2$ is the contamination rate, and $p_1 + p_2 - 1 = p_1 - (1 - p_2)$, so the adjustment factor can be restated as $1/(p - c)$, where p is the participation rate in the screening cohort and c the contamination rate in the control cohort. A similar adjustment for non-attendance and contamination has also been called a ‘causal’ estimate [Baker, 2002].

It is important to note that most studies have not reported breast cancer mortality rate reductions for the time period of maximal effect, but have rather reported the cumulative mortality experience from the first day after screening is started. Results of screening studies are typically given in terms of the ratio of the cumulative rates of breast cancer mortality in screened versus unscreened women. These cumulative rates are calculated by taking the total number of breast cancer deaths observed in each cohort over all follow-up time, i.e., from the onset of screening to the end of follow-up, and dividing by the number of person-years of observation. This practice is widely accepted: “disease-specific mortality in the two arms of the trial from the date of randomization to the end of follow-up is compared” [Paci and Alexander, 1997].

However, a screening program's effect on breast cancer mortality will only begin to appear five years or more after the beginning of screening. Counting deaths that occur in the first years of follow-up will dilute the apparent effect of screening ('early dilution'). As follow-up is extended farther, more and more of the relevant contrast (‘steady-state’ rate ratios) is included in the cumulative mortality rates. However, as follow-up continues beyond the zone of full mortality reduction, breast cancer deaths also begin to occur from cases that could not have been identified by screening during the study contrast, once again diluting the full effect of screening ('late dilution').

The effects of both early and late dilution can be remedied, although not eliminated, by examining only deaths occurring among women whose cases were diagnosed during the screening contrast. This strategy has been employed in the recent analyses of the Swedish studies, referred to by their authors as the ‘evaluation model’ [Larsson et al., 1997]. However, such an analysis can only be applied to trials in which a round of screening is offered to the control group at the conclusion of the study period. The control round of screening is needed in order to detect ‘latent’ cases, i.e., cases which are not yet symptomatic. Without such a control round, lead-time bias towards the null hypothesis would tend to arise, since cases are more likely to be detected early in the screened population [Nyström et al., 1993]. As a result, this analysis can only be applied to the TCS, Stockholm and Gothenburg trials where a control round of screening was carried out at the conclusion of the screening contrast.

The Conseil’s previous reports [CETS 1993; 1990] have attempted to correct the early dilution problem by examining only deaths from breast cancer occurring at least five years after the onset of screening, allowing for the elimination of the early years which are entirely dilutive.[8] The Swedish ‘evaluation’ model more thoroughly counters dilution, both early and late, by ignoring cases detected before the screening contrast or subsequent to it. We have extended this line of thinking by evaluating for each study the degree to which mortality reductions are likely to have been diluted by the analytic techniques used by the investigators. Further details regarding this approach can be found in Appendix A: Screening contrast and its duration, and in Appendix B: Timing of the effects of screening on breast cancer mortality.

2.4.6 Therapeutic contrast

Screening is effective if it succeeds in identifying lesions before they are detected otherwise, allowing for treatment which is more effective at an earlier stage. Thus far, we have concentrated on aspects of diagnostic performance that favour the early identification of breast cancer. However,

[8] This line of reasoning is not new. In reference to his HIP study, Shapiro [1977] explains “the series on mortality beyond year 5 following entry is limited at this time to deaths due to breast cancer among women with a diagnosis of breast cancer during the first five years. This approach reduces the effect of attenuation that would occur through the inclusion of mortality among breast cancer cases detected substantially after the cycle of screening examinations was completed. The reason is that the study group of women, in time, would return to the same status as control group women […] There is no completely satisfactory way to deal with this issue but to minimize the attenuation factor, cases diagnosed after year 5 are excluded.”

therapies introduced over the past 40 years, and in particular adjuvant chemotherapy, have dramatically improved treatment outcomes for breast cancer, and survival is now better at every stage of breast cancer. For instance, in the U.S., five-year survival for localized breast cancer has increased from 72% in the 1940s to 97% today [ACS, 2004]. These advances could either increase the usefulness of early diagnosis or, more likely, reduce it [Miller et al., 2000a]. It is thus possible that the advances associated with earlier radiological diagnosis have been diminished by better therapeutic management available for both early and later stages of cancer. While this report has not examined this issue in regard to screening trials and current programs, future evaluations may need to rely on models examining both the diagnostic stage advance obtained by screening and the relative survival rates based on current treatments at different stages of cancer.

2.4.7 Strength of contrast index

In order to compare the degree to which individual studies have been able to produce a strong contrast by their design, execution and analysis, we have developed a scale that allows for some replicable assessment of these factors. Its elements are the factors that we have discussed as pertinent to the assessment of study contrast, i.e., the technological contrast, the era in which mammography was conducted, the technical quality of the mammography performed, the participation rates, and the timing dilution implied by the method of analysis chosen. We have then applied this scale to each of the eight published mammography trials, in addition to an analogous validity scale discussed in section 2.5.5 and in Appendix D (Validity scale). Further details of this scale and its application to individual studies can be found in Appendix C (Strength of contrast scale).

2.5 BIAS OF UNKNOWN DIRECTION

The randomized controlled trial (RCT) was developed as a research tool to allow for a fair comparison between two (or more) interventions by applying the interventions to cohorts of individuals that are equal as for all characteristics with a bearing on the outcome of interest, apart from the intervention itself. Any significant difference in outcome observed in such cohorts could thus be logically attributed to the interventions and not to other factors. There is general consensus that the design features of randomized controlled trials make them least vulnerable to bias; in particular, when the object of interest is the intended effect of an intervention, there is great potential for bias when clinicians assign patients to the intervention based on its potential benefit for them (the indication) [Miettinen, 1983]. Systematic reviews of screening mammography have thus generally excluded from consideration the non-experimental studies that purport to evaluate efficacy.

Standards for designing, conducting and reporting RCTs have evolved since the landmark 1948 trial by Bradford Hill et al. [MRC, 1948], and these standards form the logical basis for evaluating the validity of the RCTs of screening mammography. In the preceding section, we have discussed aspects of the screening contrast that contribute to the strength of the contrast, and clearly any weakness in the contrast will tend to bias the measurement of efficacy differences towards zero. We examine here features of randomized controlled trials that are designed to allow for a fair comparison, such as randomization. In the absence of these features, we may not have a fair comparison; however, it is difficult to know whether any resulting bias would be in the direction of the null hypothesis or away from it. These factors, discussed below, can thus be characterized as validity factors of unknown direction.

2.5.1 Random allocation

Random allocation of patients to the screening group or to the control group should in principle ensure that all other risk factors for breast cancer mortality, both known and unknown, are similarly distributed in both groups. Random allocation has become standard because the unpredictability of random assignment makes it difficult for a patient or physician to effectively select one of the interventions [Chalmers, 2001]. Other schemes, such as assigning patients based on the day of clinic visits, birthdays, etc., are more vulnerable to this sort of selection. They cannot properly be labelled RCTs, although they are often referred to by this name [Frisell et al., 1986; Tabár et al., 1985a]. Randomization may also be applied to groups of individuals [Roberts et al., 1984], although this has implications for the analysis of results.

2.5.2 Baseline comparability

Since random allocation of women to screening and control groups is undertaken in order to create cohorts with equal distributions of risk factors for breast cancer death, the description of the distribution of these risk factors in the groups created by randomization allows for verification of its success, at least for known risk factors. The reporting of such a descriptive analysis of the baseline comparability of groups is one of the quality criteria proposed by Chalmers et al. [1981]. Parameters of particular importance in a mammography trial are the known and suspected risk factors for breast cancer death, including age at randomization, number of previous live births, menopausal status, and family history of breast cancer.

2.5.3 Treatment of exclusions

In order to maintain the equality of the cohorts, any exclusion subsequent to randomization must be applied equally to screening and control groups. A common and important exclusion in breast cancer screening trials is the previous diagnosis of breast cancer. Since approximately one third of women with breast cancer eventually die of the disease, unequal application of this exclusion criterion has the potential to create an enormous bias, typically in favour of screening. For instance, if women in the control group are not interviewed using the same protocol as women assigned to mammography, woman with unrecognized cases may be included in the study comparison. Subsequent exclusion, for instance based on hospital records, puts too much reliance on hospital records that may not include some of the information obtained by personal interview.

2.5.4 Equal information about outcome

To maintain a fair comparison, the same accuracy of information must be obtained regarding the outcome of women in both groups. This may not be the case when diagnoses among screened women are established by virtue of procedures practised within the screening program, and if similar diagnoses are missed in women belonging to the control group. This bias, which would be expected to attenuate the measure of efficacy of screening, can be partially corrected by validating end-points by the use of blind assessments of cause of death, and by validation of assigned causes of death using autopsies.

Misclassification of the cause of death is a serious threat to validity in screening studies, for several reasons. First, women dying with breast cancer are often at an age where other illnesses, including other cancers, are present and could also be given as causes of death. Autopsy may help to distinguish between competing causes, but rates of autopsy have been falling over the last 50 years and were typically between 30% and 40% in the published studies. In addition, complications of treatment of diagnosed cases are not necessarily assigned to breast cancer, and this underestimation of true breast cancer-related mortality, which may be about 9% according to one estimate

[Brown et al., 1993], may tend to favour mammography screening.

2.5.5 Validity scale

Systematic reviews of any intervention must determine criteria for the inclusion or exclusion of studies. The construction of an index allows for a replicable assessment of the degree to which individual studies respect criteria of validity. We have discussed factors that are commonly considered in the assessment of the validity of screening trials. Other reviewers have developed scales for measuring validity, but these are difficult to apply to screening mammography. For example, Jadad's scale assigns 2 points for blinding, out of a total of 5, but no screening mammography trial has involved blinding either of patients or of care providers [Jadad et al., 1996]. Our rating scale, like others, involves the subjective assessment of the degree to which these qualitative features of a well-designed trial are present. Further details of this scale and its application to individual studies can be found in Appendix D (Validity scale).

2.6 MINIMIZING THE EFFECTS OF CHANCE

The technique of meta-analysis was initially developed based on the work of Gauss and Laplace with the object of reducing 'errors of observations' by the combination of observations, first in astronomy, and subsequently in other fields such as medicine [Chalmers, 2001]. Thus trials that are too small to show statistically significant results might, by their aggregation, provide greater precision. Although screening mammography trials have all involved thousands of women in each cohort, only a small percentage develop breast cancer, and an even smaller number die of their disease, typically less than a hundred in each cohort. The problem is even more serious for sub-groups, for instance, for estimating the efficacy of mammography in younger women.

However, in comparison with the objectives of examining relevant contrasts and providing valid estimates of effect, the quest for precise estimates with narrow confidence intervals is clearly of secondary importance. In other words, a valid but imprecise result is highly preferable to a precise but invalid one. There is no need to develop an index for the precision of studies, since the standard error serves as a natural and universally available measure of imprecision. Standard techniques of synthesis using the inverse of the standard error as a weight allow for amalgamating the results of different studies, if such an estimate is desired. Unlike in the case of the studies with weak contrast or invalid results, there is no advantage in excluding imprecise results, since they contribute to strengthening the precision of the combined results, and not weakening it.

2.7 OUTCOME INDICATOR RELEVANT TO THE RESEARCH QUESTION

The aim of breast cancer screening is to reduce breast cancer morbidity and mortality without increasing the incidence of other causes of morbidity and mortality. This would also, of course, imply a reduction in overall morbidity and mortality. The eight mammography screening trials for which we have results have all compared breast cancer mortality rates between women invited to screening and women not invited. Several also reported data for other causes of death and for overall mortality.

Although screening mammography aims to reduce mortality by breast cancer, it is important to ascertain the extent of unwanted health effects arising from screening. A previous report [CETS, 1993] mentioned two of these: false positive tests, leading to needle and surgical biopsies, and radiation risk, due to the dose of radiation absorbed by women undergoing mammography. This list is far from exhaustive: in particular, screening may lead to other harms, including anxiety after positive screens [Lampic et al., 2001], diagnostic procedures arising from screening (e.g., surgical or anaesthetic misadventure), or therapeutic procedures arising from definitive diagnosis (e.g., side-effects of chemotherapy or radiotherapy for lesions identified by screening).

The meta-analysis published by Gøtzsche and Olsen [2000] has revived the old debate about whether breast cancer mortality is the appropriate indicator for measuring the performance of screening mammography. Worried that mammography may cause as many deaths from other causes as it saves from breast cancer, they claim that overall mortality is a more appropriate indicator. The use of disease-specific mortality introduces ambiguity, since deaths resulting from the screening process itself or subsequent treatment to target disease are generally not breast cancer deaths, although as Black et al. [2002] point out, "the actual rules used to determine which deaths count for disease-specific mortality are rarely published with trial results."

Obviously, reducing breast cancer mortality should not be pursued without regard to other causes of death. From the practical perspective of deciding whether to avail oneself of screening mammography, one should consider both intended effects of screening (principally, reduction of breast cancer mortality), and also unintended effects (principally, side effects of aggressive diagnosis and treatment, but also other unexpected effects such as radiation risk [Jung, 2001]). However, mixing these two categories by using overall mortality as the primary indicator of outcome clouds the issue, because death from breast cancer represents only a small proportion of overall mortality, and looking at overall mortality would require studies of prohibitive size if the requisite power is to be attained. It is better to consider two separate questions: 1) Can breast cancer mortality be reduced by regular long-term screening? and 2) Are there other causes of death with plausible links to screening (e.g., operative misadventures, cardiac deaths post-radiotherapy, suicide, etc.) which have higher rates in screened or unscreened women? In this case, the suitable interval in which to study effects may be from the moment screening begins.

Published studies have largely restricted themselves to assessing the first question, regarding breast cancer mortality only. Without question, the specific effect of mammography screening on breast cancer mortality constitutes information that should be useful for women who are considering undergoing screening mammography. However, quantitative answers to the second question about side effects are not presently available. In the following sections, we shall examine more closely the evidence of screening's efficacy[9] using breast cancer mortality as its indicator. It should, however, be kept in mind that lack of evidence regarding the harm that screening mammography might cause is not the same as evidence of lack of such harm.

[9] In this review we will generally use the term 'efficacy' to refer to the relative reduction of breast cancer mortality rates in the screened group with respect to the control group. The term 'effectiveness' is sometimes used to refer to the efficacy as it might be observed in 'real-world' situations [Last, 2001] as opposed to the efficacy measured in the context of a trial. Since we explicitly draw attention to these 'real-world' differences by reference to the notion of the strength of a study's contrast, we will not distinguish further between these two terms.

3 ANALYSIS OF PUBLISHED TRIALS AND COMPARISON TO THE QUÉBEC CONTEXT

In this section we will consider the published evidence regarding the efficacy of mammographic screening programs, with a view to assessing the strength of their contrasts and their validity. To this end, we will apply our rating scales for strength of contrast and validity to each of the eight published screening mammography trials, as well as partially rating the current UK Age Trial with the limited data available. Details on the rating scales and their application to these studies can be found in Appendix C (Strength of contrast scale), Appendix D (Validity scale), and Appendix E (Review of the nine trials). The issue of precision will be addressed in the discussion (section 5), where we will discuss the appropriateness of statistically combining measures of studies that are found to be valid and have strong contrasts. In particular, we will refer to design features of the various mammography trials and the extent to which they resemble or differ from current screening in Québec, which takes place largely within a systematic breast cancer screening program known as the *Programme québécois de dépistage du cancer du sein* (PQDCS). We begin with a discussion of Québec's program (section 3.1), and then describe the nine mammography trials' structure and process (section 3.2) in order to allow for an understanding of how the research trials resemble the Québec program or differ from it (section 3.3).

3.1 SCREENING IN QUÉBEC

Mammography became routine medical practice in Québec in the 1970s, and has been provided free of charge under universal provincial health insurance. Use of mammography has risen slowly: in 1987, only 44% of women 50 to 59 years old had ever had a mammogram, 24% in the preceding 24 months. By 1997–1998, 56% of women aged 50 to 69 had had at least one mammogram, although it is difficult to know what proportion of these exams represented screening and what proportion were diagnostic studies for the work-up of signs and symptoms of breast pathology. The PQDCS, launched in 1998, actively invites all Québec women between the ages 50 to 69 with no previous diagnosis of breast cancer to have a mammogram every two years. Screening among women younger than 50 or who are 70 and over, while not formally part of the PQDCS, is conducted by program facilities, upon prescription by a woman's physician. The program's structure, process and objectives are laid out in a reference document [MSSS, 1996] and are loosely based on established breast cancer screening programs in place in Sweden, the United Kingdom and Australia.

Screening takes place in approximately 80 screening centres, largely private clinics in urban areas and hospital departments in regions with smaller populations. These centres must meet province-wide quality assurance standards, including periodic inspection and measurement of equipment, professional qualifications, use of the programme's information technology and participation in the evaluation of performance. Screening centres do initial screening (two-view mammography) and, usually, additional radiological procedures as required. Each region also designates a smaller number of diagnostic centres, to which women are to be referred if their screening exam is positive. In these centres, the diagnostic work-up is completed and, if cancer is discovered, treatment may be initiated. Service coordination takes place at a regional level, including inviting women to screening by mail, sending them results of screening exams, and following up on positive screens. Regional centres also coordinate the use of the information system, evaluation and quality assurance. At a provincial level, technical support, information technology maintenance and development, and liaison with professional bodies and ministry policies are handled by teams based at the Ministère de la Santé et des Services sociaux and the Institut national de santé publique (INSPQ).

3.2 RESEARCH TRIALS

Nine published trials using screening mammography have been conducted in the past 40 years. As discussed in section 2, the issues relevant to this evaluation are the relevance of the contrast that has been studied by each trial, the strength of this contrast, and the absence of bias of unknown direction. In this section, salient features of these elements of the research trials will be noted and compared to their counterpart in Québec, in order to allow for later judgments as to whether the trials' results apply to the Québec context. A more detailed description of these trials is presented in Appendix E.

3.2.1 New York (HIP)

The first formal mammography trial, generally referred to as the HIP Mammography study, was conducted among clients of a private health insurance company of Greater New York, the Health Insurance Plan (HIP) beginning in 1963. Screening consisted of three or four rounds of mammography at one-year intervals, along with a clinical breast exam, while women in the control group were to make no changes in their health care practices. Using our rating scale for validity, the study scored 1.5 out of 4. This trial shared the major defect of most studies in that almost no information about the characteristics of the control cohort were recorded; in addition, the quality of follow-up information was very poor, since rapid attrition from this private health plan and the paucity of health information on people that had left the plan made it difficult to exclude cases of breast cancer with diagnosis prior to the screening contrast.

The study scored 12% on our strength of contrast scale, the lowest of all published trials. Since the HIP trial was begun before dedicated mammographic equipment was available, there are major differences from the Québec context 40 years later. On the other hand, quality assurance measures for the reading of films were considerably stronger than is the case in the Québec program. In particular, reading of films was centralized to several radiologists, and all films, both positive and negative, were double read [Strax et al., 1973].

3.2.2 Malmö

The Malmö Mammographic Screening Trial (MMST) was conducted in Malmö beginning in 1979, and is the first of four Swedish studies comparing mammography alone to usual care. In this trial, control patients were not screened but those with breast complaints ‘attended the ordinary medical service’.

Using our rating scale for validity, the trial received a score of 3 out of 4, the highest score apart from the Canadian studies. It shared, along with all other studies except the Canadian ones, an almost total lack of information about the control population, making it difficult to ascertain whether randomization was successful in creating equal groups and whether cases of breast cancer diagnosed prior to screening were properly excluded.

Using the rating scale for strength of contrast, this study scored 27%. The Malmö study differed significantly from the Québec program with respect to reader expertise. As for all Swedish studies, screening in the Malmö study was conducted by a highly centralized service whereby films taken in fixed and mobile units were interpreted by readers with annual volumes of 5,000 to 20,000 films per year. This is in contrast to the Québec program where the median volume is approximately 600 films per year, and only 20% of films are read by radiologists with a volume of over 2,000 films per year [INSPQ, 2003]. In addition, Swedish readers are responsible for the diagnostic work-up of the cases whose screening films they have read as positive, whereas the Québec program recommends that women with positive screens be referred to designated diagnostic centres which are generally distinct from the screening centres.

3.2.3 Two-County Study (TCS)

The Two-County Study (TCS), the second of four trials conducted in Sweden, began in Kopparberg county in 1977 and in Östergötland county in 1978, two counties in central south-east Sweden representing together about 8% of the total Swedish population. It was planned as one study, and indeed the same screening protocol was used in Kopparberg and in Östergötland. Validity issues included conflicting and vague information in the large number of reports in scientific journals, a complicated and sometimes arbitrary cluster randomization process, poorly documented exclusion criteria, and baseline age inequality between study and control cohorts. Using our validity scale, the study scored 0.5 out of 4.

The contrast studied was between single-view mammographic screening, without clinical breast examination, versus usual care. Also like the other Swedish studies, it used a combination of one central unit and several mobile units, with central reading of all films, in this case by the two radiologists who were the principal researchers (Fagerberg and Tabár). Average participation was approximately 87%, with moderately low contamination. The screening interval was longer than the Québec program's, two years for women 40 to 49 years old at entry and three years for women 50 to 69 years old. The contrast was of short duration; although accounts differ somewhat, it seems that a five-year intervention period was planned, although the control cohort may have only been screened as long as eight years after randomization. Full steady-state mortality reduction cannot be observed for a trial with such a short screening contrast, however, since the trial was analyzed using the Swedish evaluation model, where only women diagnosed with cancer during the study contrast were followed for mortality results, and since exceptionally long follow-up is available (20 years), the dilutional effect of including early years of follow-up in cumulative mortality figures is less important than in other studies with similarly short contrasts. Using the rating scale for strength of contrast, we gave this study a score of 42%.

3.2.4 Edinburgh

The Edinburgh Randomised Trial of Breast Cancer Screening began in 1979. Validity concerns include cluster randomization based on the clientele of general practitioners in Edinburgh, adjusted in various ways by the investigators, markedly unequal exclusion of previously diagnosed breast cancer cases, significant baseline differences in socio-economic status and all-cause mortality rates. We rated this study 0.5 out of 4 for validity.

The screening contrast involved an invitation to a series of four rounds of mammography accompanied by clinical breast examination (CBE) every two years, with CBE alone in the intervening years; mammography was two-view in the first round and one-view subsequently. Different from other studies and from the Québec program, most films were not read by radiologists, but rather by 'specially trained doctors'; radiologists read all abnormal films and a random 5% of all films. Participation rates were low, with an average rate of 52%; contamination rates were likely low but were not measured. Using the rating scale for strength of contrast, we gave this study a score of 26%.

3.2.5 Canada-1 (NBSS-1)

The Canadian National Breast Screening Study (NBSS) was begun in 1980 according to two different protocols defined for women with age at entry of 40–49 years (NBSS-1) and 50–59 years (NBSS-2). Both studies have been criticized on a number of grounds, some concerning validity, some concerning contrast.

As regards validity, an independent review of the randomization process concluded that its execution was successful [Bailar and MacMahon, 1997]. Cases identified by the initial examination were not excluded from follow-up or analysis, and post-randomization exclusions were balanced between groups. Baseline equality between groups was achieved regarding ten factors of prognostic importance [Miller et al., 2000b; Baines, 1994]. Although women in the screened group had more small node-positive cancers, this is consistent with the fact that mammography is expected to identify such tumours [Bailar and MacMahon, 1997]. Few studies have been the subject of such intense scrutiny, and fewer still have had the benefit of such an independent and thorough review, which has repudiated all significant criticisms regarding validity. Using our rating scale for validity, we gave this trial a score of 4 out of 4.

As regards the screening contrast, NBSS-1's control intervention can be summarized as baseline physical examination and teaching of breast self-examination (BSE), with mammography only as needed in the context of diagnosis, versus the screening intervention of added mammography in round 0, and additional mammography, physical examination and BSE reinforcement in subsequent rounds. The technical contrast was thus stronger than the contrast in the Québec and most other Canadian and European programs, where mammography is offered without systematic physical examination. Some reviewers have criticized this study because participation was 'volunteer-based' rather than 'population-based'. We feel that the only important consequence of this strategy is the high participation rates (92%). Although this feature in itself tends to provide a stronger contrast, this is partly countered by other features that weaken the contrast, in particular its short duration of 4.6 years. Using the rating scale for strength of contrast, this study received the highest score among mammography trials at 45%.

3.2.6 Canada-2 (NBSS-2)

The second part of the Canadian National Breast Screening Study (NBSS), concerning women 50 to 59 years old at entry, also began in January 1980. Validity concerns are identical to those discussed in regard to NBSS-1, and likewise, using our rating scale for validity, we also gave this trial a score of 4 out of 4.

As for screening contrast, different from all other studies and from the Québec program, NBSS-2 involved a contrast between two screening interventions. The first screening intervention involved systematic physical examination by trained nurses and physicians, along with training and reinforcement of breast self-examination. The second screening intervention involved these as well, along with annual two-view mammography. This contrast is substantially weaker, since women in the 'control' group received a screening regimen (physical examination and breast self-examination) that may have some efficacy in itself. As in NBSS-1, participation rates were high, with 91% in the screening group, and the duration of the contrast was short, with and average of 4.6 years. Using the rating scale for strength of contrast, we gave this trial a score of 31%.

3.2.7 Stockholm

Third of four Swedish studies, the Stockholm Mammographic Screening Trial began in 1981. Assignment of patients was not random but rather according to the day of birth, making the study vulnerable to self-selection of patients based on their risk of developing breast cancer. Other validity concerns include inappropriate rounding of quantitative information, lack of data on the baseline comparability of screening and control groups, and late exclusion of cases of breast cancer with diagnosis prior to the study. Using our rating scale for validity, we scored this trial as 2 out of 4.

The screening contrast was extremely short, consisting of only two rounds of single-view (oblique) mammography, and the (single) interval was long at 28 months. Both these features, and the fact that the mammography was single-view for both rounds, make the research contrast weaker than the Québec program's. In addition, although participation in the first round was relatively high at 80%, contamination was also relatively high, since 25% of all Stockholm women had mammography in the three years before the trial [Frisell et al., 1991]. Using the rating scale for strength of contrast, we gave this study a score of 28%.

3.2.8 Gothenburg

The Gothenburg Breast Screening Trial, initiated in 1983, is the most recent of the eight studies with published results. Validity concerns are similar to those with the other Swedish studies, and include fragmentary information about the control cohort, and late exclusion of breast cancer cases diagnosed prior to screening. We rated this trial 1.5 out of 4 on our validity scale.

The screening contrast intervention consisted of up to five rounds of screening mammograms at planned intervals of 18 months between rounds for the screening group, versus no screening for the control group except for a closing round of screening contemporaneously with the screening group. Of note, the contrast was one round shorter for women 50 and over, since screening had become routine procedure for this age group. As in the Québec program, two views were used (for most exams), and clinical breast examination was not offered. Participation and contamination rates were high, approximately 79%, and 19% respectively. Using the rating scale for strength of contrast, we gave this trial a score of 37%.

3.2.9 United Kingdom (UK Age Trial)

In 1991, the United Kingdom Coordinating Committee on Cancer Research set up a national multicentre randomized controlled trial that they refer to as the 'Age trial' [Moss et al., 2005a]. In this analysis, we will briefly describe the trial as an indication of how it may compare with the others once more results are known. Although some methodological details have not been published, it is likely that this trial will be among the strongest as far as validity, using modern techniques of randomization, measurement of baseline characteristics, and follow-up of results are concerned. Using our rating scale for validity, and making some conjectures which will need to be confirmed when more details are published, we gave this trial a potential score of 4 out of 4, based on its published design.

As for the strength of the screening contrast, this trial will have carried out the contrast of longest duration, with up to nine rounds at 12-month intervals. Indeed, given the fact that it seeks to evaluate the validity of screening women from the age of 40 to 41 for a full ten years, this will be the only study design where the contrast is of the same duration as the program that may eventually be put in place. In other words, the study evaluates the full course of a potential program, instead of evaluating several rounds of screening and extrapolating their results. Of concern will be contamination by screening among the control cohort, which will be difficult to avoid in today's environment where breast screening is widely recommended. Finally, the structure of the UK program differs significantly from the Québec program, in that it is a highly centralized, hierarchical program with emphasis not only on film quality but also on readers' expertise, second reading of films, etc. Using the rating scale for strength of contrast, and making some guesses about how this study is likely to be conducted and analyzed, we project that the strength of contrast score would be 43% at that time; further follow-up will likely give this study the highest strength of contrast among trials of screening mammography.

Tables 1 and 2 summarize the salient features of the nine screening mammography trials, as well as the validity and strength of contrast ratings using our scales. The first table summarizes major design elements of the trials' protocols.

Table 2 summarizes scores assigned to the mammography trials for the features which are important in limiting bias of unknown direction. Although all the trials refer to themselves as randomized, many were not properly randomized, and it is not surprising that baseline equality of the screening and control populations could not be established. Many trials were unable to adequately exclude cases of previously diagnosed breast cancer or to obtain full follow-up of mortality results in each group.

Finally, Table 3 summarizes the scores assigned to features we have examined which relate to trials' strength of contrast. Structure and process issues such as the technical contrast, technological era and quality assurance tended to be relatively strong, whereas participation and timing issues had marked effects in weakening most of the trials' contrasts.

TABLE 1

Summary description of 9 trials

TRIAL, INITIATION YEAR	AGE AT ENTRY	SCREENING GROUP				CONTROL GROUP		STUDY DURATION (YEARS)
		MODALITY	PARTICIPATION (INITIAL, SUBSEQUENT ROUNDS) (%)	INTERVAL (MONTHS)	AVG # OF ROUNDS	MODALITY	PARTICIPATION (ESTIMATED CONTAMINATION) (%)	
HIP (New York), 1963	40–64	2 view, + CBE	65, 52, 48, 45	12	4	‘Usual practices’	5	3.5
MMST (Malmö), 1976	45–69	2/1 view (CC+O / later just O)	74, 70 subseq.	18–24	6	‘Ordinary medical service’	10	8.8
TCS (Two-County), 1977	40–74	1 view (MLO)	91, 86, 84	24–33	2.5	Usual care	10	6.5
Edinburgh, 1979	45–64	2/1 view (CC+O / later just O)	61, N/A, N/A, 44	24	4	Usual care	10	6.4
NBSS-1 (Canada), 1980	40–49	2 view, CBE, BSE	100, 89.4, N/A, N/A, 85.6	12	4.5	Initial CBE, then ‘usual care’	12.5	4.6
NBSS-2 (Canada), 1980	50–59	2 view, CBE, BSE	100, 90.4, N/A, N/A, 86.7	12	4.5	CBE, BSE	6.5	4.6
Stockholm, 1981	40–64	1 view	81, 80	28	2	Usual care	20	4.6
Gothenburg, 1982	39–59	2 then 1 view	85, 78, 79, 77, 75	18	5	Usual care	15	6.4
UK Age Trial, 1991	40–41	2 then 1 view	61	12	9	No mammography	4	11.0

BSE = breast self-examination; CBE = clinical breast examination; CC = craniocaudal; MLO = mediolateral oblique; N/A = not available; O = oblique.

TABLE 2

Validity ratings

TRIAL, INITIATION YEAR	RANDOMIZATION	BASELINE EQUALITY	EXCLUSIONS	FOLLOW-UP	TOTAL (SUM), OUT OF 4
HIP (New York), 1963	1	0.5	0	0	1.5
MMST (Malmö), 1976	1	0.5	0.5	1	3
TCS (Two-County), 1977	0	0	0	0.5	0.5
Edinburgh, 1979	0	0	0	0.5	0.5
NBSS-1 (Canada), 1980	1	1	1	1	4
NBSS-2 (Canada), 1980	1	1	1	1	4
Stockholm, 1981	0.5	0.5	0.5	0.5	2
Gothenburg, 1982	0.5	0	0.5	0.5	1.5
UK Age Trial, 1991	1	1	1	1	4

TABLE 3

Strength of contrast ratings

TRIAL	TECHNICAL CONTRAST (C)	ERA (E)	QUALITY (Q)	STRUCTURE AND PROCESS SUB-INDEX (CxExQ)	PARTICIPATION (P)	TIMING (T)	TOTAL (CxExQxPxT)
HIP (New York)	1.25	0.70	1.00	0.88	0.49	0.28	0.12
MMST (Malmö)	1.00	0.80	0.85	0.68	0.61	0.64	0.27
TCS (Two-County)	1.00	0.80	0.85	0.68	0.82	0.76	0.42
Edinburgh	1.25	0.80	0.90	0.90	0.43	0.67	0.26
NBSS-1 (Canada)	1.25	0.90	0.95	1.07	0.75	0.56	0.45
NBSS-2 (Canada)	0.80	0.90	0.95	0.68	0.82	0.56	0.31
Stockholm	1.00	0.90	0.80	0.72	0.60	0.65	0.28
Gothenburg	1.00	0.90	0.88	0.79	0.64	0.73	0.37
UK Age Trial*	1.00	1.00	0.98	0.98	0.57	0.73	0.41

* Projected results after 13 years follow-up

3.3 COMPARISON OF TRIALS' STRUCTURE AND PROCESS TO QUÉBEC CONTEXT

In order to assess the value of a program of mammography screening, we would wish to examine the breast cancer mortality experience of a cohort of Québec women aged 50 at the outset, to whom screening would be offered every two years for a total of 20 years. This experience would be compared with that of a similar cohort receiving no screening. However, it is impossible to study this contrast for a number of reasons. First, a population-based screening program is not a research study, and has no control cohort with which to compare. Nor would the constitution of such a cohort have been feasible or ethically admissible, given the current clinical consensus regarding the efficacy of screening mammography. Second, as discussed previously, such a trial would take 30–40 years to conduct, and its results would not likely be relevant to policy-making after such a long delay. Third, it is not feasible to construct a cohort sufficiently similar to the screened population, either historically or internationally. There are too many differences in the different populations' distribution of risk factors for breast cancer (age at menarche, number of pregnancies, breast-feeding, hormone use, etc.) and rates of mortality (access to diagnostic services, treatment protocols, etc.) that prevent meaningful non-experimental comparisons with other populations. Although the PQDCS intends to evaluate its performance based in part on its success in achieving breast cancer mortality reduction of 25% over 10 years [MSSS, 1996], any such reduction could not be attributed solely to screening. For example, changes in treatment efficacy may be decreasing rates of breast cancer mortality regardless of screening [Miller et al., 2002].

Rather, we propose here to compare the contrast achieved in the context of published studies with the strength of contrast in place in the Québec program. To the extent that the Québec program achieves a contrast similar to those obtained in research studies, it can hope to also attain similar results. This is analogous to the notion of 'external validity' of these trials, except that we do not make general judgments of the external validity of the trials but rather, inversely, from the Québec standpoint, judge the pertinence of each trial's results to the Québec context.

3.3.1 Structure and process

We consider the same constituent elements of strength of contrast that we have applied to the trials examined in this analysis.

Technical contrast: As far as technical contrast goes, the Québec program is like most of the studies examined in that it proposes screening mammography without clinical breast exam or breast self-examination. It would thus score 1.0 on this aspect.

Era: Mammography equipment is generally modern, compared to the equipment used in studies over the past 40 years, so the program scores 1 on this aspect.

Quality of mammography: The PQDCS scores well on some aspects of quality and poorly on others. Quality assurance is applied rigorously in that all screening centres must be accredited with the Canadian Association of Radiologists, with criteria that test the ability to produce high-quality films. However, as far as reading of films is concerned, few standards are applied. Thus, whereas most programs and studies have ensured that the appropriate expertise is present by setting volume standards for mammography reading, typically a minimum of 2,000 to 5,000 films per year, Québec program radiologists need only read 480 films each year. Reader volume is generally well below that which applied in most research studies, with 45% of films read by readers of less than 1,000 screening films in 2002, with 36% read by readers of 1,000 to 1,999 films, and 19% by readers of 2,000 films or more [INSPQ, 2003]. Likewise, while other programs and studies have often included double reading for all positive films and a proportion of negative films, the Québec program includes no systematic double reading of positive or negative films. The Québec program uses two-view mammography systematically, whereas many trials have used single-view mammography or a combination of single- and double-view. The Québec program's interval of 24 months gives it a score of 1, in line with the average screening study, since screening trials employed intervals ranging from 12 to 33 months. Taken together, these three contrast elements, which might be considered elements of screening's structure and process, obtain a combined score of 0.90, comparing quite favourably with trials' scores that ranged from 0.68 to 1.07. Table 4 summarizes these features in the trials and in the current Québec context.

3.3.2 Participation

Québec's program began in 1998 when rates of screening were already much higher than was the case in the screening trials. In the 24-month period between January 1997 and December 1998, 56% of women aged 50 to 69 had at least one screening or diagnostic mammogram. For the most recent two-year period for which data are available, between 2001 and 2002, these rates had risen to 63% in women 50 to 69 years old. Several adjustments to these figures are necessary. First, given the fact that the actual procedure involved in screening and diagnostic mammography is identical, and that less restrictive billing conditions apply to diagnostic mammography, some mammograms billed as diagnostic mammograms actually represent mammography done for the purpose of screening. Second, as screening rates rise, there should eventually be somewhat less diagnostic mammography, as lesions are discovered earlier, before they become symptomatic. Third, there has been a trend towards increasing participation within the program over the past several years, of the order of 2% per year. At the end of 2003, the Québec program was in place in 16 of 18 regions, and measures had been taken to offer services to all Québec women via mobile screening units [MSSS, 2004].

We feel that it is more useful to consider how effective population-based mammography screening is compared to no screening. Under this approach, and assuming the program eventually achieves its objective of participation rates of 70%, the resulting participation contrast of 70% would be in the upper range of those achieved in the eight studies, which varied between 43% and 82%. The PQDCS, seen in this light, provides an opportunity to insure that mammography's structure and process is optimized in Québec and that participation rates are as high as possible, even though many of the structural and process elements may have already been in place, and significant levels of participation existed prior to the program.

Finally, it should be noted that an individual women participating regularly in a screening program will derive somewhat greater benefit than this population average, since the average is intermediate between the low value among women who are screened irregularly or not at all, and a higher value among women who are screened. In this sense, program efficacy as it is affected by participation rates is not relevant in assessing whether an individual woman might derive a benefit from screening. Conversely, since the population is the sum of its individual members, it is clear that the effect at the population level will be stronger if a large number of women participate regularly.

3.3.3 Timing dilution

This is an issue in studies with short contrast or follow-up, but does not apply to the steady-state mortality reduction that a program will eventually produce. Thus we assign a value of 1.0 to timing dilution for program screening. Research trials had timing dilution ranging from 0.28 to 0.76 (Tables 3 and B-4, Appendix B), meaning that, in addition to structure and process issues and participation rates, the data analysis of the mammography trials is likely to have further diluted their measured effect, increasing the chances that a screening program like the PQDCS will demonstrate more positive results than the studies that were all handicapped by significant timing dilution.

TABLE 4

Structure and process ratings in trials and in Québec

TRIAL	TECHNICAL CONTRAST (C)	ERA (E)	QUALITY (Q)	STRUCTURE AND PROCESS SUB-INDEX (CxExQ)
HIP (New York)	1.25	0.70	1.00	0.88
MMST (Malmö)	1.00	0.80	0.85	0.68
TCS (Two-County)	1.00	0.80	0.85	0.68
Edinburgh	1.25	0.80	0.90	0.90
NBSS-1 (Canada)	1.25	0.90	0.95	1.07
NBSS-2 (Canada)	0.80	0.90	0.95	0.68
Stockholm	1.00	0.90	0.80	0.72
Gothenburg	1.00	0.90	0.88	0.79
UK Age Trial	1.00	1.00	0.98	0.98
PQDCS	1.00	1.00	0.90	0.90

3.3.4 Overall strength of contrast

Our contrast scale is intended to measure the degree to which trials have measured the full potential of population-based screening mammography, as a proportion of the effectiveness of an ‘ideal’ program with full participation, modern quality, short intervals, etc. Using this paradigm, the contrast score for screening mammography in Québec, with participation rates at 70%, would be 0.63. This score compares very favourably with the scores for the mammography studies, which ranged from 0.12 to 0.45. This indicates that screening today is quite likely to perform somewhat better than screening in the mammography trials. Some potential improvements in the quality of mammography screening and in participation rates would likely increase this advantage.

4 SYNTHESIS OF RESULTS

4.1 ASSESSMENTS OF OTHER REVIEWERS

Although most systematic reviews and meta-analyses have included all studies without discrimination, several analyses have excluded some trials based on validity concerns. As Table 5 shows, reviewers who have made any distinctions have tended to rate the HIP and Edinburgh studies lowly. In addition, Olsen and Gøtzsche's analysis found that, while the HIP and Edinburgh studies were flawed, the three most recent Swedish studies were also of poor quality [Olsen and Gøtzsche, 2001]. Additionally, several reviewers excluded the Canadian NBSS-2 data because it examined a contrast (mammography versus clinical breast examination) that was qualitatively different from the others.

In addition to these systematic reviewers, a number of evaluators have reviewed the controversial analysis that Olsen and Gøtzsche [2001] have produced for the Nordic Cochrane Collaboration Group, without themselves conducting new reviews. While these analyses have arrived at various conclusions, in general, they have tended to sympathize with many of the criticisms raised by Olsen and Gøtzsche, while tempering or contradicting their conclusions. In particular, major reviews of the Olsen and Gøtzsche critique have since been produced by the Health Council of the Netherlands [2002], by the French Agence Nationale d'Accréditation et d'Évaluation en Santé [ANAES, 2002], Health Canada [Kakuma, 2002], the US Preventive Services Task Force [Humphrey et al., 2002] and by a group working for the International Agency for Research on Cancer (IARC) [2002] of the World Health Organization (WHO).

The French group emphasized that Olsen and Gøtzsche's findings are contradicted by nine meta-analyses conducted between 1993 and 1998. However, since these meta-analyses generally included all studies without regard for their quality, this is hardly surprising. They also disagreed with Olsen and Gøtzsche's emphasis on the importance of overall mortality as the relevant indicator, and were not convinced that Olsen and Gøtzsche had adequately justified their classification as to the methodological quality of the trials reviewed. They concluded that the meta-analysis did not constitute adequate evidence to reverse their previous recommendations in favour of mammography screening of women aged 50 to 69.

The Dutch group agreed with much of the criticism regarding the trials, and in particular those concerning the HIP and Edinburgh trials. As for the three Swedish trials which Olsen and Gøtzsche found to be of poor quality, they agreed with many of the findings but doubted that the magnitude of the problems identified justified the categorization of these trials as poor quality. They also criticized Olsen and Gøtzsche's failure to make allowances for differences between the Canadian and Swedish studies, and in particular that one of the Canadian studies (NBSS-2) involved a weaker contrast. They noted that Olsen and Gøtzsche attached the greatest priority to the (internal) validity of the RCTs, and while agreeing with this approach in principle, were critical of the way Olsen and Gøtzsche chose the quality criteria and failed to apply them consistently. They concluded that the arguments presented in the Cochrane review do not convincingly refute the evidence that breast cancer screening provides a survival benefit for women over the age of 50 [HCN, 2002].

TABLE 5

Reviewers' use of trial data

META-ANALYSIS	EXCLUDED OR UNRELIABLE	INCLUDED, BUT POOR	INCLUDED, MEDIUM OR HIGH QUALITY, OR QUALITY NOT ASSESSED
CETS, 1993; 1990	HIP	-	All but HIP
Fletcher et al., 1993	NBSS-2 (CBE* in control group)	-	All but NBSS-2
Kerlikowske et al., 1995	None	-	All
Glasziou, 1997	None	-	All
Hendrick et al.,1997	None	-	All
Olsen and Gøtzsche, 2001	HIP; Edinburgh	TCS; Stockholm; Gothenburg	NBSS-1 and -2; Malmö
US Preventive Task Force [Humphrey et al., 2002]	Edinburgh	-	All but Edinburgh
Health Council of the Netherlands, 2002	Edinburgh	-	All but Edinburgh
International Agency for Research in Cancer (IARC), 2002	Edinburgh (confounding); NBSS-2 (CBE* in control group); HIP; NBSS-1 (CBE* in screening group)	-	Swedish trials
This report: AETMIS, 2005	NBSS-2 (CBE* in control group)	Gothenburg; Stockholm; Edinburgh; TCS; HIP	NBSS-1; Malmö; (UK Age Trial)

* CBE = clinical breast examination.

Kakuma [2002], in a report prepared for Health Canada, reviewed the quality of the mammography screening trials and identified many of the same problems raised by Olsen and Gøtzsche. She also reviewed other meta-analyses conducted by Fletcher et al. [1993] and by Kerlikowske et al. [1995]. In the first case, Fletcher identified numerous concerns with the validity of screening studies, and chose not to carry out a formal meta-analysis, but nonetheless accepted the results of all trials except the Canadian one. Kerlikowske, on the other hand, did not assess the quality of the trials at all before pooling their data. Kakuma concluded that Olsen and Gøtzsche's review is more comprehensive, and that trials' individual limitations are sufficiently numerous so as to jeopardize the validity of their results, so that pooling the data from all trials would tend to produce results that are misleading.

The American group (US Preventive Services Task Force) criticized Olsen and Gøtzsche's meta-analysis for not considering the age of women studied, the number, type and quality of screenings, the interval between rounds, the rates of compliance and contamination, nor qualitative differences in the contrasts studied. They retained the view that the evidence is of fair quality, but that the limited evidence suggests mortality reductions of 20 to 35%.

Finally, the IARC [2002] group, in a chapter of its handbook on the efficacy of screening by mammography, reviewed all trials and excluded the HIP and NBSS-1 trials because they included CBE in the screened cohort, the Edinburgh trial because of concerns regarding confounding factors, and the NBSS-2 trial because of the different nature of its design (inclusion of CBE and BSE in the control cohort). They judged that the four Swedish trials are valid, despite the criticisms raised by Olsen and Gøtzsche, and concluded that screening in these trials had been shown to reduce breast cancer mortality by about 25% in women aged 50–69 and by 19% in younger women.

4.2 ANALYSIS OF EVIDENCE

Our examination of published mammography trials indicates that studies differ markedly as regards the quality of their design and execution. Traditional systematic reviews have generally considered that the evidence for mammography screening is of the highest order, since many review agencies rate the evidence based on the existence of at least one well-conducted randomized controlled trial or of a meta-analysis of more than one such trial. We believe that such an analysis is simplistic and overstates the quality of the evidence. On the one hand, mammography trials have tended to fall short of modern quality standards for trials. Some were not randomized at all, but rather allocated by researchers to suit practical need. Most studies have been poorly or inconsistently documented. In particular, most have not provided baseline characteristics of women in screening and control groups, often because next to nothing was known about the control cohort. Exclusion of previously diagnosed cancers has been inconsistent. Blinding has not been attempted in any trial, either of patients or of care providers. These defects led one reviewer [Baum, 2004] to decry a "double standard in that clinical trials of screening are acceptable at a quality that would completely invalidate trials for the prevention or treatment of breast cancer."

On the other hand, we have argued that no study has been designed and conducted in such a way that the full potential of mammography screening could be determined, and that it is important to also examine this aspect of validity. Our approach, rather than qualifying the body of evidence as weak or strong, is to consider each study separately, considering its relevance to the evaluation of a modern screening program, and rating its strength of contrast and the other aspects of validity which protect against systematic bias. This approach can be summarized as:

Step 1: Select relevant trials, based on the nature of the study contrast.

Step 2: Order these trials, by the degree of protection against bias of unknown direction.

Step 3: Pool trials in this order, successively adding trials of lesser validity, keeping track of the strength of contrast of these poolings and the results regarding breast cancer mortality.

First, based on the relevance of its contrast, we believe the NBSS-2 study should be considered separately, since it compared two screening regimens (mammography, clinical breast exam and breast self-examination versus clinical breast exam and breast self-examination alone). Trials, in chronological order, are thus classified in Table 6.

Next, we have re-ordered the remaining trials by the validity score (Table 7), since we feel that trials with low scores may have biased results, but we do not know in which direction this bias is likely to have occurred. As for these validity concerns, our rating scale indicates that two trials, Edinburgh and the Two-County trial, are particularly problematic. In addition, the HIP trial, while it achieved a moderate score, might also be considered unusable, since the loss of as well over half of patients to follow-up is probably inadequately penalized by our scale. To the extent that other reviewers have judged the quality of individual trials, most have excluded one or several of these three trials as flawed. Although we have included the UK Age Trial in tables 6 and 7 for the sake of comparison, it is not included in subsequent analysis, since it has not yet published any mortality results.

Finally, we have assessed the strength of the screening contrast in each of these studies, using our quantitative scale. We present in Table 8 the results of our assessment of the strength of the screening contrast, in the same order of declining validity. Using the inverse of the variance as a weighting factor, we also give the cumulative value of the strength of contrast of these studies, as studies of lesser validity are progressively included.

TABLE 6

Step 1: Relevant trials

INITIATION YEAR	TRIALS
1991	UK Age Trial
1982	Gothenburg
1981	Stockholm
1980	NBSS-1
1979	Edinburgh
1977	Two-County (TCS)
1976	Malmö (MMST)
1963	HIP

TABLE 7

Step 2: Trials by order of validity against bias of unknown direction

RANKING	TRIALS	VALIDITY
1	NBSS-1	4.0
2	Malmö (MMST)	3.0
3	Stockholm	2.0
4	Gothenburg	1.5
5	HIP	1.5
6	Two-County (TCS)	0.5
7	Edinburgh	0.5
	UK Age Trial*	4.0

* For comparison purposes.

TABLE 8

Step 3: Strength of contrast

RANKING	TRIALS	STRENGTH OF CONTRAST	CUMULATIVE STRENGTH OF CONTRAST *
1	NBSS-1	0.45	0.45
2	Malmö (MMST)	0.27	0.38
3	Stockholm	0.28	0.36
4	Gothenburg	0.37	0.36
5	HIP	0.12	0.29
6	Two-County (TCS)	0.42	0.34
7	Edinburgh	0.26	0.32

* Weighted by inverse of the variance of the log odds ratio, i.e., 1/VAR(ln(OR)).

No study came close to the standard of a hypothetical study with many years of regular screening with modern equipment, quality assurance, two-view mammography at intervals of two years or less, and with full participation. Compared to this standard, the seven published mammography screening trials only achieved strength of contrast of between 12 and 45% of what would be possible. Using this same standard, modern programs are likely to achieve considerably greater strength of contrast, with the Québec program estimated to be 63% of full potential, if it attains a participation rate of 70%, when compared to the option of no screening. Better performance in screening may or may not be associated with greater mortality reductions, depending on the relative effectiveness of treatment regimens for early versus late cancers. However, this strength of contrast score indicates that a modern program is likely to identify lesions earlier, maximizing the potential of a screening program given today's treatment options.

Several further observations can be made related to previous systematic reviews conducted by other investigators. First, as we have noted, most systematic reviews have included all studies without consideration of their validity or, in particular, their strength of contrast. While this has the advantage of avoiding arbitrary choices and gives narrower statistical confidence intervals, it implies that validity and strength of contrast are not important considerations. In the case of mammography screening trials, we feel this sanguine view is not justified, since we have found that all trials leave much to be desired and differ markedly in their individual characteristics. Second, when criteria for inclusion have been stated, they have usually concerned only certain aspects of validity that may have introduced bias of unknown magnitude and direction. While these aspects are important, they are not sufficient in the case where important failures in establishing study contrast may have systematically biased the results in the direction of negative findings. This situation certainly prevails with regard to mammography screening trials. Moreover, to the extent that either sort of validity criteria are applied, we believe that they should be explicit, allowing readers to judge whether they have been appropriately applied.

In the following sections, we will examine the reductions of breast cancer mortality obtained in the various trials, with due consideration given both to their strength of contrast and other aspects of validity. To do so, we present the mortality results of each study, once again in order of declining validity. We also present the cumulative mortality reduction of all studies of equal or greater validity on this measure, allowing the reader to judge how far down the list he wishes to go. The validity score thus allows for an ordering of studies by virtue of aspects of validity which protect against bias of unknown direction, and the strength of contrast score allows for an assessment of the degree to which the results of each study or combination of studies is likely to have been diluted.

4.3 BREAST CANCER MORTALITY RESULTS FOR ALL AGES

Although the primary aim of this analysis was not to arrive at a single estimate of mammography's ability to reduce breast cancer mortality, it is interesting to compare the strength of the various studies and their reported results. In this section and the following two sections, we will examine published breast cancer mortality results for all ages, for younger women (primarily under 50 years old), and for older women (primarily over 50 years old). Although we provide the results for all trials, we will focus on the four that we believe to be the most valid. The two trials that score highest with regard to validity both produced results showing little reduction of breast cancer mortality, if any. In particular, the large Canadian study showed similar mortality in the screened group and in the control group, with a risk ratio (RR) of 0.97 (95% CI: 0.74–1.27). Likewise, the smaller Malmö study showed a non-significant reduction, with a risk ratio of 0.82 (95% CI: 0.59–1.15). The UK Age Trial, also of high validity, has not yet published results.

The Stockholm, Gothenburg and HIP results fall in a group with a small range of poor scores for both strength of contrast and validity, and yet they showed greater mortality reductions, with RR of 0.74 (95% CI: 0.50–1.10), 0.79 (95% CI: 0.58–1.08), and 0.77 (95% CI: 0.61–0.98), respectively. Finally, the TCS and Edinburgh trials, of very poor quality, showed better respective risk ratios of 0.67 (95% CI: 0.56–0.79) and 0.78 (95% CI: 0.62–0.97). Some researchers in the field of meta-analysis have suggested that studies with low validity tend to overstate the efficacy of the interventions they evaluate. The seven studies examined here tend to support this conclusion. If one were to consider only traditional aspects of study validity and take no account of contrast issues, we feel it would be fair to cast strong doubt on the efficacy of mammography screening.

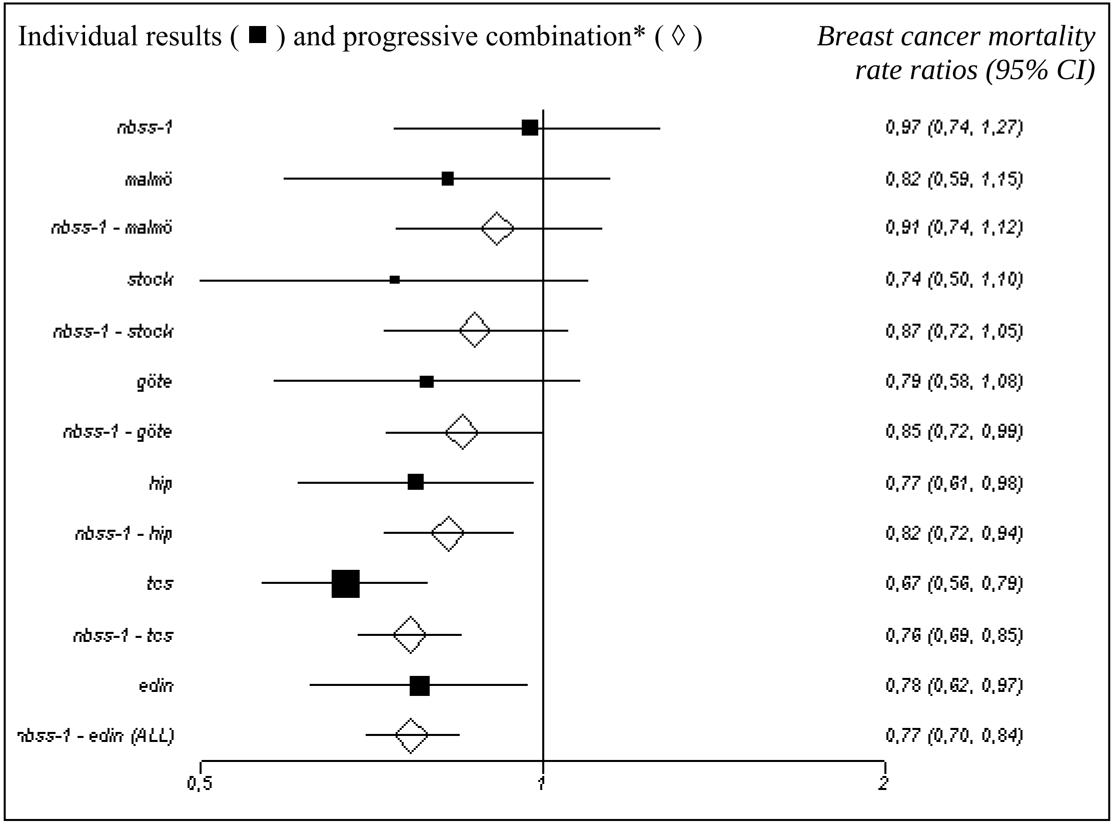
Figure 1 summarizes the breast cancer mortality results of each trial, in descending order of validity scores along with the weighted average cumulative result as successive studies with lower validity are added.

Although we might expect studies with weak contrasts to show negative findings, the opposite is true for the HIP, Stockholm and Edinburgh trials that scored poorly on strength of contrast, yet reported significant mortality reductions. This might suggest that these discrepant findings can be explained by problems with the validity of these latter trials. An alternative explanation is that, although early trials such as HIP did not provide a strong screening contrast, therapies available at that time may have meant that a small advance in identifying early tumours made a bigger difference then than in later trials, when therapies had become more effective for both early and more advanced tumours. In either case, these results should be treated with caution, and may not apply to today's context.

Overall, in the two studies found to be most valid by our analysis, the cumulative RR is 0.91 (95% CI: 0.74–1.12). Including data from two further studies with weaker validity, the overall RR would be 0.85 (0.72–0.99). Successive addition of study results consistently show some reduction in breast cancer mortality, ranging from 9% (medium-quality studies), to 15% (medium- and poor-quality studies), to 23% (all studies); in other words, the more valid studies tend to show lesser reductions, and confidence intervals sometimes include the null value.

FIGURE 1

Synthesis of breast cancer mortality results: All ages



Each line with the symbol -◊- represents the cumulative effect of all preceding individual studies.

4.4 BREAST CANCER MORTALITY RESULTS IN YOUNGER AGE GROUPS

Studies have included women between 39 and 74 at randomization, but most individual studies are too small to allow conclusions about relative efficacy in the different age groups to be reliably drawn. Despite the studies' numerous qualitative differences, it is thus natural that attempts have been made to aggregate the experience of various age groups across studies. In particular, interest has focused on the efficacy of mammography in age groups that correspond roughly to pre- and post-menopausal periods, usually using the age of 50 as the dividing point. With regard to younger women, two major points should be made before looking at the evidence of efficacy. First, a study enrolling 40- to 49-year-old women is not at all the same as a study enrolling women at age 40 and screening them up to age 49. If the question is to know how efficacious screening might be if it were to be begun at age 40, then one would adopt the latter strategy; this is the design of the UK Age Trial. However, all other studies that included younger women enrolled women of all ages within this age group, i.e., enrolled some women at 40, but also some at 41, 42, etc., up to 49. For instance, in the Gothenburg study, each of the 11 one-year age groups from 39 to 49 years old were approximately equally represented, constituting between 8 and 11% of the total cohort. This age structure is thus about five years older, on average, than that which would be necessary to show the effects of beginning screening at age 40.

Secondly, over the course of the study contrast, many women who begin in a younger age group age graduate into an older one. For instance, the ‘younger women’ age category in the Malmö study was composed of women 45 to 54 years old. The youngest of these women, who began screening at age 45, were 54 years old by the time the screening period ended; those who were 54 years old at study onset were 63 years old at the end of the contrast. As the HIP investigators observed, it is difficult to be sure that any effectiveness of screening mammography among these women was not due to the screening that took place when they had aged into an older category. In addition, of the seven trials that included younger women, two (Malmö and Edinburgh) included no women below the age of 45 [Shapiro et al., 1985].

Much larger confidence intervals around the results of all seven trials can be attributed both to the fact that smaller numbers of women were enrolled in these age groups, and to the fact that a smaller number of mortality events are to be expected among these women, where the incidence of breast cancer is lower.

Again, these studies are too heterogeneous to be aggregated without detailed examination of their differences, but in general, mortality reduction is lesser than in older women. In the two studies with higher validity scores, both showed equal mortality in the screened group, with NBSS-1 showing a RR of 0.97 (95% CI: 0.74–1.27), and Malmö showing a RR of 1.01 (95% CI: 0.58–1.77). The studies with poor validity scores showed more variable results, with a RR of 1.08 in the Stockholm trial (95% CI: 0.54–2.17) and 0.65 in the Gothenburg study (95% CI: 0.40–1.05).

Figure 2 summarizes the results available for these younger age groups, both for individual study results and for cumulative results of studies added in order of their validity scores.

Overall, in the medium-quality studies, the cumulative RR is 0.98 (95% CI: 0.77–1.25). Including data from studies with weaker validity, the overall RR would be 0.92 (95% CI: 0.74–1.13). The better data thus show minimal mortality reduction if any, and do not reach traditional levels of statistical significance. Successive addition of study results consistently shows no significant reduction in breast cancer mortality, ranging between 2% (medium-quality studies), 8% (medium- and poor-quality studies), and 13% (all studies). Once again, the more valid studies tend to show lesser reductions. In addition, the mortality reduction is much smaller in younger women, and confidence intervals include the null value for all combinations of studies.

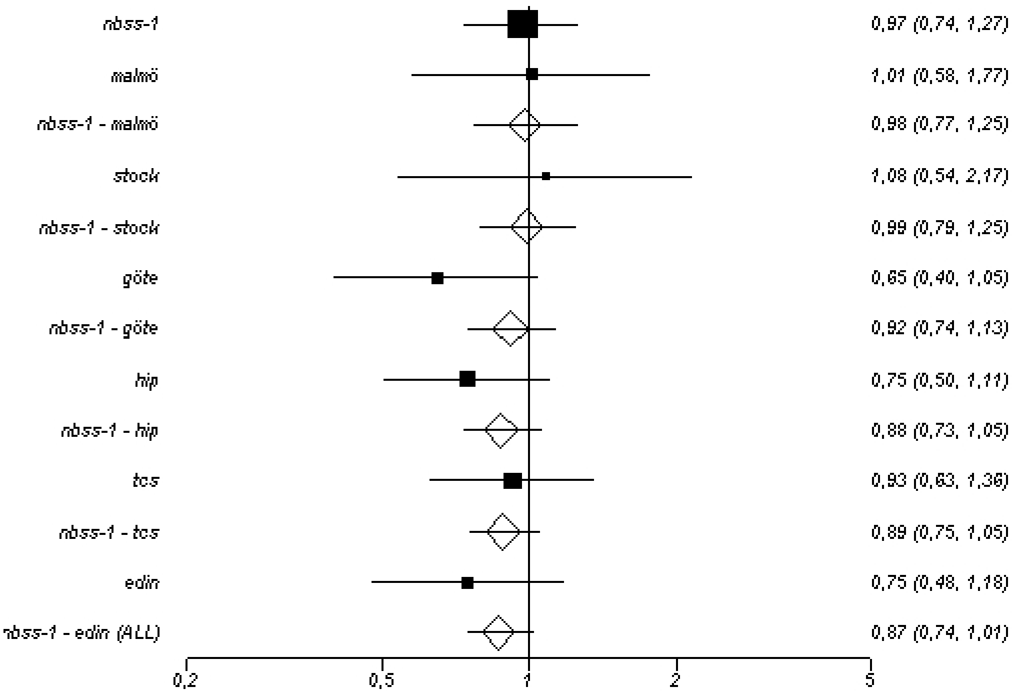
4.5 BREAST CANCER MORTALITY RESULTS IN OLDER AGE GROUPS

A second sub-group analysis can be applied to data pertaining to women at least 50 years old at enrolment (55 years in the Malmö trial). Again the studies are probably too heterogeneous to be properly aggregated. Weaker trials from Stockholm and Gothenburg showed risk ratios of 0.62 (95% CI: 0.38–1.00) and 0.91 (95% CI: 0.61–1.36), respectively. Figure 3 summarizes the results available for these older age groups, both for individual study results and for cumulative results of studies added in order of their validity scores.

FIGURE 2

Synthesis of breast cancer mortality results: Younger women

Individual results (■) and progressive combination* (◊) *Breast cancer mortality rate ratios (95% CI)*



* Each line with the symbol -◊- represents the cumulative effect of all preceding individual studies.

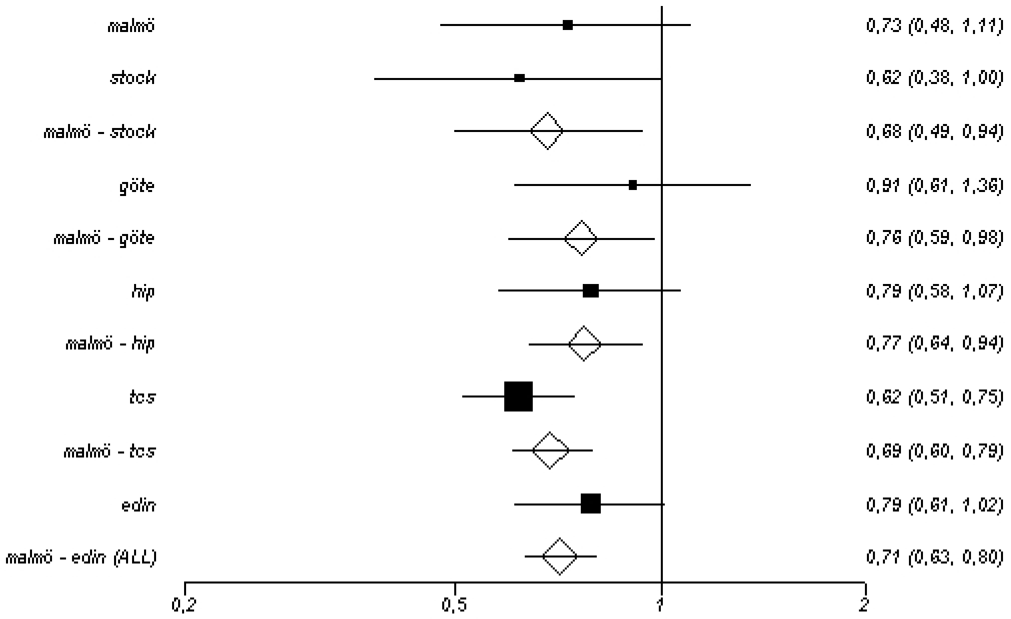
Overall, in the Malmö study, the only study of medium validity, the RR is 0.73 (95% CI: 0.48–1.11). Including data from studies of poor quality, the overall RR is 0.76 (95% CI: 0.59–0.98), and including all data irrespective of validity gives an overall RR of 0.71 (95% CI: 0.63–0.80). The data thus show mortality reductions between 24 and 29%, a range substantially higher than what is observed when all age groups are combined. Confidence intervals in this sub-group are naturally wider, but no combination of studies includes the null value.

While we have not combined the results of NBSS-2 with those of the other studies because of the qualitatively different technical contrast, that trial showed no reduction in breast cancer mortality in the group screened with mammography, with a RR of 1.02 (95% CI: 0.78–1.33). In other words, it did not demonstrate any additional value of screening mammography beyond that which could be achieved with intensive screening using regular, high-quality clinical breast examination and breast self-examination. Whether this is better explained by some efficacy of clinical breast examination or of breast self-examination or both, or, as the principal investigator suggests, by improved treatment options for more advanced cancers [Miller et al., 2000a], or by some other factor, is beyond the purview of this paper.

FIGURE 3

Synthesis of breast cancer mortality results: Older women

Individual results (■) and progressive combination* (◊) *Breast cancer mortality rate ratios (95% CI)*



* Each line with the symbol -◊- represents the cumulative effect of all preceding individual studies.

5 DISCUSSION AND CONCLUSIONS

We have reviewed all published screening mammography trials and have examined in detail their methodological strengths and weaknesses. Our findings are in a sense intermediate between those of most reviewers, who have concluded that the effectiveness of screening mammography is well demonstrated, and those of Olsen and Gøtzsche, who conclude there is no evidence supporting screening.

Based on analysis of the trials' validity, we found that the Canadian and Malmö studies are of the highest validity, and these studies do not support the hypothesis that screening mammography is effective. Studies with less valid design and conduct have tended to produce more favourable conclusions, and so their inclusion clearly leads to estimates of greater mortality reduction, but tends to weaken the confidence with which these conclusions can be stated. However, many of the studies also showed weak contrast, including the Canadian and Malmö trials, and these weaknesses would tend to make it harder for studies to demonstrate the full potential of expertly conducted mammography using modern equipment and carried out over adequate periods of time. One of these, timing dilution, could be partially remedied in some studies by censoring mortality results from the first years of follow-up, but this adjustment would not fundamentally alter the above conclusions. Other weaknesses in establishing or maintaining contrast in these studies can not be corrected by analysis, but should nevertheless be noted since they may have obscured the potential for mammography to reduce mortality. In comparison, a modern screening program such as the *Programme québécois de dépistage du cancer du sein* (PQDCS), if it achieves a participation rate of 70%, would constitute a stronger contrast, compared to no screening, than that which applied in any of the studies.

It is thus likely that the PQDCS will eventually permit an earlier diagnosis of breast cancer than what screening trials have been able to achieve. This is not a sufficient cause for greater reductions in mortality, since advances in therapeutic efficacy may have cancelled out some of this advantage. However, strong diagnostic performance is a necessary condition for achieving maximal gains that screening may allow.

This review was undertaken to answer three questions:

Question 1: “What is the strength of the scientific evidence on which screening mammography programs are based?”

We believe that there are serious concerns regarding the validity of most of the trials supporting mammography screening. Some of these trials hardly deserve the moniker ‘randomized’, and the degree of control of the execution of the contrast makes the second adjective in ‘randomized controlled trials’ rather dubious as well. In most of the studies, including all four from Sweden, little was known about the control population except the age of its members, many of the studies used non-random assignment to screening and control groups, exclusion of previously diagnosed breast cancer was almost always done years after the study contrast using hospital records, blinding was never attempted, and so on. Reporting of important aspects of randomized controlled trials has generally not been up to modern standards, with aspects such as randomization techniques, baseline equality of screening and control cohorts, exclusion criteria, etc. often not mentioned, and methods of analysis are not well described for most studies. Criticism of these aspects of the mammography studies [Olsen and Gøtzsche, 2001] has not yet succeeded in eliciting further details or convincing explanations from the studies’ principal investigators [Nyström et al., 2002].

In addition to these validity concerns, we have emphasized that the eight trials are highly heterogeneous with regard to the strength of the contrast that they studied. Mammography is a complex technique whose success depends on a long chain of actions that must each be carried out with precision. We have identified numerous weaknesses in all the major studies, meaning that the full potential of screening mammography has perhaps not been thoroughly explored. It is plausible that a well-designed, valid trial of modern mammography, in conditions which guarantee high-quality execution of the technique, over a long period of time, with high rates of participation and low rates of contamination and with adequate follow-up and analysis of the results, could demonstrate convincing evidence of the efficacy of screening mammography. The UK Age Trial may fulfill most of these conditions, but complete results are not likely to be available before 2008 or 2010. It is likewise plausible that a modern program of screening mammography, operated under the same high-quality conditions, would provide results significantly better than what studies have demonstrated in the past.

Analysis of the scientific literature indicates that screening mammography, as practised in the best studies, resulted in a moderate reduction of breast cancer mortality, of the order of 9 to 15%, for women of all ages. Data restricted to women over the age of 50 show somewhat greater reductions, of the order of 24 to 29%. Furthermore, our analysis has demonstrated that modern mammography, carried out in circumstances that maximize its performance, has the potential to identify cancerous lesions earlier in their progression, and this may allow for some further reduction in mortality. In addition, the benefit to the individual woman participating regularly in a screening program should be greater than the average benefit demonstrated by population trials.

As we have discussed, there are also qualitative reasons to worry that mammography screening causes adverse effects, but very little quantitative data that might allow a policy-maker to assess whether the harms are sufficiently important to nullify the measured benefits of screening. It should be noted that, while false positive rates were around 2–5% in the Swedish trials and ongoing screening programs [Nyström et al., 2002; Elmore et al., 1998], they are typically two to three times higher in North American practice [INSPQ, 2003; Health Canada, 2003; Elmore et al., 1998], making the potential trade-off between potential positive and negative effects of screening less favourable. The inconvenience of mammography, the anxiety associated with false positives, the extra diagnostic workups and the cost of screening are more certain than the hoped-for efficacy. Although there is little evidence of long-term anxiety produced by screening [Lampic et al., 2001], these side effects are difficult to measure and their gravity is difficult to quantify on the same scale as the potential benefits of mammography. The possibility that mammography may be associated with more serious adverse effects cannot be adequately addressed using data from any of the available trials thus far, since they were generally designed only to test the hypothesis of a reduction in breast cancer mortality. Although general acceptance of mammography screening may have made it impossible to conduct new randomized studies, it should be possible to study the unintended effects of mammography using a non-experimental design [Miettinen, 1983], given that the risk of experiencing these side effects is probably not associated with participation in screening programs. Adverse effects of mammography are a problem of unknown magnitude, and might counterbalance any breast cancer mortality advantage which mammography confers.

Further updates of the eight trials are unlikely to alter the results already published, since most of the relevant periods where mortality would be affected by screening have already passed. The major exception is the UK Age Trial, for which no results have yet been published. Due to its size and methodological quality, the results of this trial could well have a material effect on the overall corpus of evidence. No further trials are likely to be conducted on screening mammography for women 50 to 69 years old, since any such trial would not be feasible in the current context of widespread use of screening, nor would it meet usual ethical requirements, given the clinical consensus in favour of mammography. Future improvements to mammographic screening will likely need to be evaluated based on measures of diagnostic performance combined with rates of grade- and stage-specific survival.

Conclusion 1: *Existing scientific trials, despite their flaws, support mammography screening programs. In addition, there are good reasons to believe that modern, well-conducted screening programs may achieve earlier detection and diagnosis of breast cancer and, perhaps, greater reductions in breast cancer mortality than what has been found in screening trials.*

Question 2: "What is the evidence in support of screening mammography for women aged 40 to 49 years?"

There is much less data available to answer the question, since most study experience is in women over 50, even though some women in some of the studies started screening several years earlier than their fiftieth birthday. If the younger subgroups of these studies did show efficacy, one could only conclude that screening is efficacious in women 45 to 54 years old. In any case, the best data available show that screening younger women provides no significant reduction in breast cancer mortality. We conclude that there is virtually no evidence that screening is effective in women 40 to 49 years old, and indeed there is some limited evidence that it is ineffective. Fortunately, the UK Age Trial will vastly increase the amount of data about women in this age group. Indeed, if the latter trial does show efficacy, then this would substantially bolster the argument that mammography may be efficacious in older women as well.

Side effects are a particular concern in younger women, since most programs have observed that rates of false positive mammograms tend to be higher. This may be because younger women's breasts are more difficult to evaluate, both because they tend to be denser and because of their physiologic response to hormones throughout the menstrual cycle. In the absence of any convincing data that mammography is efficacious in this age group, harmful effects may well outweigh any possible positive effects.

Conclusion 2: *Trial data published to date do not provide scientific justification to recommend screening for women younger than 50. However, this conclusion does not exclude the possibility that screening of individual women, based on a personalized risk assessment, could be of benefit. These conclusions should be reviewed when results from the UK Trial become available.*

Question 3: "What are the implications of research studies for maximizing the effectiveness of modern programs such as the *Programme québécois de dépistage du cancer du sein* (PQDCS)?"

We have concluded that there is some uncertainty as to the efficacy of screening mammography, but that significant reductions have been observed, primarily in women over 50. It is possible that the screening trials' poor design, poor execution of the contrast and inadequate analysis of the results may have handicapped these results. Indeed, participation over many years in a modern era program such as Québec's PQDCS, when compared to the option of no screening, constitutes a contrast which is, in many regards, stronger than that which has been studied in the eight screening trials. There is thus potential for this program to produce results which are better than those reported in the scientific literature.

Although the PQDCS already includes rigorous control of the quality of films produced, certain aspects of the structure and process of trials examined under the rubric of strength of contrast can be transposed as additional quality norms. Notable among these are double reading of films and an annual reading volume sufficient to allow each radiologist to acquire and maintain the necessary expertise to detect breast cancer in its early stages. These aspects should also allow for a reduction in false positive rates and subsequent unnecessary diagnostic procedures.

Moreover, high participation rates at each screening round will contribute to achieving and perhaps exceeding the mortality reductions obtained by screening trials.

Finally, trials have provided little information on the possible negative effects of screening, which may counterbalance screening's positive effects. In particular, false positive rates in Québec's program are much higher than those observed in most trials, and many of the negative effects of screening are likely to be exacerbated by high rates of false positive screens and unnecessary interventions.

Conclusion 3: *Modern screening programs such as the PQDCS may produce outcomes comparable or even superior to those observed in screening trials if they achieve a standard of quality equal to or better than the standard achieved by trials. Measures that should reduce false positive rates and assure high-quality screening include making sure that high-quality mammographic films are being produced, that readers have the necessary expertise to detect early cancer and avoid false positives, and double reading of a proportion of films. While participation rates should be as high as possible, efforts to increase participation should not overstate the benefits of mammography nor understate the risks and uncertainties which remain.*

APPENDIX A

SCREENING CONTRAST AND ITS DURATION

INTRODUCTION

Mammography screening trials that have been conducted have typically involved a screening contrast consisting of 2 to 7 rounds of mammography, separated by intervals of 12 to 28 months. These trials are useful to decision makers to the extent that they allow for extrapolation of the effect of such short programs (typically, 4 to 10 years of screening) to potential programs of much longer duration. Present screening programs are typically offered to women 50 to 69 years old, and sometimes also to younger women (40–49) and older (70 and over). Organized programs thus have a duration which may be 20, 30 or even 40 years for any given participant. The analysis of the results of screening trials requires consideration of the duration of the screening contrast, since longer trials will typically provide a stronger contrast and thus are more relevant to decision making about screening programs. To take an extreme example, a trial involving two rounds of screening at a one-year interval would be a weak contrast, and would not be expected to produce as much of an impact on breast cancer mortality as a 20- or 30-year program. A 10-year trial, appropriately conducted and analyzed, has a much better chance of demonstrating a screening program's potential, and this qualitative difference should be recognized when reviewing the effects of trials of markedly different length.

Published trials have produced contrasts varying between 3–4 years (HIP study) and 8–9 years (Malmö), with the UK Age Trial currently underway, designed to produce a 10- to 11-year contrast. Unfortunately, the exact duration of the screening contrast is not unambiguously described in the published reports of screening trials. This report aims to describe the concepts necessary to an analysis of the duration of contrast to allow for a more complete evaluation of the timing of the expected effects of a screening trial.

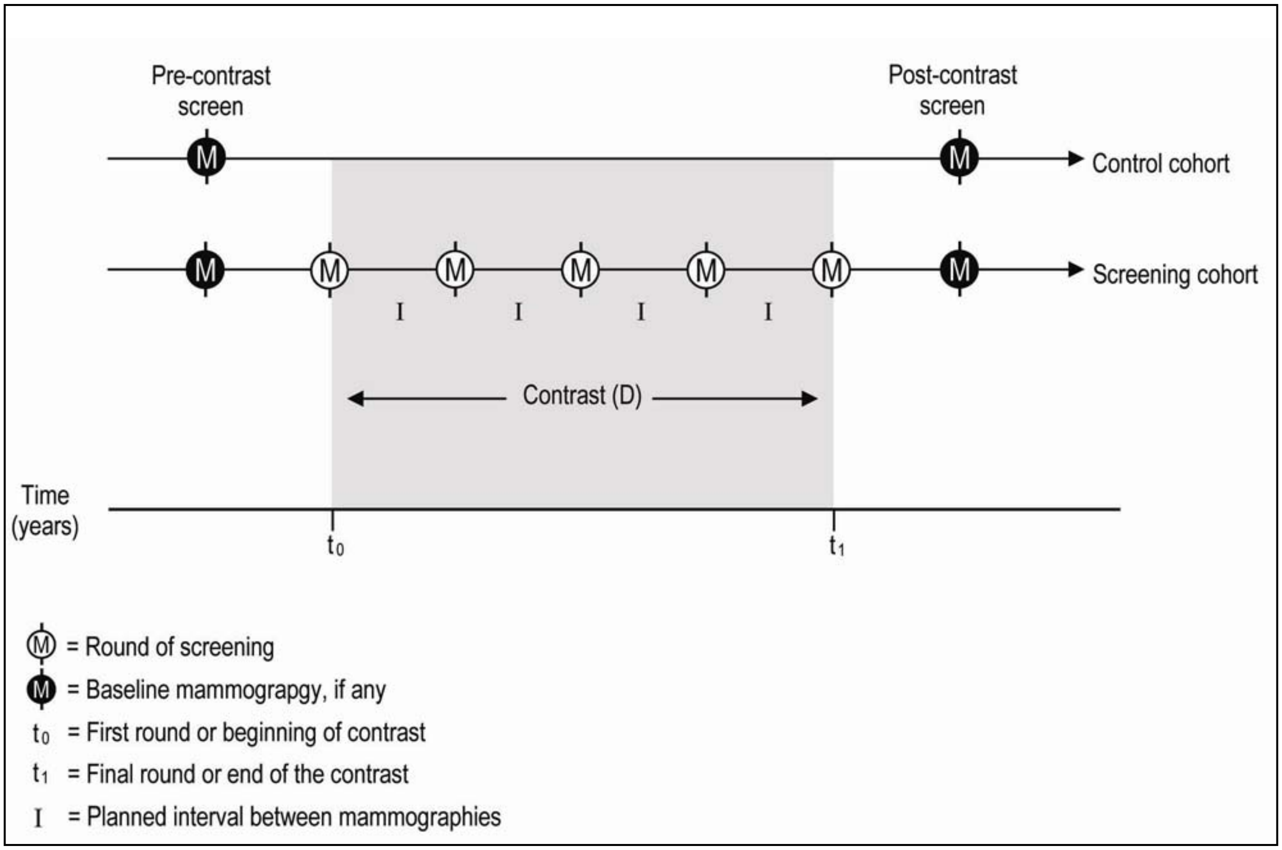
DEFINITION OF CONCEPTS

Clinical trials involve the assignment of comparable groups of individuals to at least two different regimens for a certain duration of time. Most trials involve the comparison of two cohorts, which we will refer to as the screening cohort and the control cohort, and this is the case for all nine mammography trials described in the published literature. Random allocation of individuals to these groups is recommended with the aim of producing comparable cohorts, i.e., cohorts whose individuals have similar profiles of known and unknown risk factors for the outcome of interest. In particular, these two cohorts should have similar profiles with regard to past screening activities. The intention is that the screening and control cohorts will differ only as regards their assignment to screening or control intervention, so that any differences in outcomes can be attributed to this contrast. It is important to note that subsequent to the screening contrast, the two groups should also have similar screening experience, if we are to attribute any difference in outcome to the defined screening contrast. We can thus define the screening contrast to be the temporary divergence between the experience of the screening cohort and the control cohort.

The beginning of the contrast is conceptually fairly clear, since it is the first screen offered to a given woman in the screening cohort. The date of randomization is often used as the beginning of the contrast, since according to the intention-to-treat principle, the contrast continues throughout the study period, regardless of whether an individual woman participates in all or any screening rounds. It should be noted that a recent previous mammography should be an exclusion criterion for immediate entry into the study, since the mammography would be clinically unnecessary and such a woman would not experience a complete contrast with her control counterpart (also recently screened, in principle), making the timing of the contrast ambiguous. However, such an exclusion criterion does not seem to have been applied in any of the studies examined.

Starting at the onset of the contrast, a woman in the screening cohort will be offered a series of **n** screening mammograms, separated by n-1 intervals of length **I**. The contrast will end on the day when her counterpart in the control cohort begins to have the same screening regimen as her, and thus screening experience reconverges. We will refer to the total period during which screening experience is contrasted as the duration (D) of the screening contrast.

FIGURE A-1
Basic screening contrast


Pre-contrast screen
Post-contrast screen
M
Control cohort
Screening cohort
I
I
I
I
Contrast (D)
Time (years)
t_0
t_1
M = Round of screening
M = Baseline mammograpgy, if any
t_0 = First round or beginning of contrast
t_1 = Final round or end of the contrast
I = Planned interval between mammographies

DURATION

The duration **D** is the total period of time during which this screening contrast takes place. It will in principle be equal to the number of intervals between rounds of screening in the contrast, multiplied by the average length of these intervals. However, the definition of the end of the contrast is somewhat more subtle, since it depends on the protocol for screening of the control cohort. Three cases may be distinguished, according to whether the screening contrast is terminated passively, actively and simultaneously with the final round or actively and subsequent to the final round.

In the first case, the last of **n** rounds of screening is offered to women in the screening group, but not to women in the control group, and no further screening is actively offered to either group. This was the case, for example, in the HIP study. It is presumed that, after the contrast, women in the screening and control groups will have mammograms at the same rate, but in the case of women in the screening group, since they have just been screened, they will generally not be screened again until the end of another interval **I**. Control group women, on the other hand, have no such constraint, and will have their mammogram at any time during the interval, and on average halfway through. Since the first round of screening is at time 0 and the n^{th} round at (n-1)•I, the total duration D of the screening contrast would thus be (n – 1)•I + (1/2)•I = (n – 1/2)•I.

In the second case, the screening contrast may be actively terminated by systematically offering screening to all women in the control group as was the case in all of the Swedish trials except for the Malmö trial. In this case, the contrast will end somewhat earlier; its duration will be (n-1)•I if the control group screen is offered at the time of the final round of screening for the screening group.

In the final case, the screening contrast may be actively terminated by systematically offering screening to all women in the control group, but only at the end of the n^{th} screening round offered to the screening group. In this case, the control group would effectively receive its screen simultaneously with the screening group's $(n + 1)^{st}$ screen; its screening contrast duration would then be n•I. This strategy was not used by any of the eight published mammography trials.

Three further points are relevant to the evaluation of the trials' duration of study contrast. First is that at issue is study time and not calendar time. Study time is counted from time zero, which is in principle the moment at which a woman in the screening group is invited to a series of screening mammograms, whereas her counterpart in the control group is not invited. Although women may enter the trial at different points of calendar time, they all begin their contrast at study time zero. The length of the screening contrast is evidently the study time between time zero and the end of the contrast, and not the calendar time between the first randomization of a woman and the last screen conducted within the study. In other words, a study may be conducted over a period of time which is substantially longer than the study time during which individual subjects experience the screening contrast, if all participating women do not begin their contrast at the outset of the study.

Secondly, it should be noted that contrast duration will often not be identical for all study participants. In particular, many trials have enrolled new subjects until late in the study, sometimes so late that the study was closed before later entrants could receive all the rounds of screening that early entrants received. This is usually documented, and in these cases we have calculated the average length of contrast, weighted by the number of women who received contrasts of each length.

A final point is that we have used the planned length of intervals between each round of screening, and not the average interval between screens that women actually received. Although an interval or a range of intervals are always proposed by the study protocol, observed intervals between actual screening exams can be both shorter (if allowed) and, more commonly, longer, since study subjects have considerable freedom to participate when it is convenient for them. Any delays between screens or missed screens should be reflected in participation rates, and using average real intervals between screening rounds would constitute double counting of this phenomenon.

CONCLUSION

In order to adequately consider the length of a screening contrast which a trial has designed and executed, the notions of the beginning and the end of the screening contrast must be defined, and information about the number of intervals, the length between intervals, and how many women received each duration of contrast should be available. Fortunately, this information is available for the eight published trials. The method chosen by investigators for concluding the screening contrast can alter the total length of contrast by as much as one inter-screen interval I. We have assumed that rates of screening in screened and control groups were identical before and after the contrast; to ensure validity and proper analysis, this should be reported in investigators' publications, but it has been universally neglected in the reporting of trial results.

Concerning rates of screening prior to the research contrast, it is likely that rates of mammography in younger women have been low, and in any case equal in screening and control groups if randomization has been conducted properly. We have also assumed that women recently screened were excluded from entry into trials, although this was also not reported. If it were not the case, it would have the effect of reducing the real duration of contrast by a small amount.

As for screening after the contrast, rates of mammography might not be equal, for example if screened women had acquired the habit of screening mammography. This would tend to increase the effective duration of contrast. Since this information is not available, and would likely not have a large impact in any case, we have assumed in this analysis that rates of screening were similar before and after the screening contrast.

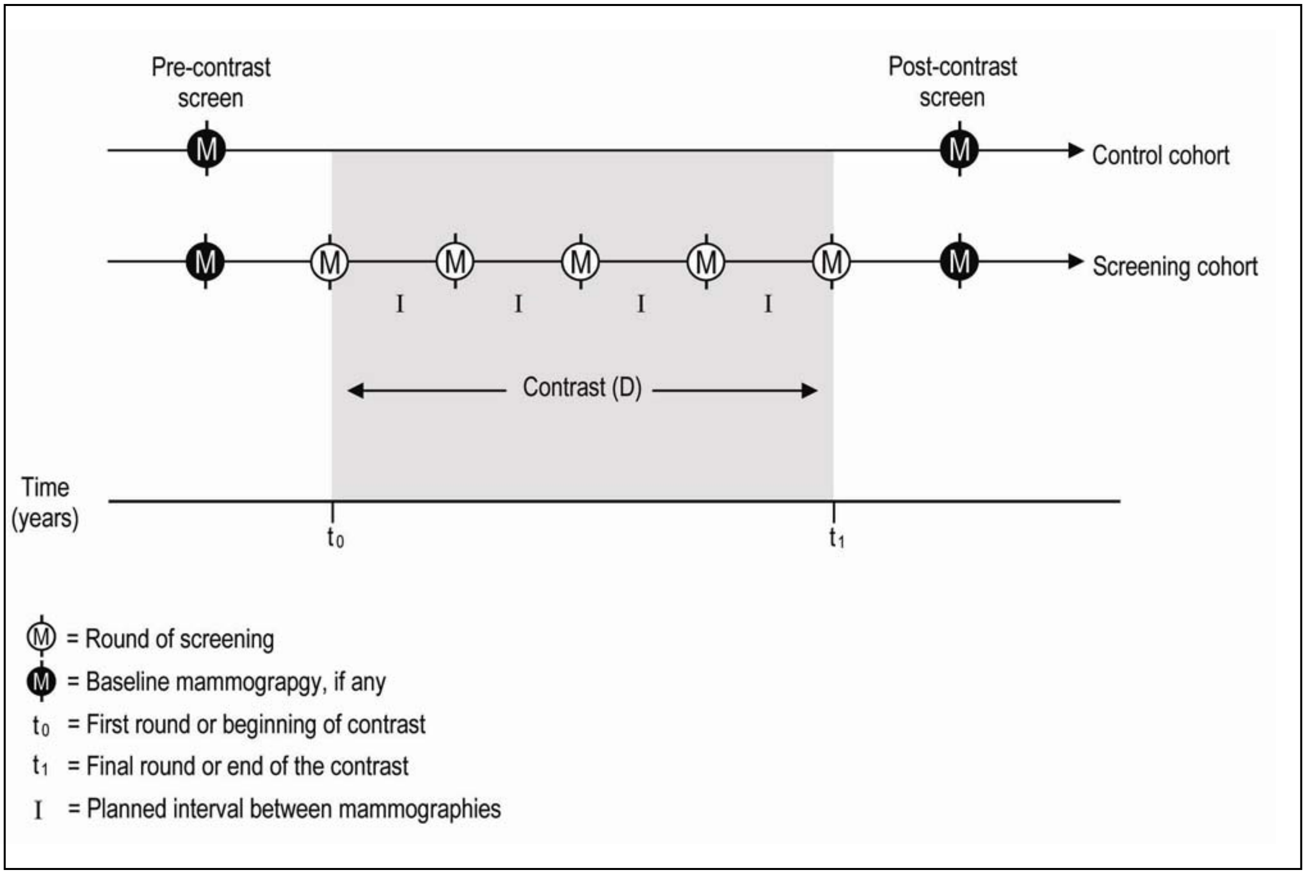
APPENDIX B

TIMING OF THE EFFECTS OF SCREENING ON BREAST CANCER MORTALITY

INTRODUCTION

We have previously defined the screening contrast to be the temporary divergence in screening experience between the screening cohort and the control cohort (see Appendix A: Screening contrast and its duration). Successful randomization implies that, prior to the screening contrast both cohorts should have had similar experiences with respect to screening. Starting at the onset of the contrast, a woman in the screening cohort will be offered a series of **n** screening mammograms, separated by interval **I**. The contrast will end on the day when her counterpart in the control cohort begins to have the same screening regimen as her, and thus screening experiences reconverge. We will refer to the total period during which screening experience is contrasted with the control experience as the duration of the screening contrast (D) (see Figure B-1).

FIGURE B-1
Basic screening contrast


Pre-contrast screen
Post-contrast screen
M
Control cohort
Screening cohort
I
I
I
I
Contrast (D)
Time (years)
t_0
t_1
= Round of screening
= Baseline mammograpgy, if any
t_0 = First round or beginning of contrast
t_1 = Final round or end of the contrast
I = Planned interval between mammographies

Basic paradigm

Breast cancer screening will only be effective if there are a significant number of tumours which are both fatal when detected at the symptomatic stage but curable if diagnosis and treatment are advanced. In such cases, we imagine the prototypical contrast between two hypothetical series of events:

1) Control group: a potentially fatal tumour begins to grow, it becomes symptomatic, the diagnosis is made, the tumour may be treated, but death nevertheless occurs a number of years later.
2) Screening group: a potentially fatal tumour begins to grow, it becomes screen-detectable before it becomes symptomatic, it is detected by the next round of mammography, it is diagnosed and treated at this early stage, and death is avoided.

We define the **lag time L** to be the time between the moment when a potentially fatal case could be detected by screening and, in the absence of screening, death from breast cancer. This cannot be directly observed in an ethically acceptable screening trial, because screen-diagnosed cases must always be treated. However, it can be deduced from a randomized controlled trial, since we presume that the same cancers are arising in the screened and unscreened groups, and that mortality differences are attributable to the differing therapeutic interventions that screening has made possible.

If the lag time L between screen-detectability and death was known and if it was the same in all cases, we could measure the mortality experience in the period between L years after the onset of screening and L years after the end of screening. As Miettinen et al. [2002] have pointed out, to achieve the full mortality reduction, a trial would need to have a duration at least as long as this lag time; inclusion of mortality data prior to year L or subsequent to year D + L would have dilutive effects, since some of the deaths occurring in these periods could not be affected by screening.[10]

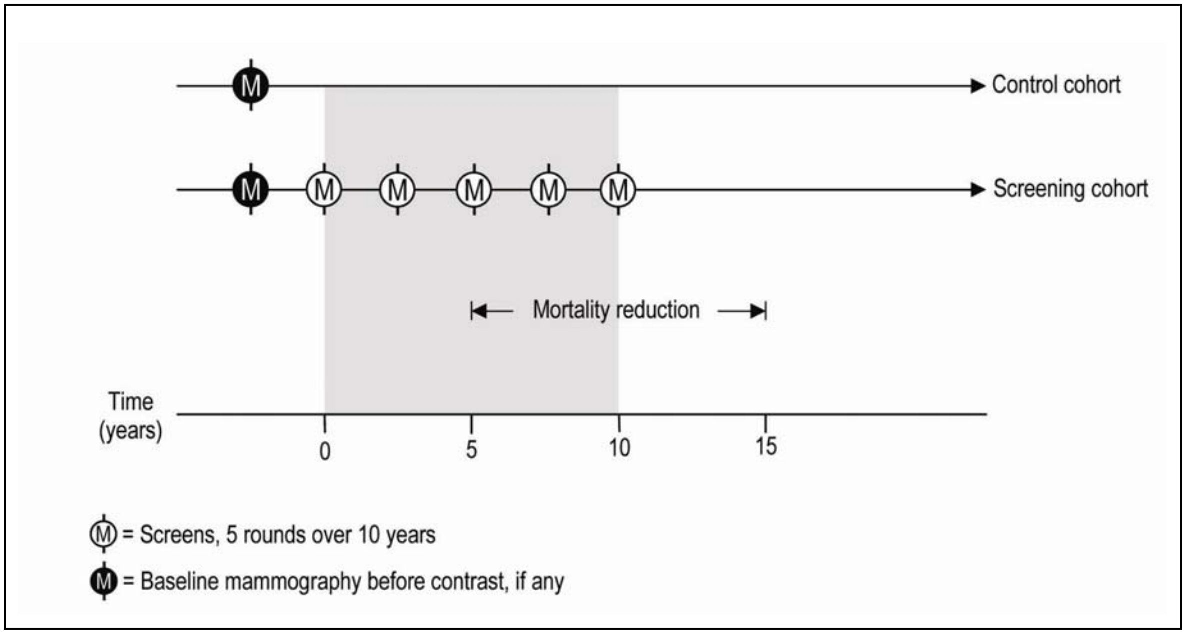
For example, if the lag time was always L = 5 years, and screening was conducted for D = 10 years, the period of mortality reduction would look like the diagram of Figure B-2.

Similarly, if the lag time was always 10 years, the mortality rate reduction would be progressively shifted 5 years farther into the future, as can be seen in Figure B-3.

In fact, breast cancer is a family of many different diseases arising in the same organ but from different structures, under different hormonal, immunologic and genetic circumstances, in different individuals, and with different screening and treatment regimens. Thus the lag time between the moment a cancer is detected by screening and the time it would become fatal is likely to show considerable variation. If we define L_{min} to be the shortest lag time, and L_{max} to be the longest, then the different lag times would be expected to vary continuously from L_{min} to L_{max}. We can thus expect that any reductions in mortality rates will only begin to appear L_{min} years after the initiation of screening, and that the full reduction will not be manifest before L_{max} years after initiation of screening.

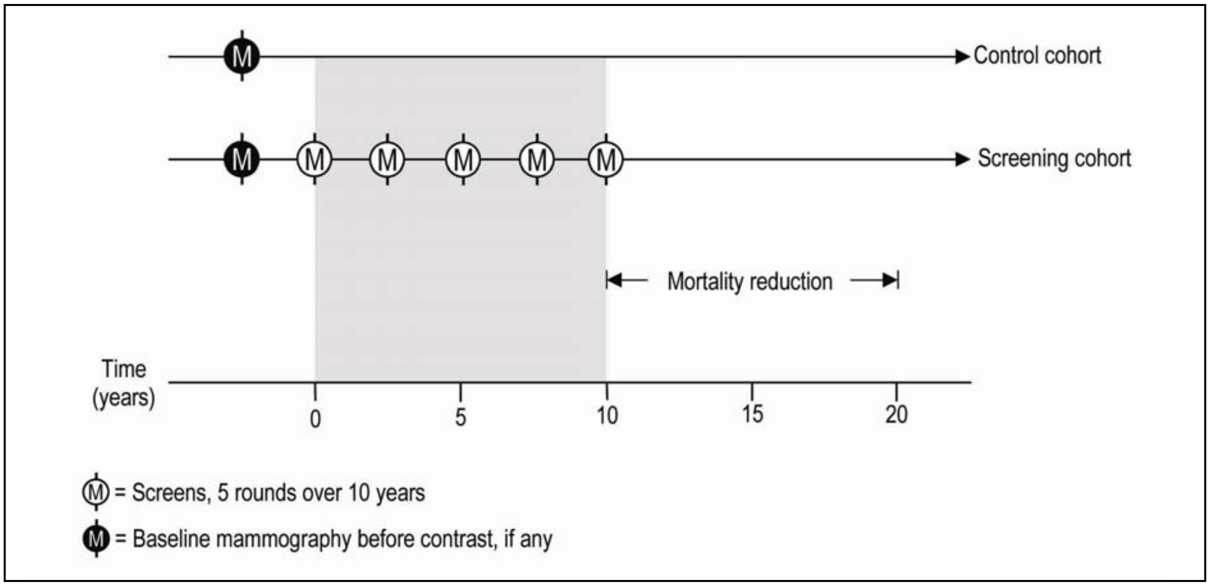
[10] Although screening may provide some benefit in detecting cancers later in their course, this effect is likely to be negligible.

FIGURE B-2
Paradigm for lag time = 5 years



M
Control cohort
M
M
M
M
M
M
Screening cohort
Mortality reduction
Time
(years)
0
5
10
15
M = Screens, 5 rounds over 10 years
M = Baseline mammography before contrast, if any

FIGURE B-3
Paradigm for lag time = 10 years

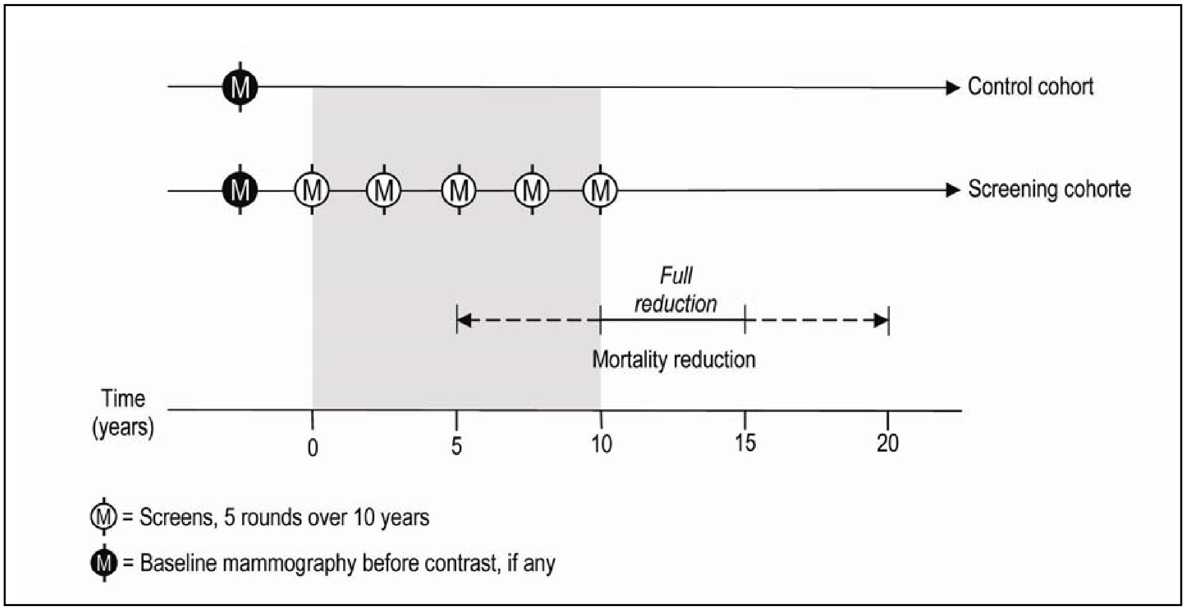


M
Control cohort
M
M
M
M
M
M
Screening cohort
Mortality reduction
Time
(years)
0
5
10
15
20
M = Screens, 5 rounds over 10 years
M = Baseline mammography before contrast, if any

When screening is stopped, the mortality rate reduction will continue to apply for some time, since new deaths will be arising from cases which would have been identified during the screening period. After L_{min} years, however, deaths will begin to appear from cases with L_{min} lag time and since these cases will have arisen (been detectible) after the contrast period D, there should be no fatality rate difference between the screened and control cohorts. By L_{max} years after screening, all new deaths will be from cases arising since the end of the screening period D, and the mortality rates of screened and unscreened should be equal. Once again, this return to normal rates will be not precipitous, but gradual. Using L_{min} = 5 and L_{max} = 10, we would expect the period of maximal reduction of mortality rates to be as in Figure B-4.

FIGURE B-4

Mortality pattern for lag time varying continuously between 5 and 10 years



Depending on the length of the screening period and the length of follow-up, follow-up time may thus be logically divided into five periods:

1. The first L_{min} years, with little or no difference in mortality rates.
2. The next $L_{max} – L_{min}$ years, with gradually increasing effect of screening.
3. The next $D – L_{max} + L_{min}$ years, with full effect.
4. The next $L_{max} – L_{min}$ years, with gradually decreasing effect.
5. Years subsequent to $D + L_{max}$, with little or no effect.

It is apparent from the preceding analysis that, although some reduction in mortality rates will appear between L_{min} and L_{max} years, the full reduction will not appear until L_{max} years after screening initiation. This was what the previous CETS [1993] report referred to as the program's 'steady state'. Furthermore, the full reduction ('steady-state') will be over by $D + L_{min}$ years, i.e., L_{min} years after screening termination. Since the steady-state period is between $D + L_{min}$ and L_{max}, the length of this period can be calculated as being simply $D + L_{min} – L_{max}$.

Whereas D can be fairly well approximated from the description of the study protocol, L_{max} and L_{min} are features of breast cancer in the population to whom screening is offered, and can be estimated by observing the point in screening studies where relative mortality rates for screened women begin to fall (L_{min}) and the point at which they stabilize at their lowest point (L_{max}).

ESTIMATION OF LAG TIME

An examination of survival curves of breast cancer in the absence of screening indicates that deaths occur in the first year and continue to occur as long as 20 years after initial diagnosis. However, since lag time L refers to the interval between screen-detection and death from screening-preventable disease, the relevant source of information for estimates of L must come from studies comparing screen-detected cases and their counterparts in a control group, i.e., from trial data, not natural history data.

A number of studies provide information which is helpful for the estimation of L_{min} and L_{max}. Data from the HIP study indicate that rates of breast cancer deaths began to diverge in year 4; the gap between the curves widened until year 6 or 7 and then remained fairly constant [Shapiro et al., 1982, p. 351]. This would correspond to $L_{min} = 4$ and $L_{max} = 6$ or 7. Interestingly, the favourable effect of screening appeared later among women aged 40–49 at entry than among older women [Shapiro, 1997, p. 30]. A similar result was observed in the Malmö study, where breast cancer death rates observed among women over 50 at entry (MMST-II) diverged as early as year 2, and the gap between the rates became steady around year 7. On the other hand, for women under 50 at entry (MMST-I), rates only diverged around year 6 and the gap became constant around year 11. The Two-County Study reports that "the mortality curves begin to separate at 5-6 years after randomization" [Tabár et al., 1999], which would correspond to $L_{min} = 5$ or 6. The Stockholm study [Frisell et al., 1997, Figure 1] showed mortality curves that were virtually identical up to 5 years after randomization, diverging thereafter up until around year 10 or 11, suggesting $L_{min} = 5$ and $L_{max} = 10$ or 11. In the most recent overview of all the Swedish trials, mortality curves diverged around year 5 or 6, and the absolute effect increased up to 12 years after randomization, whereafter it was maintained [Larsson et al., 1997]. This would correspond to $L_{min} = 5$ and $L_{max} = 12$. In this case, although mortality curves diverged much less in younger women, no clear pattern emerged as to whether the lag times before divergence were earlier or later than for older women.

Because the numbers of deaths observed in each study were fairly small, in some cases less than 10 deaths each year in each group, the estimation of lag times should not be entirely based on any one study.

While it is difficult to be precise about the minimum lag time, it is probably within the range of 2 to 6 years; likewise, L_{max} is probably within the range of 6 to 12 years. This imprecision is nonetheless preferable to the traditional approach, which ignores lag time and counts all mortality events occurring since randomization, making the implicit assumption that L_{min} is 0. In addition, the traditional model, which accumulates all mortality events occurring in follow-up time, involves the implicit assumption that L_{max} is indefinitely large. In fact, by giving equal importance to each period, from randomization to the end of follow-up time, the traditional model gives equal importance to each year of follow-up, implicitly assuming that mortality reductions in the first, fifth, tenth and even 15th years are of equal magnitude. Under the traditional model, dilution is sure to occur unless these extreme and unlikely assumptions apply. We believe it is more realistic to estimate L_{min} and L_{max} based on the best available evidence rather than assume that $L_{min} = 0$ and L_{max} is indefinitely long. Thus for the purposes of the following analysis, we will assume that $L_{min} = 5$ and $L_{max} = 10$.

INDEX OF TIMING DILUTION

Based on the above considerations, we have attempted to estimate the extent to which each study has been diluted in the analysis of its results. This analysis requires several assumptions about the distribution of mortality effects, and requires the estimation of the duration of each contrast, the number of years of follow-up in the latest published report from each trial, and the analytical technique used to calculate results. These aspects are discussed in more detail below.

An important assumption required in order to estimate dilutive effects is the lag time between screening and its effects on mortality. As we have discussed, we assume that $L_{min} = 5$ and $L_{max} = 10$; in addition, we assume that cases are equally distributed between these two extremes. We have discussed in Appendix A a number of considerations concerning the duration of the study contrast, and we apply these to each study to arrive at an estimate of each study's duration of contrast. Most trialists have reported results on numerous occasions as follow-up has continued; we use the longest follow-up reported for each trial.

For each year of follow-up, we calculate the proportion of mortality effects occurring in that year which could be attributable to the screening. Thus, for years of follow-up prior to lag time L, we do not anticipate any effect; the same is true for years of follow-up after D + L. The calculation is complicated somewhat because we allow L to vary between $L_{min} = 5$ and $L_{max} = 10$. Finally, in order to calculate what proportion of observed deaths occurred in years when they might have been preventable, we must assume the number of years that diagnosis of breast cancer is advanced by screening: we have estimated this to be two years, consistent with the discussion above. In addition to the year-by-year analysis, we compute, for each year of follow-up, the cumulative dilution which would have occurred from time zero up to that year.

Two major analytic techniques used by researchers in the screening trials must be considered: in the first, naïve analysis, all mortality events from time zero are considered. This inevitably produces considerable timing dilution. A partial refinement, dropping the first five years' follow-up data, is possible in some studies where the data are reported by year of mortality events. The second analytic technique, adopted by the three Swedish studies after Malmö, is to consider only deaths arising from cases diagnosed during the period of study contrast. This provides for lesser timing dilution, particularly after the screening contrast, although considerable dilution remains because of the diagnosis of advanced cases of cancer during the screening contrast. The results of an example of this analysis of timing dilution are given in Tables B-1 and B-2.

Parameters for calculation of timing dilution

Lag time (L)	
L_{min}	5 years
L_{max}	10 years
Advance of diagnosis (A)	2 years

TABLE B-1

Specific parameters of trials

TRIALS	CONTRAST DURATION (D) (IN YEARS)	FOLLOW-UP DURATION (IN YEARS)	ANALYSIS*	TIMING DILUTION
HIP	3.5	18	N	0.28
Malmö (MMST)	8.8	11	N	0.64
TCS	6.5	20	S	0.76
Edinburgh	6.4	14	N	0.67
NBSS-1	4.6	11–16	N	0.56
NBSS-2	4.6	13	N	0.56
Stockholm	4.6	11.4	N	0.65
Gothenburg	**6.4**	**14**	**S**	**0.73**
UK Age Trial	11.0	13	N?	0.73

*N: Naïve (diagnoses previous to randomization excluded, all breast cancer deaths during follow-up counted); S: Swedish evaluation model (counts only breast cancer mortality occurring in cases diagnosed during study period).

TABLE B-2

Calculation of timing dilution (Gothenburg example)

Follow-up year	RELEVANT FOLLOW-UP YEARS							FOLLOW-UP YEARS COUNTED							DILUTION*	
	LAG-TIME (L)*							LAG-TIME MINUS ADVANCED DIAGNOSTIC (L - A)							YEAR'S RELEVANCE†	CUMULATIVE‡
	5	6	7	8	9	10	Total index	3	4	5	6	7	8	Total index		
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
5	0	0	0	0	0	0	0	1	1	0	0	0	0	2	0	0
6	1	0	0	0	0	0	1	1	1	1	0	0	0	3	0.33	0.17
7	1	1	0	0	0	0	2	1	1	1	1	0	0	4	0.50	0.30
8	1	1	1	0	0	0	3	1	1	1	1	1	0	5	0.60	0.40
9	1	1	1	1	0	0	4	1	1	1	1	1	1	6	0.67	0.48
10	1	1	1	1	1	0	5	1	1	1	1	1	1	6	0.83	0.56
11	1	1	1	1	1	1	6	1	1	1	1	1	1	6	1.00	0.64
12	0	1	1	1	1	1	5	1	1	1	1	1	1	6	0.83	0.68
13	0	0	1	1	1	1	4	1	1	1	1	1	1	6	0.67	0.71
14	**0**	**0**	**0**	**1**	**1**	**1**	**3**	**1**	**1**	**1**	**1**	**1**	**1**	**6**	**0.50**	**0.73**
15	0	0	0	0	1	1	2	1	1	1	1	1	1	6	0.33	0.74
16	0	0	0	0	0	1	1	1	1	1	1	1	1	6	0.17	0.75
17	0	0	0	0	0	0	0	1	1	1	1	1	1	6	0	0.75
18	0	0	0	0	0	0	0	1	1	1	1	1	1	6	0	0.75
19	0	0	0	0	0	0	0	1	1	1	1	1	1	6	0	0.75
20	0	0	0	0	0	0	0	1	1	1	1	1	1	6	0	0.75

*0 = Total dilution; 1 = no dilution.

† The annual index corresponds to the quotient of the total (L) divided by the total (D – A) for each year of follow-up.

‡ The cumulative index corresponds to the quotient of the cumulative total (L) divided by the cumulative total (D – A); for example, in follow-up year 6, we divide (0 + 0 + 0 + 0 + 0 +1) by (0 + 0 + 0 + 1 + 2 + 3), giving 0.17.

Note: Numbers used in this example are for the Gothenburg study.

APPENDIX C

STRENGTH OF CONTRAST SCALE

INTRODUCTION

Screening mammography can only be effective if it is at least some minimum level of quality, employing proper equipment with trained personnel, is offered at an appropriate interval, and succeeds in eliciting the participation of the women to whom it is offered. Three additional conditions apply in the context of a screening trial: that the control population is not also screened (contamination) and does not receive some other effective screening regimen, and that the results are calculated by an analysis of mortality data from the correct follow-up period. We have considered these aspects under five categories, which are the technical contrast examined by each trial, the era, the quality of screening, the participation and contamination rates, and the timing of follow-up. We explain these concepts further in the sections below, justifying the aspects of a rating scale developed to measure the strength of contrast of each of the nine screening mammography trials. Finally, we show the results of the application of this scale to these nine trials.

Technical contrast (C)

Published mammography trials actually involve three different contrasts which are qualitatively different, which we have rated according to the intensity of their contrast. The three levels are:

3) mammography and clinical exam vs. no screening (strongest contrast);
4) mammography alone vs. no screening (average contrast); and
5) mammography screening vs. screening by clinical exam (weak contrast).

Era (E)

Mammography screening was first introduced in the 1950s, and the first dedicated mammography machine was developed in 1966, with continuing evolution of the technology since then. Some experts believe that major qualitative advances were made at the beginning of the 1980s [Sickles, 1997]. Four categories are considered: 1) before 1970; 2) 1970–1979; 3) 1980–1989; 4) since 1990.

Quality (Q)

As for the quality of the technical contrast, we have identified four components which can be assessed from information published by trial investigators. These components are:

6) quality control measures put in place by the researchers (including use of accredited facilities and double reading of positive films);
7) whether or not mammography readers are generally high volume (defined as readers who interpret more than 3,000 mammograms per year);
8) whether or not the screening involved one mammographic view, two, or some combination of one-view and two-view, such as the process whereby initial screens are with two views and

subsequent screens are with one view when this view has been deemed adequate at the initial screen; and

9) whether the interval between rounds of screening is large or small (with three categories defined: less than 18 months, 18 months up to but not including 24 months, and 24 months and greater).

The quality score was calculated as shown in Table C-1.

TABLE C-1
Quality rating

COMPONENT	SCORE
Quality control (2nd reading)	
Adequate	1
Inadequate	0
Reader expertise	
High volume	1
Low volume	0
Number of views	
2 (each round)	1
2 then 1 (according to round)	0.5
1 (each round)	0
Interval between mammographies	
< 18 months	1
18 < 24 months	0.5
> 24 months	0
TOTAL SCORE	**(between 0 and 4)**

Participation and contamination (P)

Participation and contamination are naturally expressed as percentages, and their difference, also a percentage, constitutes a reasonable estimate of the percentage of the screening contrast which the two cohorts actually experience. For example, if half of the screening cohort were to participate and half of the control cohort were to also obtain screening mammography outside the trial, it would be reasonable to say that there had been no contrast (50% – 50% = 0%); 100% contrast would require full participation in the screening cohort with no contamination in the control cohort. For any intermediate combination, the difference would give a contrast between 0% and 100%, as one would wish. This approach is consistent with a proposal by Glasziou [1992], who essentially proposes an approach (attributed to Newcombe) whereby the intention-to-treat estimate can be adjusted by a factor $1/\Delta$, where $\Delta = p_1 + p_2 - 1$, and p_1 and p_2 are compliance rates in the screened and control groups. Since p_2 is the compliance rate in the control group, $1\text{-}p_2$ is the contamination rate, and $p_1 + p_2 - 1 = p_1 - (1 - p_2)$, so the adjustment factor can be restated as $1/(p - c)$, where p is the participation rate in the screening cohort and c the contamination rate in the control cohort. A similar adjustment for non-attendance and contamination has also been called a ‘causal’ estimate [Baker et al., 2002].

Timing (T)

The timing index represents the expected degree of mortality reduction, expressed as a proportion of the full potential steady-state mortality reduction which the screening contrast can produce. This index is also between 0 and 1, with 0 signifying that none of the mortality potential reduction should be observable in the measured time period (e.g., in the first few years of follow-up), and 1 signifying that the full reduction should be observed (e.g., in the time period between L_{max} and $D – L_{max} + L_{min}$, as we have argued). Intermediate values can be thought of as a proportion of the potential mortality reduction, diluted by the deaths which cannot be affected to the screening contrast because of their timing. Appendix B (Timing of the effects of screening on breast cancer mortality) explains how this dilution can be estimated given assumptions about the lag time between screening and mortality effects.

COMBINATION OF SCORES

As for the combination of component scores, we believe that the choice of how to score each of these components requires judgment but should not be arbitrary. Since the point of the exercise is to differentiate between studies with an eye to concentrating on the most relevant ones (those with the contrast most like the screening program which might be implemented), it seems useful in principal to assign a score which corresponds to the degree with which a trial reflects the potential of a program of steady-state screening mammography to be efficacious in the current environment.

Participation rates and timing dilution rates are already expressed naturally as percentage rates. In addition, intuition suggests that zero-percentage participation would give zero impact and a 100% participation rate would give full impact. Likewise, full timing dilution (e.g., using only the first year’s follow-up data) would give zero impact, and no timing dilution would give full impact. It thus seems natural to express the three other components as percentages and to design a strength-of-contrast score as the product of the five component scores. In an additive scale, any given component scoring 0 would decrease the overall rating by 1 point out of 5, or 20%. In a multiplicative scale, any component scoring 0 would mean that the overall score is also 0. To maintain differentiation between scores, it is preferable to allow each component to vary only by the same percentage, i.e., 20%.

We thus assign a value of 1 to a component which corresponds to the contrast which is proposed in our original research question, i.e., for a screening mammography program without clinical exam, with modern equipment, quality control and two-view mammography at 24-month intervals, with participation at 100%, and considering the full steady-state effect with no timing dilution. Following this approach, we arrive at the following scores for the categories identified in Table C-2.

RATING THE STUDIES

Using the scale and scoring system shown in Table C-2, each study was independently assigned a rating by two researchers (WD and RK) with experience in the analysis of epidemiologic studies and familiarity with the field of mammography trials. For each of the nine published studies, all published articles were exhaustively reviewed. Information was obtained in the introductions (general conditions of the study, particularly baseline mammography rates), methods sections (mammographic equipment and techniques, quality control measures, intervals) and results sections (participation rates). In general, we assigned the lowest score when no information could be found about a given category. As for participation rates in particular, the average participation rate was calculated using the rates for the different rounds of screening. Interpolation was used when rates were only reported for first and last rounds. Participation rates in the control cohort have usually not been reported, since most studies did not obtain any information from unscreened women except their age; however, some information about general population rates of screening was usually available. In other cases, control participation (contamination) rates were estimated from general information about population rates of screening at the time and place where the study was conducted.

Ratings by the two reviewers were compared, and discrepancies were resolved by consensus with reference to the original research reports. Tables C-3, C-4 and C-5 contain the results of this rating exercise.

TABLE C-2
Strength of contrast ratings

COMPONENTS	CRITERIA	MULTIPLIER
Technical contrast (C)	Mammography + clinical breast exam vs. no screening	1.25
	Mammography vs. no screening	1
	Mammography vs. clinical breast exam	0.8
BEGINNING OF STUDY		
Era (E)	Before 1970	0.7
	1970–1979	0.8
	1980–1989	0.9
	Since 1990	1
TOTAL QUALITY SCORE		
Quality (Q) (See Table C-1)	0	0.8
	0.5	0.825
	1	0.85
	1.5	0.875
	2	0.9
	2.5	0.925
	3	0.95
	3.5	0.975
	4	1
DIFFERENCE BETWEEN COHORTS		
Participation – contamination (P)	Participation rate in screening cohort (rate in study, see Table C-4)	Observed difference
	Contamination rate in control cohort (rate in study, see Table C-4)	
TIME DILUTION INDEX		
Timing index (T)	Rates calculated as in Tables B-1 and B-2.	See Table B-1

TABLE C-3

Rating of each trial: Quality score

TRIALS	HIP	Malmö	TCS	Edinburgh	NBSS-1	NBSS-2	Stockholm	Gothenburg	*UK Age*	**Comparison PQDCS**
Quality control	1	0	0	1	1	1	0	0.5	1	0.5
Readers' expertise	1	0	1	0	0	0	0	0	1	0
Number of views	1	0.5	0	0.5	1	1	0	0.5	0.5	1
Interval between mammographies	1	0.5	0	0.5	1	1	0	0.5	1	0.5
Total quality score	4	1	1	2	3	3	0	1.5	3.5	2
Quality (Q)	**1**	**0.85**	**0.85**	**0.90**	**0.95**	**0.95**	**0.80**	**0.88**	**0.98**	**0.90**

TABLE C-4

Rating of each trial: Participation – contamination rate

TRIALS	HIP	Malmö	TCS	Edinburgh	NBSS-1	NBSS-2	Stockholm	Gothenburg	*UK Age*	**Comparison PQDCS**
Participation (screening cohort)	0.54	0.74	0.87	0.527	0.875	0.883	0.805	0.788	0.61	0.70
Contamination (control cohort)	0.05	0.10	0.05	0.10	0.125	0.065	0.20	0.15	0.04	0
Difference	**0.49**	**0.64**	**0.82**	**0.43**	**0.75**	**0.82**	**0.61**	**0.64**	**0.57**	**0.70**

TABLE C-5

Rating of each trial: Contrast strength

TRIALS	HIP	Malmö	TCS	Edinburgh	NBSS-1	NBSS-2	Stockholm	Gothenburg	*UK Age*	**Comparison PQDCS**
	(1963)	(1976)	(1977)	(1979)	(1980)	(1980)	(1981)	(1982)	(1991)	
Technical contrast (C)	1.25	1	1	1.25	1.25	0.80	1	1	1	1
Era (E)	0,70	0.80	0.80	0.80	0.90	0.90	0.90	0.90	1	1
Quality (Q)	1	0.85	0.85	0.90	0.95	0.95	0.80	0.88	0.98	0.90
Participation – contamination (P)	0.49	0.61	0.82	0.43	0.75	0.82	0.61	0.64	0.57	0.70
Timing index (T)	0.28	0.64	0.76	0.67	0.56	0.56	0.65	0.73	0.73	1
Total (C x E x Q x P x T)	**0.12**	**0.27**	**0.42**	**0.26**	**0.45**	**0.31**	**0.29**	**0.37**	**0.41**	**0.63**

APPENDIX D

VALIDITY SCALE

INTRODUCTION

We have discussed, in the main document, the principal design components of randomized controlled trials of screening mammography which are designed to protect the validity of their results. These are: the correct randomization to screening or control cohorts, the baseline equality of these two groups, the equality of subsequent exclusions from either cohort, and the equality of follow-up information obtained regarding breast cancer mortality. These criteria are similar to those which are commonly used in systematic reviews. For example, Jadad et al. [1996] have proposed a five-point scale, based on randomization, double blinding, and the equality of dropouts and withdrawals with regard to risk of developing the outcome in question. Similarly, Olsen and Gøtzsche [2001], in their examination of the mammography screening trials, assigned points based on the randomized method, baseline comparability, exclusions and randomization, unbiased outcome assessment, and screening in the control population [Mocharnuk, 2002].

The Jadad scale, which is widely used, assigns 2 out of 5 possible points based on double blinding. Since no mammography study has attempted blinding, either of patients or investigators, they all would rate 3 or less using this scale. Likewise, the equality of dropouts and withdrawals is difficult to assess in the mammography trials, since little is known about control cohorts in most studies, the Canadian studies being a notable exception. Most mammography studies would thus rate 0 or 1 out of 5. This scale would thus be of little use in differentiating between trials. It is, however, worth noting that all mammography studies fail to meet the highest standards that are generally applied to randomized trials.

The major purpose of blinding in clinical trials is to properly constrain the initial allocation of participants to receive the screening or control intervention so that this allocation is done randomly, i.e., without regard to their risk [Chalmers, 2001]. Indeed, some would feel that blinding procedures are the most important determinants of the quality of a randomized control trial [Chalmers et al*.*, 1981]. The major consequence of unblinded assignment to screening or control groups is that patients and their physicians may ‘self-select’ their exposure to mammography, based perhaps on their perceived risk of breast cancer or their expected benefit from mammography. This may account for the high levels of non-participation in the screening group and contamination in the control group that most studies have observed. We have already considered the issue of control screening as an issue of strength of contrast, discussed in Appendices A and C. The remaining four issues raised here also have the potential to bias the results of a trial, but unlike participation-related bias, it is difficult to know whether this bias would be toward the null hypothesis or away from it. Bias of unknown direction is considered here under the rubric of validity (even though we recognize that systematic bias towards the null hypothesis, and principally weaknesses related to the strength of contrast, are also a 'validity' issue). These aspects of validity, and their inclusion in our rating tool, are described below.

Randomization (R)

Three levels are defined: adequate, poor, and inadequate or undocumented, with associated point scores of 1, 0.5 and 0, respectively. It is generally believed to be preferable to randomize individuals, not practice groups or communities; if the latter is employed, then the analysis and confidence intervals should reflect this feature of the study design. In addition, randomization by definition involves random assignment (e.g., random number tables, coin toss) rather than systematic assignment (e.g., based on date of birth).

Baseline equality (B)

Three levels are defined: adequate, poor, and inadequate or undocumented, with associated point scores of 1, 0.5 and 0, respectively. An assessment that baseline equality is adequate requires some measurement of characteristics of screening and control cohorts that correspond to breast cancer risk or potential benefits of mammography. Most studies obtained virtually no information about the control cohort, apart from the dates of birth (and thus ages) of the women in that cohort. The equality of age alone is a necessary but insufficient measure of baseline equality, especially given the fact that blinding was not attempted and thus participants had de facto considerable control over the intervention they received. Olsen and Gøtzsche [2001] observed differences in average age of up to 3 months between screening and control cohorts, and used this criterion in judging that some trials were not successfully randomized. Some critics of their work have claimed that these age differences could not have accounted for major differences in breast cancer rates, and even that these age differences would tend to conceal the effectiveness of screening [Law et al., 2000, for example]. This criticism, we feel, misses the point that average age is the only evidence we have of the success of randomization in obtaining baseline equality. If age is not equal between groups, then randomization is unlikely to have adequately balanced other relevant risk factors between the two cohorts. Likewise, major differences in socio-economic status, measured in one study by obtaining information about random samples of the screening and control cohorts, would be evidence of inadequate baseline equality.

Equal exclusion (E)

Three levels are defined: adequate, poor, and inadequate or undocumented, with associated point scores of 1, 0.5 and 0, respectively. Of particular concern in these trials is the exclusion of cases of breast cancer with diagnoses prior to the screening contrast. Due to the regular visits of women in the screening cohort, investigators typically had more opportunities to obtain clinical information about previous diagnoses of breast cancer. Women in control cohorts were also excluded if a prior diagnosis of breast cancer was found, but since this was often 20 years later, the available clinical information may have been less thorough for these women, potentially resulting in a significant bias in favour of screening. For instance, Olsen and Gøtzsche [2001] calculated that, in the HIP study, 853 women were excluded from the screened group on this basis, but only 336 from the control group. As they state, if only a small proportion of these excluded cases are added as breast cancer deaths, breast cancer mortality becomes as high in the screened group as in the control group, since the difference between the two groups was only 44 deaths.

Follow-up (F)

Three levels are defined: adequate, poor, and inadequate or undocumented, with associated point scores of 1, 0.5 and 0, respectively. Factors considered in this assessment included the existence of a national mortality database for follow-up, and the validation of end-points using autopsy or case review.

COMBINATION OF SCORES

Once trials have been assessed on these dimensions, there remain the challenges of combining the component scores and using the total score in the analysis. We propose a simple sum of the four components, for want of any more compelling solution. This is consistent with frequently used techniques such as those proposed by Jadad et al. [1996].

RATING THE STUDIES

Once the components of a validity index have been chosen, it remains to measure each study according to these components. Of course, there remains a degree of subjectivity in deciding whether a given study meets a given component criterion, for example, whether random allocation, baseline comparisons or the treatment of exclusions are 'adequate' or whether the often summary descriptions of methods are sufficient to justify calling them 'documented' or not.

Using the scale and scoring system developed above, each study was independently assigned a rating by two researchers (WD and RK) with experience in the analysis of epidemiologic studies and familiarity with the field of mammography trials. For each of the nine published studies, all published articles were exhaustively reviewed. Information was obtained in the methods sections (randomization) and results sections (baseline equality). In general, we assigned the lowest score when no information could be found about a given category.

Ratings by the two reviewers were compared, and discrepancies were resolved by consensus with reference to the original research reports. Table D-1 contains the results of this rating exercise.

TABLE D-1
Validity ratings of published mammography trials

VALIDITY INDEX	TRIALS								
	HIP	Malmö	TCS	Edinburgh	NBSS-1	NBSS-2	Stockholm	Gothenburg	*UK Age*
Randomization (R)	1	1	0	0	1	1	0.5	0.5	1
Baseline equality (B)	0.5	0.5	0	0	1	1	0.5	0	1
Exclusions (E)	0	0.5	0	0	1	1	0.5	0.5	1
Follow-up (F)	0	1	0.5	0.5	1	1	0.5	0.5	1
R + B + E + F	1.5	3	0.5	0.5	4	4	2	1.5	4
Olsen and Gotzsche [2001]	"*Flawed*"	"*Medium*"	"*Poor, likely flawed*"	"*Flawed*"	"*Medium*"	"*Medium*"	"*Poor*"	"*Poor*"	
US Preventive Services Task Force [Humphrey et al., 2002]	"*Fair*"	"*Fair*"	"*Fair*"	"*Poor*"	"*Fair +*"	"*Fair +*"	"*Fair*"	"*Fair*"	

1 = adequate
0.5 = poor
0 = inadequate or undocumented

APPENDIX E

REVIEW OF THE NINE TRIALS

1. NEW YORK (HIP)

The first formal mammography trial, generally referred to as the HIP Mammography study, was conducted among clients of a private health insurance company of Greater New York, the Health Insurance Plan (HIP). It was launched in December 1963, with continuing enrolment until June 1966. Screening ended in June 1970, and follow-up of results continued up to 1986, when the study was formally closed. Major reports on study design, conduct and results were published between 1966 and 1997 [Shapiro, 1997; 1994; 1989; Shapiro et al., 1988; 1985; 1982; Shapiro, 1977; Strax et al., 1973; Fink et al., 1972; Shapiro et al., 1966].

Contrast

The screening intervention studied was three to four annual two-view mammographies with clinical breast examination (CBE). Participation rates for the first round of screening have been variously reported as being 64% [Shapiro et al., 1966], 65% [Fink et al., 1972] and 67% [Shapiro et al., 1985]. Three subsequent rounds at one-year intervals were planned. Only women who participated in the first round were invited to subsequent rounds, and so rates in these rounds fell to 52%, 48% and 45%, for an average of 52%. Women in the control group "followed their usual practices in obtaining medical care" [Shapiro, 1989]. Their rates of use of mammography or breast exam were not measured, but the authors state that use of screening mammography was low in the control group, since this was not a covered benefit in the HIP at that time, nor was regular mammography a standard component of clinical care in the 1960s.

Aware of the dilutive effects of further and further follow-up, the HIP's most recent report appropriately offers analyses restricted to cases diagnosed in the five or seven years following screening. This has the effect of removing much of the dilutive effect of later years, when cases identified subsequent to the contrast progress to death. However, early dilutive effects, due to cases diagnosed too late in their course for mammography to have any protective effect, would remain.

The study scored 12% on our strength of contrast scale, the lowest score of all published trials. In particular, the mammography equipment, which preceded the introduction of dedicated mammographs, the short contrast, timing dilution effects and low participation rates all contributed to the low rating.

Validity

Randomization was conducted using individual allocation within pairs matched by age, size of insured family and employment group [Shapiro et al., 1966]. The methods for generation of the allocation sequence were unclear: one report describes the method by which, within each medical group, "2 systematic random samples" were selected, and "every nth woman was placed in the study group, the paired (n + 1) woman in the control group." The pairs of subjects and controls in a medical group were randomized, and study women were drawn in sequence from the list [Shapiro et al., 1985]. The most important post-randomization exclusion was for women with previous breast cancer, but this was unequally applied, since "this status was most completely ascertained for the screened women" [Shapiro et al., 1985]. Olsen and Gøtzsche [2001] calculate that 853 women were excluded from the screened group on this basis, but only 336 from the control group. As they state, if only a small proportion of these excluded cases are added as breast cancer deaths, breast cancer mortality becomes as high in the screened group as in the control group, since the difference between groups was only 44 deaths. In reply, Miller [2001], one of the investigators, claims that patients in the control group with cancers diagnosed previous to randomization were excluded at the 18-year follow-up, based on hospital records. However, he did not mention the number of patients excluded on this basis, and conceded that the late exclusion did not eliminate the inequality identified by Olsen and Gøtzsche. The authors of the study themselves acknowledge the problems with follow-up, pointing out that "the utility of HIP medical records is reduced as subscribers leave the Plan because of changes in employment, move out of the service areas, and transfer to other types of coverage." By 15 years, the proportion no longer in HIP was well over one-half [Shapiro et al., 1985].

Other baseline differences found by Gøtzsche and Olsen [2000] included previous lump in the breast, menopause and education. Many responses to the Gøtzsche and Olsen meta-analysis have advanced the argument that small baseline differences could not have explained the major reductions in mortality rates, and this is likely true [McTiernan, 2002]. However, they miss the point, which is not that these baseline differences *per se* alter the comparison, but rather that they reveal flaws in the randomization process, putting in doubt the adequacy of randomization for controlling other more important risk factors for breast cancer which are difficult to measure, such as family history of breast cancer, reproductive histories, hormone use, etc.

Using our rating scale for validity, the study scored 1.5 out of 4, largely because of difficulties in equally excluding previous cases of breast cancer and adequately following up future cases in a database of insurance plan members.

2. MALMÖ (MMST)

The Malmö Mammographic Screening Trial (MMST), conducted in Malmö, Sweden, began enrolment of patients in the Fall of 1976. Published reports on study design, conduct and results were released in 1979, 1981, 1988 and, for women under 50, in 1997 [Andersson and Janzon, 1997; Andersson et al., 1988; Andersson, 1981; Andersson et al., 1979]. The first report indicates that the population eligible for the study was women 50–69 years old living in Malmö [Andersson et al., 1979], but the second report states that participation in the trial was offered to women 45–69 years old. A later publication, referring to the MMST, mentions that separate analyses were conducted on women 55 and over at entry and younger women [Andersson et al., 1988]. A final report [Andersson and Janzon, 1997] refers to two separate cohorts, MMST-I (women under 50 at entry, i.e., birth cohorts 1927–1932) and MMST-II (women born between 1933 and 1945). Finally, the recent Swedish overview refers to the "continuation of the Malmö trial (MMST-II)" [Nyström et al. 2002], explaining that MMST-I was the part of the trial for which enrolment closed in 1978, whereas MMST-II was conducted from 1978 to 1990 and comprised women who, as they reached age 45, were continuously randomized according to the same protocol. MMST-II was conducted without a formal protocol [Olsen and Gøtzsche, 2001] and results have never been separately reported. Although the distinction between MMST-I and -II is certainly reported in a confusing way, they do not seem to be separate trials but rather cohorts distinguished well after the conclusion of the study contrast. MMST-I had more rounds of screening than MMST-II merely because the younger MMST-II cohort could only begin screening upon reaching the age of 45. For instance, the 1945 birth cohort could only begin screening in 1990, the year the study ended, and thus could only participate in one round of screening.

Contrast

The screening intervention involved double-view mammography using craniocaudal and oblique views in the first two rounds; in subsequent rounds, patients with fatty parenchymal pattern on previous mammography received only oblique views. No clinical exam was included in the screening intervention. Participation rates were 74% in the first round, and are reported indiscriminately as being 70% in subsequent rounds [Andersson et al., 1988]; another report indicates that the "attendance rate varied between 75% and 80%" [Andersson and Janzon, 1997]. We will assume an average participation rate of 71%. Control patients were not screened but those with breast complaints 'attended the ordinary medical service' [Andersson, 1981]; rates of mammography use in this group were not measured, but were presumably low, since mammography screening was not routinely available in Sweden at that time, and 24% of controls had at least one mammogram during the study period, most only once [Andersson et al., 1988]. We have estimated the rate of regular screening in the control group to be 10%. Women in the first cohort (birth years 1927–1932) had six rounds of screening; women in the second cohort (1933–1945) had less, the youngest having one round only. The interval between rounds of screening was 18–24 months, and the average length of the screening contrast was a total of 8.8 years. The reporting of mortality results by the study year in which they occurred [Andersson et al., 1988, Table IX] allows for further analysis of undiluted results in these years.

Using the rating scale for strength of contrast, the MMST-I trial scored 27%. It is unfortunate that results have not been separately reported with longer follow-up, since the relatively long study contrast would allow for longer follow-up with less dilution.

Validity

All women 45–69 living in the city of Malmö were randomly allocated by computer on an individual basis within each year of birth [Andersson, 1981]. Half were assigned to screening (21,242) and the other 50% (21,240) to the control group. Women with a case of breast cancer diagnosed before randomization were excluded, reducing the screened group by 393 and the control group by 412 [Gøtzsche and Olsen, 2000]. Baseline age data were similar in the screening and control groups, which argues in favour of adequate randomization, but baseline data are not available for other risk factors. Cause of death assessments were blinded for women with a diagnosis of breast cancer up to 1988 [Gøtzsche and Olsen, 2000], and autopsy rates were relatively high, at 76%, as reported in 1988, although they dropped to 40% in a later report [Andersson and Janzon, 1997].

MMST-II, conducted with no formal protocol, has not been separately reported, and it is much more difficult to assess the validity of this semi-official trial. Important administrative errors have been described, both in reports of the trial [Andersson and Janzon, 1997] and based on personal communications with the authors [Gøtzsche and Olsen, 2000]. It was not originally included in the Swedish review (Nyström et al., 1993], but has been included in the most recent review [Nyström et al., 2002], with the mention that the protocol, identical to the MMST-I protocol, could not be strictly adhered to. No baseline data are available, nor detailed information about exclusions, nor information about end-point validation such as autopsy rates. We have thus considered only the MMST-I part of the trial for which full information is available.

Using our rating scale for validity, the MMST-I trial received a score of 3 out of 4. This was one of only two trials rated as high as 'medium quality' by Gøtzsche and Olsen, as regards validity concerns.

3. TWO-COUNTY STUDY

The Two-County Study (TCS) was originally described in 1981 [Tabár and Gad, 1981] as the Swedish Trial, and is the second of four trials conducted in Sweden. It was conducted beginning in Kopparberg County in 1977 and in Östergötland County in 1978, two counties in central south-east Sweden representing together about 8% of the total Swedish population. The study is also sometimes referred to as the WE study, Kopparberg being the W (Western) study and Östergötland the E (Eastern) of the two. Its authors insist that it was planned as one study, and indeed the screening protocols were identical in all important respects. Reports were published between 1981 and 2001 [Tabár et al., 2001; 2000; 1999; 1997; 1995a; 1995b; 1992a; 1992b; 1992c; 1989; Fagerberg and Tabár, 1988; Tabár et al., 1985a; 1985b; Fagerberg et al., 1985; Tabár and Gad, 1981].

Contrast

Like the other Swedish studies, this study sought to evaluate single-view mammographic screening, using the medio-lateral oblique incidence, without clinical breast examination. Also like the other Swedish studies, it used a combination of one central unit and several mobile units, with central reading of all films, in this case by the two radiologists who were the principal researchers (Fagerberg and Tabár]. The eligible population was women 40–79 living in the two counties, although due to low participation among women over 75, a subsequent report restricted the analysis to women 40–74 [Fagerberg et al., 1985], and the restriction was later extended to women 40–69 [Tabár et al., 2000]. Participation by study women was high, with 91% of women 40–74 (both counties combined) complying with the program in the first round, 86% in the second and 84% in the third [Fagerberg and Tabár, 1988]. Average participation was thus approximately 87%. Women in the control group were not offered mammography, and their rate of use of mammography was not verified. Contamination rates may have been as high as 13% for this study, since "13% of women in the control group had a mammogram as part of routine medical care up to the end of 1984" [Tabár et al., 1985a], that is, prior to the screening contrast. The screening interval was two years for women 40–49 at entry, and three years for older women. Although accounts differ somewhat, it seems that a five-year intervention period was planned, with a stopping-rule based on statistical significance testing every six months.

The controversy surrounding the details of the screening protocol and its execution are testimony to the inadequacy of the documentation of this study. It is impossible to interpret with any confidence the start or finish of the programs in either Kopparberg or Östergötland based on published reports. It seems, however, that the Kopparberg arm began in October 1977 and continued until September 1982, when screening was offered to the control group. The Östergötland arm began in May 1978, lasted five years as well, and screening was offered to the control group in 1984, but perhaps as late as 1986. Further analysis of the timing is made difficult by the fact that results are reported by study calendar year, despite the fact that the Östergötland arm began one year later; thus results at any given follow-up time are a composite of two arms of different length. In any case, it seems clear that the time between randomization and the screening of the control cohort was between 5 and 8 years. As was the case for the HIP study, full steady-state mortality reduction cannot be observed for a trial with such a short screening contrast. However, since the trial was analyzed using the Swedish evaluation model, where only women diagnosed with cancer during the study contrast were followed for mortality results, and since exceptionally long follow-up is available (20 years), the dilutive effect of including early years of follow-up in cumulative mortality figures is less important than in other studies with short contrasts.

Using the rating scale for strength of contrast, we gave this study a score of 42%.

Validity

Allocation of women to study or control groups was described by the authors as randomization, but the actual process used was cluster allocation by a complicated process which can better be described as creative than random. The two counties of Kopparberg and Östergötland were first divided into blocks according to area of residence, and each of these blocks was further divided into smaller units based on parishes and municipalities. These smaller units were constituted in such a way as to ensure that all units inside each block had "similar socioeconomic and geographic circumstances" [Tabár and Gad, 1981]. The smaller units were 'randomly assigned' to screening or control group, by a process not described by the authors, although a coin toss may have been used for the clusters in Östergötland [Olsen and Gøtzsche, 2001]. Exclusion rules have not been described in published reports, and may explain the differences in the number of women randomized in both counties which appear from one published report to the next. Data on baseline comparability of the two cohorts have not been published. Calculations done by Gøtzsche and Olsen, based on the different numbers reported in different reports, indicate minor differences in baseline characteristics such as pre-trial breast cancers (more of them excluded in the control group) and age (study group 0.27 to 0.45 years older). Both would tend to reduce the demonstrated effect of screening on mortality reduction, but they indicate serious problems in the execution of the trial and its reporting.

Using our rating scale for validity, we gave this trial a score of 0.5 out of 4.

4. EDINBURGH

Edinburgh was one of the centres in the non-randomized UK Seven-year Trial of Breast Screening, begun in 1979, in which Edinburgh and another district were assigned to mammography screening, with other centres assigned to clinical examination or no intervention. The Edinburgh portion of this study was subsequently extended to become a randomized trial with its own control population within the city [Roberts et al., 1984], and given the name Edinburgh Randomised Trial of Screening for Breast Cancer. Publications describing the study and its results were released between 1981 and 1999 [Alexander et al., 1999; Alexander, 1997; Alexander et al., 1994; Roberts et al., 1990; UK Trial of Early Detection of Breast Cancer Group, 1988; Roberts et al., 1984; UK Trial of Early Detection of Breast Cancer Group, 1981].

Contrast

The screening contrast involved an invitation to a series of four rounds of screening mammography accompanied by clinical breast examination (CBE) every two years, with CBE alone in the intervening years. Over a period of 2,5 years between 1979 and 1981, all women 45–64 in the screening group were offered screening, and later entrants were invited to join the program during the study as they turned 45 or upon moving into the city or joining a study practice: the latter women had fewer episodes of screening. Mammography was two-view on the initial screen (oblique and craniocaudal views), with oblique view only in subsequent rounds. Different from other studies, most films were not read by radiologists, but rather by 'specially trained doctors'; radiologists read all abnormal films and a random 5% of all films. Screening group women received CBE 'as a standardized procedure' by a doctor or a nurse, and they were taught breast self-examination (BSE) and encouraged to carry it out once a month between visits. For women in control practices, no specific protocol is mentioned, although they were eligible to receive a leaflet about BSE. Participation rates in the screening group were only 61% in the first round, and only 44% attended the final screen [Alexander et al., 1999]; participation rates in intermediate rounds have not been reported. We estimate an average participation rate of 52%.

Rates of screening for women in the control group were not measured, although the investigators believed “that very few women in the control group arranged screening for themselves’, although they had ‘no way of confirming this assumption” [Alexander et al., 1999]. The mammography screening schedule was for years 0, 2, 4 and 6, but since some women enrolled later in the study the weighted average length of contrast was 6.4 years; follow-up was conducted for up to 14 years post randomization [Alexander et al., 1999].

Using the rating scale for strength of contrast, we gave this study a score of 26%.

Validity

Women were recruited through 348 general practitioners in Edinburgh who agreed to participate in the trial. Practices were stratified by district, and ‘randomly selected’ by one of the investigators. Subsequent adjustments included assigning practices operating from the same premises to the same trial status, and unintentional errors in communicating trial status to practitioners were allowed to stand. Some practices initially selected as controls had their status changed in order to increase the screening population. Exclusion of previously diagnosed breast cancer cases appears to have been markedly unequal, as 338 such women were excluded from the study group and only 177 from the control group. Comparison of baseline information is not possible, since epidemiological or demographic factors were “recorded only for cancer cases and in those women who attend for screening” [Roberts et al., 1984]. However, when an imbalance in all-cause mortality rates of screening and control groups was observed in 1983, socio-economic status was estimated based on census data derived from postal codes [Alexander et al., 1999]. These analyses suggested major differences in the two groups, with 26% of women in the control group and 53% in the screening group belonging to the highest socio-economic status level [Alexander et al., 1994]. Furthermore, women in the screening group had 15% lower all-cause mortality rates, suggesting better health status.

Using our rating scale for validity, we gave the trial a score of 0.5 out of 4, the lowest rating of all trials along with the Two-County Study.

5. CANADA-1 (NBSS-1)

The Canadian National Breast Screening Study (NBSS) was conducted beginning in January 1980 at 10 centres in 6 Canadian provinces, according to two different protocols defined for women 40–49 years old at entry (NBSS-1) or 50–59 years old at entry (NBSS-2). Articles about NBSS-1 were published between 1981 and 2002 [Miller et al., 2002; 1997; Bailar and MacMahon, 1997; Miller et al., 1992a; 1991; Baines et al., 1990; Miller, 1988; Baines et al., 1986; Miller et al., 1981].

Contrast

The screening contrast in NBSS-1 was between screening and control interventions defined as follows. The screening intervention consisted of an initial physical examination (P) conducted by specially trained nurses or physicians, followed immediately by two-view screening mammography (M). Repeat screens involving both physical examination and mammography, offered at one-year intervals, for a total series of four or five rounds of screening. The cohort to whom this intervention was offered was referred to as the “MP group”. Women assigned to the control intervention received an initial physical examination (prior to randomization) with annual follow-up through a mailed, self-administered questionnaire. The cohort to whom this intervention was offered was referred to as the “usual care group”, or “UC group”. In addition, women in both groups were taught breast self-examination (BSE); for those who returned to the screening centres for examination (presumably women in the MP group), teaching of BSE was reinforced [Miller et al., 1992b]. The screening contrast can thus be summarized as baseline physical examination and teaching of BSE, with mammography only as needed in the context of diagnosis, versus the addition of mammography in round 0, and additional mammography, physical examination and BSE reinforcement in subsequent rounds.

The proportion of MP-group women participating in at least part of the full proposed intervention (mammography and physical examination) was 100% in the first round, 89.4% in the second round, declining to 85.6% in the fifth round, although 2 to 3% of MP-group women accepted physical examination but refused mammography. Between 93 and 95% of the participants in the UC group returned their annual questionnaires, and 26.4% of UC-group women had at least one mammogram outside the study protocol during the study period, increasing from 7.0% between years 1 and 2 to 18.1% between years 4 and 5. The first 62% of women entering the study began early enough to participate in a four-year program (all five rounds), and the remaining 38% were offered a three-year program. We can thus estimate the length of the screening contrast as an average of 4.6 years. The longest published follow-up for NBSS-1 is for an average follow-up of 13 years, with a range of 11 to 16 years [Miller et al., 2002].

Using the rating scale for strength of contrast, we gave this study a score of 45%. Data is presented which allows for an analysis using something like the Swedish model, where only cases diagnosed within the first five years are followed up for breast cancer mortality. When data is censored in this way, there is much less timing dilution, and the strength of contrast using this analysis would be 75%. However, this model may cause some lead-time bias, since women in the screening group would be more likely to be diagnosed with breast cancer because of their greater exposure to mammography, and we have thus not used the latter analysis here.

Validity

Volunteers recruited by various means were individually randomized to the MP or UC groups. Previously diagnosed cases of breast cancer were excluded, based on the initial questionnaire that was administered prior to randomization. An independent review of the randomization process concluded that its execution was successful [Bailar and MacMahon, 1997]. Cases identified by the initial examination were not excluded from follow-up or analysis, and post-randomization exclusions were balanced between groups. Baseline equality between groups was achieved regarding 10 risk factors of importance [Miller et al., 2000a; 1992a]. Although the study has been criticized because women in the screened group had more small node-positive cancers, this is consistent with the fact that mammography is expected to identify such tumours [Bailar and MacMahon, 1997]. Women with breast cancer identified by CBE prior to randomization were excluded from both groups, but women with breast cancer subsequently detected by mammography should not be excluded in the screened group, since mammography is conducted precisely with the objective of identifying cancers, and any exclusion criteria would have to apply equally to each group to maintain an unbiased comparison [Kakuma, 2002]. Follow-up during the study contrast period was conducted at the time of examination (MP group) or by mailed questionnaire (UC group). In both groups, women with a diagnosis of breast cancer were followed up annually by the NBSS central office, and new diagnoses for other women were retrieved by linkage with provincial cancer registries and mortality databases. All cancer deaths were reviewed blindly by an expert panel using pathological reports, hospital records and autopsy reports.

Using our rating scale for validity, we gave this trial a score of 4 out of 4.

6. CANADA-2 (NBSS-2)

The second part of the Canadian National Breast Screening Study (NBSS), concerning women 50 to 59 years old at entry, also began in January 1980. NBSS-2 was conducted in the same locations and over the same time period as NBSS-1. Articles on NBSS-2 were published in 1981, 1991, 1992, and 2000 [Miller et al., 2000b; 1992b; 1991; 1981].

Contrast

In NBSS-2, different from the NBSS-1, the design of the trial involved a contrast between two screening interventions. The first screening intervention involving mammography, as in NBSS-1, consisted of an initial physical examination (P) followed immediately by screening mammography (M), with repeat screens involving both physical examination and two-view mammography, offered at one year intervals, for a total series of four or five rounds of screening. The cohort to whom this intervention was offered was referred to as the “MP group”. Women assigned to the contrasted screening intervention received an initial physical examination (prior to randomization) with subsequent annual physical examination, but without mammography. The cohort to whom this intervention was offered was referred to as the “PO group” (Physical examination Only). In addition, women in both groups were taught breast self-examination; and, unlike the NBSS-1, teaching of BSE was reinforced for both groups at the time of annual physical examination [Miller et al., 1992b]. The screening contrast can thus be summarized as baseline physical examination and teaching of BSE, along with annual high-quality physical examination, with mammography only as needed in the context of diagnosis (control group), versus the addition of mammography in each round (screening group). Compared to NBSS-1, the contrast is substantially weaker, since women in the control group received a screening regimen which may be efficacious in itself.

The proportion of MP-group women participating in the full proposed intervention (mammography and physical examination) was 100% in the first round, 90.4% in the second round, declining to 86.7% in the fifth round. In addition, from 2 to 3% of MP-group women accepted physical examination but refused mammography. In the PO group, compliance with annual physical examination was 100% in the first round, 89.1% in the second round, declining to 85.4% in the fifth round; 16.9% of PO-group women had at least one mammogram outside the study protocol during the study period, increasing from 5.3% between years one and two to 8.0% between years four and five. The timing of the screening contrast was identical with that of NBSS-1, giving a contrast of 4.6 years’ duration, resulting in substantial dilution in all years of follow-up.

Using the rating scale for strength of contrast, we gave this trial a score of 31%. Data is presented which allows for an analysis using something like the Swedish model, where only cases diagnosed within the first five years are followed up for breast cancer mortality. When data is censored in this way, there is much less timing dilution, and the total strength of contrast using this analysis would be 54%. However, this model may cause some lead-time bias, since women in the screening group would be more likely to be diagnosed with breast cancer because of their greater exposure to mammography, and we have thus not used the latter analysis here.

Validity

The randomization, exclusion procedures, baseline comparison of the two groups and outcome assessment were virtually identical to those used in NBSS-1, and the same analysis mentioned above for NBSS-1 applies to NBSS-2 in all regards.

Using our rating scale for validity, we also gave this trial a score of 4 out of 4.

7. STOCKHOLM

The Stockholm Mammographic Screening Trial was initiated in March, 1981 in Stockholm, Sweden. Articles on this trial were published between 1986 and 1997 [Frisell et al., 1997; Frisell and Lidbrink, 1997; Frisell et al., 1991; 1989; 1986].

Contrast

The screening intervention consisted of only two rounds of single-view (oblique) mammography, with no physical exam, 28 months apart. Participation in the first round was 81%, with 80% participating in the second round. Women in the control group were not offered mammography or physical examination, but were invited to mammography screening after the completion of the second round of mammography in the screening group. The participation in this control round of mammography was 77% [Frisell et al., 1997]. It is not known what percentage of women in the control group had screening mammography during the study contrast, but it may have been at least 25%, since this was the rate of screening mammography in Stockholm women in the three years before the trial [Frisell et al., 1991]. The first round of screening began on March 9, 1981 and took 2.5 years to complete; the second round began on September 1, 1983 and took 2.1 years to complete [Frisell et al., 1991]. During 1986, the control group was invited to screening mammography. The screening contrast was thus 4.6 years.

Analysis of the extent of dilution is complicated by the decision that “After 31 Dec. 1986, no more cases were brought into the study”, i.e., 1 to 3.5 years after the end of the screening contrast. This means that accrual of cancer cases was truncated by calendar time (a little more than one year after the end of the second round), but in study time, this amounts to anywhere from one year after the second screen (for women screened at the end of the second round) to 3.5 years (for women screened at the beginning of the second round). This truncation avoids most of the problem of dilution by late cases, i.e., cases occurring after the screening contrast. This method, referred to by the Swedish trialists as the ‘evaluation method’ [Nyström et al., 1993] is also used in the two overviews of the Swedish studies. As a result, further follow-up should not be expected to significantly dilute mortality results.

Using the rating scale for strength of contrast, we gave this study a score of 28%.

Validity

Women were systematically assigned to screening and control groups on the basis of the day of their birth, with women born in the first 10 days of the month or between the 21st and 30th (or 31st) [11] assigned to screening. Technically speaking, this study is not randomized, although all reports describe it as such. While date of birth is evidently not a risk factor for breast cancer, such

[11] “Women born on the 1st to the 10th of each month or on the 21st to the 31st were assigned to the screening group” [Frisell et al., 1986], but in another publication the assignment is described as applying to women born on the 21st to the 30th of each month [Frisell et al., 1989]. This reporting error may account for part of the slight excess of women in the screening group noted by Olsen and Gøtzsche [2001] (40,318), compared to the 40,000 that would be expected if the latter assignment strategy were used.

a transparent assignment protocol lends itself to self-selection both because the reported date of birth may not be adequately verified and because women who want to be screened and who are born between the 11th and the 20th of the month may be less likely to participate in the study if they or their physician are aware of the assignment schedule. Since most studies have concluded that women who participate in screening programs have higher risk, this potential bias would be expected to be in favour of the null hypothesis.

What appears to be rounding in the reporting of this study makes it difficult to analyse the data from this study. For instance, according to Frisell et al. [1986], “the study was based on all the 60,000 women in the SE part of Greater Stockholm”. However, from the same article, “the study population (SP) totalled 40,318 women”; the only mention of the size of the control group is contained in Figure 1, where those not invited are labelled “control group c. 20,000”. A later paper mentions that “the Stockholm study comprises a total of 60,261 women […]” and that 19,943 women make up the control population [Frisell et al., 1989]. Further (Table 6), we are given the number of new cases per size of population for different years after randomization, and all 20 denominators happen to be multiples of 100. In a third report [Frisell et al., 1991], we still have a study group of 40,318 women, but the control population is down to 19,343 women, without explanation. This may be a typographical error, since the number 19,943 appears in Table 1. While rounding of quantitative information is appropriate in informal use or to express approximate information where absolute precision does not matter, its use in the reporting of study conduct and results impair the inspection of reported trial data for such aspects as the adequacy of randomization, participation, the completeness of follow-up and the results obtained.

The baseline comparability of screening and control groups has not been reported. Exclusions after randomization were made based on a diagnosis of breast cancer made before the study began and after 1986 [Frisell et al., 1997], but the number of such cases in screening and control groups was not provided. As in the Two-County Study, breast cancer mortality is defined as death with breast cancer present at death. This outcome measure seems biased against screening, since cases are more likely to have been diagnosed in women who have had mammography, whether or not these cases were actually responsible for mortality. The autopsy rate of 23% is the lowest of the Swedish trials, although it compares favourably with the Canadian trials.

Using our rating scale for validity, we gave this trial a score of 2 out of 4.

8. GOTHENBURG

The trial conducted in Gothenburg, Sweden, also referred to as the Gothenburg Breast Screening Trial, was initiated in December, 1982. Results for women 39–49 years old were published in 1997, and more recent results for women 39–59 years old were published in 2003 [Bjurstam et al., 2003; 1997].

Contrast

The screening intervention consisted of up to five rounds of screening mammograms at planned intervals of 18 months between rounds. Two views were used for 70% of exams, and one view in 30% where a previous two-view screen indicated that single-view mammography would be adequate [Bjurstam et al., 1997]. No clinical breast examination, training or encouragement in breast self-examination was offered. Participation rates were 85%, 78%, 79%, 77% and 75% in the five rounds. Women assigned to the control group were not invited to mammography, but a survey of 1,655 women in the study group indicated that 28% had had a mammogram before the study began and 13% in the prior two years. In a similar survey of 1,641 women from the control group, 51% reported a previous mammogram and 19% in the previous two years. At the end of the trial, i.e., after approximately seven years of screening in the study group, women in the control group were offered screening mammography, and 66% participated. There is some doubt about the exact date of this screening of controls however: it may not have been December 1988 [Bjurstam et al., 1997] but rather November 1987 [Nyström et al., 2002] or even earlier, at three and six years after randomization [Nyström et al., 1993] or between four and five years after randomization [Nyström et al., 1995]. It is thus not entirely clear whether the duration of the screening contrast was 5, 6, or 7 years. The most recent review [Nyström et al., 2002] gives 6.7 years as the median trial time.

Using the rating scale for strength of contrast, we gave this trial a score of 37%.

Validity

Women were randomized by cluster (18%) and, later, individually (82%), based on date of birth, with a ratio of women randomized to screening and control groups which was 1:1.2 in the 39–49 year age group and 1:1.6 in the 50–59 age group. Exact randomization ratios were in fact different even within these groups [Olsen and Gøtzsche, 2001]. According to a personal communication cited by Olsen and Gøtzsche [2001], a similar proportion of women (1.2%) were excluded from each group based on a previous diagnosis of cancer, although this exclusion was made subsequent to the trial procedure [Bjurstam et al., 1997]. Baseline characteristics of the study and control groups are not available except for age, and even for age, the variable randomization ratios applied to different age groups makes it impossible to use average age as a check for the success of randomization. Other baseline characteristics relevant to risk were not published. Screening and control groups were followed up for cases of in situ carcinoma, including DCIS but excluding LCIS. Cases diagnosed from randomization up to immediately after the round of screening offered to the control group were followed up for outcome, which was mortality from breast cancer identified through the Swedish cause-of-death register. The autopsy rate was 31% [Olsen and Gøtzsche, 2001].

Using our rating scale for validity, we gave this trial a score of 1.5 out of 4.

9. UK AGE TRIAL

In 1991, the United Kingdom Coordinating Committee on Cancer Research set up a national multicentre randomized controlled trial that they refer to as the “Age trial” [Moss, 1999]. This trial has been described in one article describing the nature of the trial [Moss, 1999] and in an article providing additional details and results from a follow-up of the first ten years [Moss et al., 2005a]. There are still too few breast cancer mortality results to conclude that mortality is reduced, but estimates based on interim results such as cancer detection and size and grade of tumours detected made it possible to project mortality rates among the 1,287 women who received a diagnosis of breast cancer up to December 31, 1999 [Moss et al., 2005b].

Contrast

The screening intervention consists of nine rounds of mammography screens starting at age 40–41 at annual intervals until the age of 48. Two-view mammography is used for the first screen, “with single view thereafter unless otherwise indicated” [Moss, 1999]. No CBE nor training nor encouragement in breast self-examination is offered. The control intervention has not been described, but presumably involves no invitation to screening and thus “usual care”. Participation of 70% was expected in the study group, but after five rounds, the average rate observed was 61%. The rate of screening among women in the control cohort was 3.9%, but this figure overestimates the contamination rate, since it represents the proportion of women with at least one mammogram in the preceding three years, whereas the frequency of mammography among women in the screening group was annual. Women from both screening and control groups will be routinely invited to the existing three-yearly mammography screening program once they reach 50. The total duration of the screening contrast should thus be 11 years, making it the trial with the longest contrast by far and, consequently, the one least vulnerable to timing dilution, after a sufficiently long period of follow-up.

Using the rating scale for strength of contrast, we gave this study a preliminary score of 41% at 13-year follow-up.

Validity

Women are randomized into two groups: a study group of 65,000 women and a control group of 130,000 with no intervention. Individual randomization is carried out, stratified by GP practice. Details regarding exclusion of women with previously diagnosed breast cancer and baseline characteristics of the two groups have not been published. Follow-up for breast cancer incidence and mortality will be based on reports from trial centres, pathology laboratories and cancer registries, with a pathology review undertaken for all breast cancers identified in the trial.

Using our rating scale for validity, and making some conjectures which will need to be confirmed when more details are published, we gave this trial a potential score of 4 out of 4 based on its published design and first data on the conduct and analysis of the trial.

RÉFÉRENCES

Agence Nationale d'Accréditation et d'Évaluation en Santé (ANAES). Dépistage du cancer du sein par mammographie : évaluation de la méta-analyse de Gøtzsche et Olsen. Paris: ANAES; 2002.

Alexander FE. The Edinburgh randomised trial of breast cancer screening. J Natl Cancer Inst Monogr 1997;(22):31–5.

Alexander FE, Anderson TJ, Brown HK, Forrest AP, Hepburn W, Kirkpatrick AE, et al. A. 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. Lancet 1999;353(9168):1903–8.

Alexander FE, Anderson TJ, Brown HK, Forrest AP, Hepburn W, Kirkpatrick AE, et al. The Edinburgh randomised trial of breast cancer screening: Results after 10 years of follow-up. Br J Cancer 1994;70(3):542–8.

American Cancer Society (ACS). Cancer facts and figures. Atlanta, GA: ACS, 2004.

Andersson I. Radiographic screening for breast carcinoma. I. Program and primary findings in 45-69 year old women. Acta Radiol Diagn (Stockh) 1981;22(2):185–94.

Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmo Mammographic Screening Trial. BMJ 1988;297(6654):943–8.

Andersson I, Hellstrom L, Bjurstam N, Lundgren B, Fagerberg G, Tabár L. [Breast cancer screening by mammography in Sweden]. Lakartidningen 1983;80(25):2559–62 (article in Swedish).

Andersson I, Andren L, Hildell J, Linell F, Ljungqvist U, Pettersson H. Breast cancer screening with mammography: A population-based, randomized trial with mammography as the only screening mode. Radiology 1979;132(2):273–6.

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

Bailar JC and MacMahon B. Randomization in the Canadian National Breast Screening Study: A review for evidence of subversion. CMAJ 1997;156(2):193–9.

Baines CJ. A different view on what is known about breast screening and the Canadian National Breast Screening Study. Cancer 1994;74 (4):1207–11.

Baines CJ, Miller AB, Kopans DB, Moskowitz M, Sanders DE, Sickles EA, et al. Canadian National Breast Screening Study: Assessment of technical quality by external review. Am J Roentgenol 1990;155(4):743–9.

Baines CJ, McFarlane DV, Wall C. Audit procedures in the National Breast Screening Study: Mammography interpretation. Can Assoc Radiol J 1986;37(4):256–60.

Baker SG, Kramer BS, Prorok PC. Statistical issues in randomized trials of cancer screening. BMC Med Res Methodol 2002;2(1):11.

Baum M. Commentary: False premises, false promises and false positives—The case against mammographic screening for breast cancer. Int J Epidemiol 2004;33(1):66–73.

Bjurstam N, Bjorneld L, Warwick J, Sala E, Duffy SW, Nystrom L, et al. The Gothenburg Breast Screening Trial. Cancer 2003;97(10):2387–96.

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(11):2091–9.

Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. J Natl Cancer Inst 2002;94(3):167–73.

Brown BW, Brauner C, Minnotte MC. Noncancer deaths in white adult cancer patients. J Natl Cancer Inst 1993;85(12):979–87.

Chalmers I. Comparing like with like: Some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. Int J Epidemiol 2001;30(5):1156–64.

Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. Control Clin Trials 1981;2(1):31–49.

Conseil d'évaluation des technologies de la santé du Québec (CETS). Screening for breast cancer in women aged 40-49 years. Montréal, Qc: CETS; 1993.

Conseil d'évaluation des technologies de la santé du Québec (CETS). Screening for breast cancer in Quebec: Estimates of health effects and of cost. Montreal, Qc: CETS; 1990.

DerSimonian R and Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7(3):177–88.

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(16):1089–96.

Fagerberg G and Tabár L. The results of periodic one-view mammography screening in a randomized, controlled trial in Sweden. Part 1: Background, organization, screening program, tumor findings. In: Day NE, Miller AB. Screening for breast cancer. Toronto, Ont.: Hans Huber Publishers; 1988: 33–8.

Fagerberg G, Baldetorp L, Grontoft O, Lundstrom B, Manson JC, Nordenskjold B. Effects of repeated mammographic screening on breast cancer stage distribution. Results from a randomised study of 92 934 women in a Swedish county. Acta Radiol Oncol 1985;24(6):465–73.

Fink R, Shapiro S, Roester R. Impact of efforts to increase participation in repetitive screenings for early breast cancer detection. Am J Public Health 1972;62(3):328–36.

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(20):1644–56.

Freedman DA, Petitti DB, Robins JM. On the efficacy of screening for breast cancer. Int J Epidemiol 2004;33(1):43–55.

Frisell J, Lidbrink E, Hellstrom L, Rutqvist LE. Followup after 11 years—Update of mortality results in the Stockholm Mammographic Screening Trial. Breast Cancer Res Treat 1997;45(3):263–70.

Frisell J, Eklund G, Hellstrom L, Lidbrink E, Rutqvist LE, Somell A. Randomized study of mammography screening—Preliminary report on mortality in the Stockholm trial. Breast Cancer Res Treat 1991;18(1):49–56.

Frisell J, Eklund G, Hellstrom L, Glas U, Somell A. The Stockholm breast cancer screening trial—5-year results and stage at discovery. Breast Cancer Res Treat 1989;13(1):79–87.

Frisell J, Glas U, Hellstrom L, Somell A. Randomized mammographic screening for breast cancer in Stockholm. Design, first round results and comparisons. Breast Cancer Res Treat 1986;8(1):45–54.

Frisell J and Lidbrink E. The Stockholm mammographic screening trial: Risks and benefits in age group 40-49 years. J Natl Cancer Inst Monogr 1997;(22):49–51.

GE Medical Systems. Mammography: History of the mammography, 2004. Available at: http:// www.gemedicalsystems.com/rad/whc/mswhhis.html (accessed on July 28,’2004).

Glasziou PP. Meta-analysis adjusting for compliance: The example of screening for breast cancer. J Clin Epidemiol 1992;45:1251–6.

Gøtzsche PC and Olsen O. Is screening for breast cancer with mammography justifiable? Lancet 2000;355(9198):129–34.

Health Canada. Organized breast cancer screening programs in Canada: 1999 and 2000 report. Ottawa, Ont.: Minister of Public Works and Government Services Canada, 2003.

Health Council of the Netherlands. The benefit of population screening for breast cancer with mammography. La Haye: HCN; 2002.

Hendrick RE, Smith RA, Rutledge JH, Smart CR. Benefit of screening mammography in women aged 40-49: A new meta-analysis of randomized controlled trials. J Natl Cancer Inst Monogr 1997;(22):87–92.

Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2002;137 (5 Pt 1):347–60.

Institut national de santé publique du Québec (INSPQ), Brisson J, Hébert-Croteau N, Langlois A. Déterminants du taux de référence lors d'une première mammographie de dépistage – Programme québécois de dépistage du cancer du sein 1999. Québec: INSPQ; 2003.

International Agency for Research on Cancer (IARC). IARC handbooks of cancer prevention. Vol. 7: Breast cancer screening. Lyon, France: IARC Press; 2002.

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17(1):1–12.

Jung H. [Assessment of usefulness and risk of mammography screening with exclusive attention to radiation risk]. Radiologe 2001;41(4):385–95 (article in German).

Kakuma R. Screening for breast cancer with mammography among women aged 50-69 years. Report prepared for the Breast Cancer Screening Unit. Ottawa, Ont.: Health Canada, 2002.

Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273(2):149–54.

Lampic C, Thurfjell E, Bergh J, Sjoden PO. Short- and long-term anxiety and depression in women recalled after breast cancer screening. Eur J Cancer 2001;37(4):463–9.

Larsson LG, Andersson I, Bjurstam N, Fagerberg G, Frisell J, Tabár L, Nystrom L. 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.

Larsson LG, Nystrom L, Wall S, Rutqvist L, Andersson I, Bjurstam N, Fagerberg G, Frisell J, Tabár L. The Swedish randomised mammography screening trials: Analysis of their effect on the breast cancer related excess mortality. J Med Screen 1996;3(3):129–32.

Last JM. A dictionary of epidemiology. 4th edition. New York, NY: Oxford University Press; 2001.

Law M, Hackshaw A, Wald N. Screening mammography re-evaluated. Lancet 2000;355(9205): 749–52.

Madlensky L, Goel V, Polzer J, Ashbury FD. Assessing the evidence for organised cancer screening programmes. Eur J Cancer 2003;39(12):1648–53.

McTiernan A. Recent controversies in mammography screening for breast cancer. Medscape Womens Health 2002;7(2):3.

Medical Research Council (MRC). Streptomycin treatment of pulmonary tuberculosis. BMJ 1948;769–82.

Miettinen OS. The need for randomization in the study of intended effects. Stat Med 1983;2(2): 267–71.

Miettinen OS, Henschke CI, Pasmantier MW, Smith JP, Libby DM, Yankelevitz DF. Mammographic screening: No reliable supporting evidence? Lancet 2002;359(9304): 404–6.

Miller AB. Screening for breast cancer with mammography. Lancet 2001;358(9299):2164, 2167–8.

Miller AB. Screening for breast cancer: A review. Eur J Cancer Clin Oncol 1988;24(1):49–53.

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(5 Part 1):305–12.

Miller AB, Baines CJ, To T, Wall C. Screening mammography re-evaluated. Lancet 2000a;355 (9205):747–52.

Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. J Natl Cancer Inst 2000b;92 (18):1490–9.

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

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. CMAJ 1992a;147(10): 1459–76.

Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. CMAJ 1992b;147(10): 1477–88.

Miller AB, Baines CJ, Turnbull C. The role of the nurse-examiner in the National Breast Screening Study. Can J Public Health 1991;82(3):162–7.

Miller AB, Howe GR, Wall C. The national study of breast cancer screening protocol for a Canadian randomized controlled trial of screening for breast cancer in women. Clin Invest Med 1981;4(3-4):227–58.

Ministère de la Santé et des Services sociaux (MSSS). Bilan 1998-2003 — Programme québécois de dépistage du cancer du sein. Québec: MSSS; 2004.

Ministère de la Santé et des Services sociaux (MSSS). Programme québécois de dépistage du cancer du sein – Rapport d'activité 2000-2001. Québec: MSSS; 2003.

Ministère de la Santé et des Services sociaux (MSSS). Programme québécois de dépistage du cancer du sein – Cadre de référence. Québec: MSSS, 1996.

Mocharnuk RS. Screening mammography: The controversy continues. 25th Annual San Antonio Breast Cancer Symposium; 2002. Available at: www.medscape.com/viewprogram/2193_pnt.

Morrison BJ. Screening for breast cancer. In: Canadian Task Force on the Periodic Health Examination. Canadian Guide to Clinical Preventive Health Care. Ottawa: Health Canada; 1994:892-900.

Morrone D, Giorgi D, Ciatto S, Frigerio A, Catarzi S, Rosselli Del Turco M. [Assessment of diagnostic accuracy of mammography carried out for secondary prevention. Results of a test with a sample caseload conducted by 75 Italian radiologists]. Radiol Med (Torino) 2001;101(1-2):44–7 (article in Italian).

Moss S. A trial to study the effect on breast cancer mortality of annual mammographic screening in women starting at age 40. Trial Steering Group. J Med Screen 1999;6(3):144–8.

Moss S, Thomas I, Evans A, Thomas B, Johns L, for the Trial Management Group. Randomised controlled trial of mammographic screening in women from age 40: Results of screening in the first 10 years. Br J Cancer 2005a;92:949–54.

Moss S, Waller M, Anderson TJ, Cuckle H, for the Trial Management Group. Randomised controlled trial of mammographic screening in women from age 40: Predicted mortality based on surrogate outcome measures. Br J Cancer 2005b;92:955–60.

Nystrom L, Andersson I, Bjurstam N, Frisell J, Nordenskjold B, Rutqvist LE. Long-term effects of mammography screening: Updated overview of the Swedish randomised trials. Lancet 2002;359(9310):909–19.

Nystrom L, Larsson LG, Wall S, Rutqvist LE, Andersson I, Bjurstam N, 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(2):85–7.

Nystrom L, Larsson LG, Rutqvist LE, Lindgren A, Lindqvist M, Ryden S, et al. Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials. A comparison between official statistics and validation by an endpoint committee. Acta Oncol 1995;34(2):145–52.

Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: Overview of Swedish randomised trials. Lancet 1993;341 (8851): 973–8.

Olsen O and Gøtzsche PC. Screening for breast cancer with mammography. Cochrane Database Syst Rev 2001;(4):CD001877.

Paci E and Alexander FE. Study design of randomized controlled clinical trials of breast cancer screening. J Natl Cancer Inst Monogr 1997;(22):21–5.

Ringash J. Preventive health care, 2001 update: Screening mammography among women aged 40-49 years at average risk of breast cancer. CMAJ 2001;164(4):469–76.

Roberts MM, Alexander FE, Anderson TJ, Chetty U, Donnan PT, Forrest P, et al. Edinburgh trial of screening for breast cancer: Mortality at seven years. Lancet 1990;335(8684):241–6.

Roberts MM, Alexander FE, Anderson TJ, Forrest AP, Hepburn W, Huggins A, et al. The Edinburgh randomised trial of screening for breast cancer: Description of method. Br J Cancer 1984;50(1):1–6.

Roberts RA and Birch NJ. A comparison of breast cancer secondary prevention activities and satisfaction with access and communication issues in women 50 and over. Prev Med 2001;32(4):348–58.

Shapiro S. Periodic screening for breast cancer: The HIP randomized controlled trial. Health Insurance Plan. J Natl Cancer Inst Monogr 1997;(22):27–30.

Shapiro S. Screening: Assessment of current studies. Cancer 1994;74(Suppl 1):231–8.

Shapiro S. The status of breast cancer screening: A quarter of a century of research. World J Surg 1989;13(1):9–18.

Shapiro S. Evidence on screening for breast cancer from a randomized trial. Cancer 1977;39 (Suppl 6):2772–82.

Shapiro S, Venet W, Strax P, Venet L. Periodic screening for breast cancer: The Health Insurance Plan project and its sequelae, 1963–1986. Baltimore, MD: Johns Hopkins University Press; 1988.

Shapiro S, Venet W, Strax P, Venet L, Roeser R. Selection, follow-up, and analysis in the Health Insurance Plan study: a randomized trial with breast cancer screening. Natl Cancer Inst Monogr 1985;67:65–74.

Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. J Natl Cancer Inst 1982;69(2):349–55.

Shapiro S, Strax P, Venet L. Evaluation of periodic breast cancer screening with mammography. Methodology and early observations. JAMA 1966;195(9):731–8.

Sickles EA. Screening outcomes: Clinical experience with service screening using modern mammography. In: National Institute of Health. NIH Consensus Development Conference, breast cancer screening for women ages 40-49, program and abstracts. Bethesda, MD: NIH; 1997: 105–10.

Strax P, Venet L, Shapiro S. Value of mammography in reduction of mortality from breast cancer in mass screening. Am J Roentgenol Radium Ther Nucl Med 1973;117(3):686–9.

Tabár L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality. Cancer 2001;91(9):1724–31.

Tabár L, Vitak B, Chen HH, Duffy SW, Yen MF, Chiang CF, 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(4):625–51.

Tabár L, Vitak B, Chen HH, Prevost TC, Duffy SW. Update of the Swedish Two-county trial of breast cancer screening: Histologic grade-specific and age-specific results. Swiss Surg 1999;5(5):199–204.

Tabár 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. J Natl Cancer Inst Monogr 1997;(22):43–7.

Tabár 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 1995a;2(2): 94–8.

Tabár L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, Smith RA. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995b;75(10):2507–17.

Tabár L, Fagerberg G, Day NE, Duffy SW. The Swedish two-county trial of mammographic screening for breast cancer: Recent results on mortality and tumor characteristics. Pathol Biol (Paris) 1992a;39(9):846.

Tabár L, Fagerberg G, Day NE, Duffy SW, Kitchin RM. Natural history of breast cancer. Lancet 1992b;339(8801):1108.

Tabár L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992c;30(1):187–210.

Tabár L, Fagerberg G, Duffy SW, Day NE. The Swedish two county trial of mammographic screening for breast cancer: Recent results and calculation of benefit. J Epidemiol Community Health 1989;43(2):107–14.

Tabár L, Fagerberg CJ, Gad A, Baldetorp L, Holmberg LH, Grontoft O, 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 1985a;1(8433):829–32.

Tabár L, Gad A, Holmberg L, Ljungquist U. Significant reduction in advanced breast cancer. Results of the first seven years of mammography screening in Kopparberg, Sweden. Diagn Imaging Clin Med 1985b;54(3-4):158–64.

Tabár L and Gad A. Screening for breast cancer: The Swedish trial. Radiology 1981;138(1):219–22.

UK Trial of Early Detection of Breast Cancer Group. First results on mortality reduction in the UK Trial of Early Detection of Breast Cancer. Lancet 1988;2(8608):411–6.

UK Trial of Early Detection of Breast Cancer Group. Trial of Early Detection of Breast Cancer: Description of method. Br J Cancer 1981;44(5):618–27.

Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. Am J Roentgenol 2001;177(3): 543–9.

Zahl PH, Kopjar B, Maehlen J. [Norwegian breast cancer mortality rates and validity of Swedish mammographic studies]. Tidsskr Nor Laegeforen 2001;121(16):1928–31 (article in Norwegian).

Agence d'évaluation
des technologies
et des modes
d'intervention en santé

Québec