Measuring the Mortality Impact of Breast Cancer Screening

James A. Hanley, PhD,[1,2] Maurice McGregor, MD,[2] Zhihui Liu, MSc,[1] Erin C. Strumpf, PhD,[1,3] Nandini Dendukuri, PhD,[1,2]

ABSTRACT

OBJECTIVE: To i) estimate how large the mortality reductions would be if women were offered screening from age 50 until age 69; ii) to do so using the same trials and participation rates considered by the Canadian Task Force; iii) but to be guided in our analyses by the critical differences between cancer screening and therapeutics, by the time-pattern that characterizes the mortality reductions produced by a limited number of screens, and by the year-by-year mortality data in the appropriate segment of follow-up within each trial; and thereby iv) to avoid the serious underestimates that stem from including inappropriate segments of follow-up, i.e., too soon after study entry and too late after discontinuation of screening.

METHODS: We focused on yearly mortality rate ratios in the follow-up years where, based on the screening regimen employed, mortality deficits would be expected. Because the regimens differed from trial to trial, we did not aggregate the yearly data across trials. To avoid statistical extremes arising from the small numbers of yearly deaths in each trial, we calculated rate ratios for 3-year moving windows.

RESULTS: We were able to extract year-specific data from the reports of five of the trials. The data are limited for the most part by the few rounds of screening. Nevertheless, they suggest that screening from age 50 until age 69 would, at each age from 55 to 74, result in breast cancer mortality reductions much larger than the estimate of 21% that the Canadian Task Force report is based on.

DISCUSSION: By ignoring key features of cancer screening, several of the contemporary analyses have seriously underestimated the impact to be expected from such a program of breast cancer screening.

KEY WORDS: Cancer screening; early diagnosis; randomized trials; mortality

La traduction du résumé se trouve à la fin de l'article.

Can J Public Health 2013;104(7):e437-e442.

Whether or not to implement a screening program for breast cancer requires weighing the health benefits (cancer deaths averted) against the harms (overdiagnosis) and the costs. Essential to such a decision is an accurate estimation of the *extent* of the health benefits and harms in question. We avoid the larger debate, to screen or not to screen, and focus instead on how the benefit is typically calculated in reports. We show that this method contains conceptual errors and leads to serious underestimates. Although other reports[1,2] are also based on analyses that contain these same errors, and use the same trials, we will for simplicity focus on the recent report of the Canadian Task Force on screening for breast cancer in average-risk women aged 50-69 years.[3] Before we address this report, we first briefly consider some important characteristics of screening for cancers.

Unlike most medical interventions (that produce rapid effects), cancer screening, by its very nature, generates mortality reductions that only manifest *several years after the onset of screening*.[4,5] Illustrated in Figure 1 are hypothetical examples of the yearly percentage mortality reductions that might be expected from screening for cancer every year for a) just three years (as some trials did) or b) twenty years (as a screening program might do). Screening leads to earlier treatment of otherwise fatal cancers, but can only save lives (produce a mortality "deficit" or "reduction") at the time when the deaths averted as a result of screening would have (otherwise) occurred. Thus, in the *trial*, illustrated in scenario a), the mortality of the screened population, relative to that of the unscreened, only starts to fall perceptibly by the third year, when the earliest effect of the first screen is expressed; it continues to fall for three more years, with the greatest reduction (35%) attained in the sixth year; mortality then rises again and returns to the level in the unscreened population after year nine when the last effect of the third and final screen is expressed. In contrast, in the 20-year screening *program*, illustrated in scenario b), the (relative) mortality in the screened population would again start to decrease by the third year, and the reductions would reach an *asymptote* (largest possible magnitude of benefit) of 46% in the seventh year; mortality would only rise again and return to that in the unscreened population after year twenty-six, the year when the last effect of the twentieth screen in the program is expressed.

Thus, when our objective is to deduce the size of the reduction in breast cancer mortality that would result from instituting a program of regular screening, we must identify the "asymptote": the *annual mortality reduction* that would be achieved each year after an *adequate period* of regular screening. One could not determine this

Author Affiliations

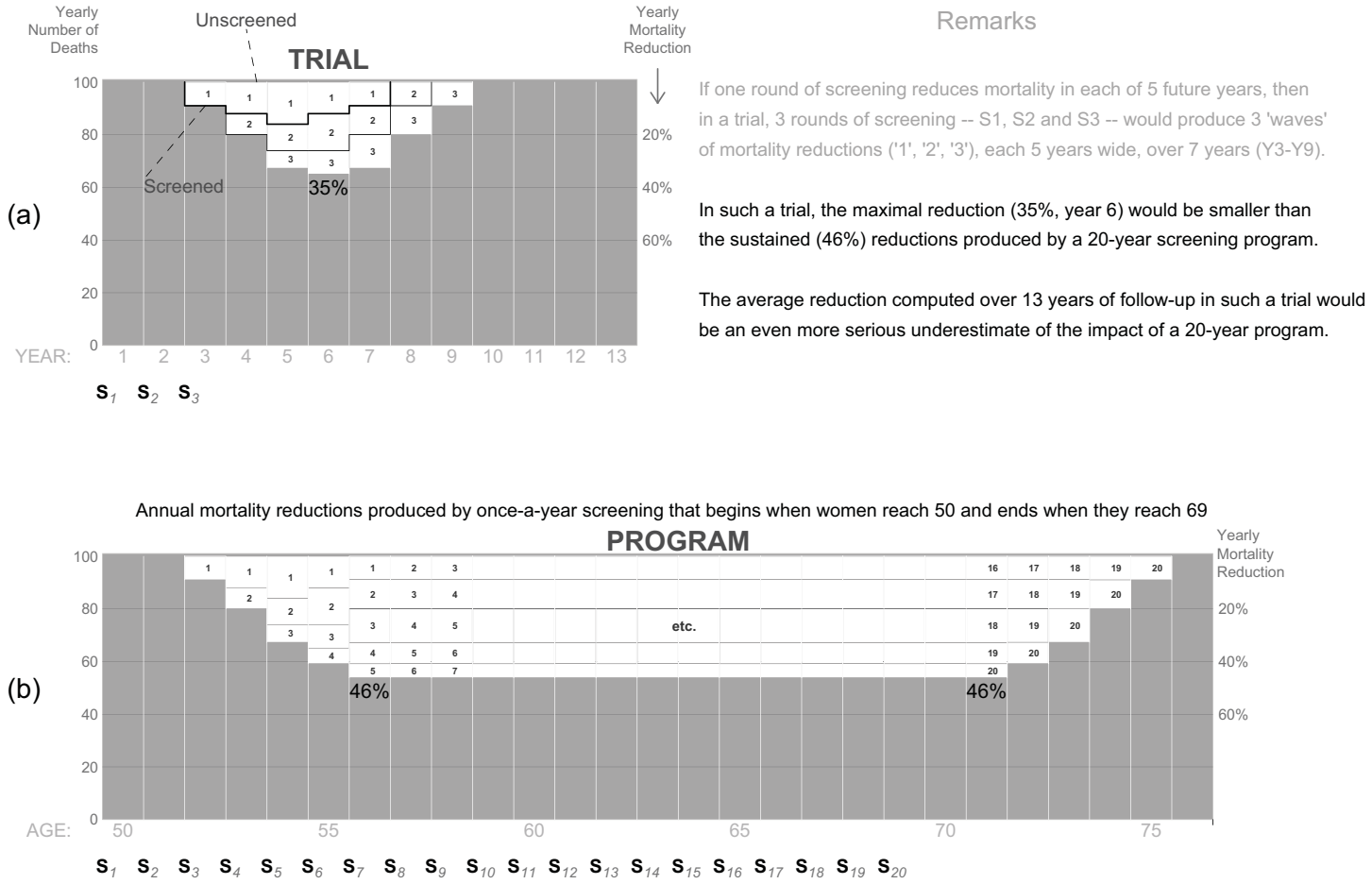
McGill University, Montreal, QC
1. Department of Epidemiology, Biostatistics and Occupational Health
2. Department of Medicine
3. Department of Economics

Correspondence: James Hanley, Dept. of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, QC H3A 1A2, E-mail: james.hanley@mcgill.ca

Acknowledgements: This work was supported by the Canadian Institutes of Health Research.

Conflict of Interest: None to declare.

Figure 1. The difference between a screening trial and a screening program



a) A hypothetical *trial* of 3 annual rounds of cancer screening (S_1, S_2, S_3) compared with no screening. The depth of the white rectangle in each year represents the percentage mortality reduction, relative to an unscreened group, for the year shown on the horizontal axis. Annual mortality reductions (100 x [mortality rate if no screening - mortality rate if screening]/mortality rate if no screening) produced by screening have an expected delay and only begin to be expressed in year three (when the first effect of S_1 is discernable); they are greater in years 4 and 5, reaching a maximum of 35% in year 6 (when the combined effect of S_1, S_2 and S_3, denoted by '1' , '2' and '3' respectively, is maximal); in year 7 the combined effects begin to wear off, and the mortality in the screening arm begins to revert to that in the non-screening arm; in year 9, the last effect of S_3 is discernable. Thus the maximum reduction is 35% and it would have been greater if screening had not been discontinued at year three. By contrast, the *average* effect of screening over the 13 years of observation (the metric used by the Canadian Task Force) would be 12%.

b) A hypothetical screening *program* with annual screening beginning at age 50 and continuing until age 69, compared with no screening. Again, the depth of the white rectangle represents the percentage mortality reduction for the age shown on the horizontal axis. The mortality reduction reaches 46% at age 56 and is maintained at that level for many age bins – until three years after the last screen when it starts to increase again. The calculated *average* reduction in mortality from age 50 to 76 would in this instance be 35%.

value if screening were discontinued "prematurely" (i.e., before the maximum annual mortality reduction of 46% was achieved), and any estimate from a trial with a limited number of rounds of screening will be an underestimate of what the program could achieve.

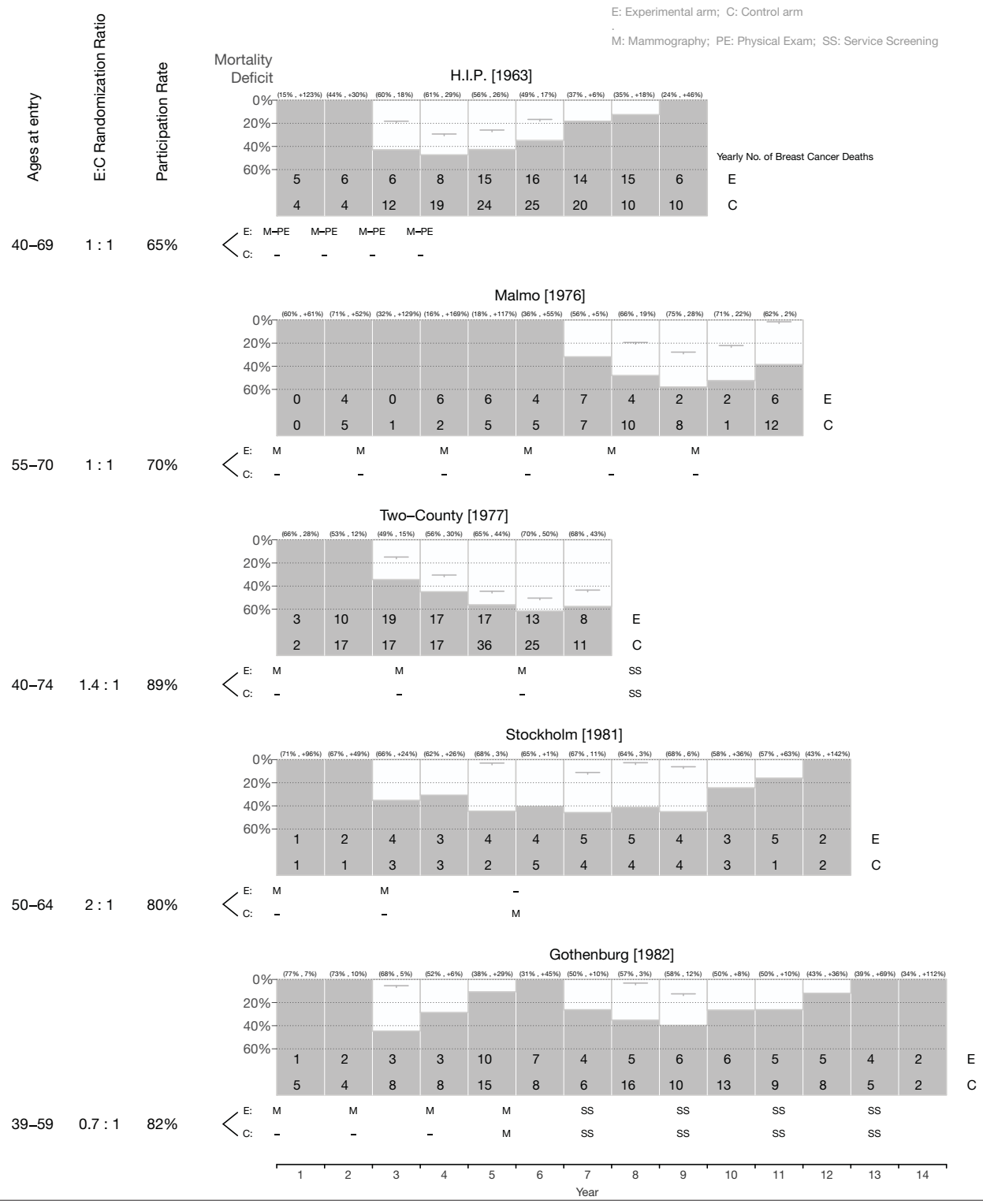
However, many of the screening trials on which the Canadian and other reports are based were terminated prematurely (either by ending screening of the intervention population, or by initiating screening of the control population). Furthermore, most of these studies do not report the mortality deficits observed in each year of the trial, but give their results as a single rate ratio, and thus a single mortality deficit, calculated from the cumulative numbers of deaths. This metric includes all deaths from the very onset of screening to the end of the follow-up, however long or short, or arbitrary, that duration may be. This overall duration includes the early years in which little or no reduction in mortality can be expected, and sometimes also the late years in which the effects of screening are diminishing as a result of its discontinuation. By relying on this overall measure, task forces inevitably arrive at results that are smaller than the reduction achievable by a program (46% in our hypothetical example) by an amount dependent on the number of years included in the average in which mortality reduction was zero or less than maximal.

Although these features of screening have long been recognized,[4-13] they are still frequently overlooked, as they were in the recent report of the Canadian Task Force on Preventive Health Care. Its guidelines are primarily based on a meta-analysis of six breast cancer screening trials,[14-19] which found that the expected mortality reduction that would result from breast screening was 21%. Our primary objective in this paper is to display the yearly mortality data in each trial and deduce the reduction expected from a screening program, using an approach that respects the features referred to above.

METHODS

Five of the six trials subjected to meta-analysis by the Task Force are briefly summarized below. It was necessary to exclude the

Figure 2. The number of rounds of screening, and the (approximate) timing of the beginning of each round, in each of the trials of breast cancer screening yearly, together with the yearly numbers of deaths in the experimental and control groups, and the year-specific mortality deficits



The year each study began is shown in square brackets.
In order to reduce the statistical noise, each yearly mortality deficit was calculated from the mortality rate ratio based on the year in question and the year before and after it, i.e., using a 3-year window. For example, the rate ratio shown in year 9 of the Malmö trial is (4+2+2)/(10+8+1) = 0.42, so the calculated deficit shown is 1 – 0.42 = 58% (limits of the 80% CI are a 75% deficit and a 28% deficit). In the same year in the Gothenburg trial, where the allocation ratio was 0.7:1, the rate ratio is ([5+6+6]/0.7)(16+10+13) = 0.62, the mortality deficit is 1 – 0.62 = 38% (limits 58% and 12%). The smoothing provides more reliable and more realistic estimates of the nadir (asymptote) one would expect with a sustained screening program, the 'estimand' of interest. For year 3 onwards, i.e., where reductions might be expected, they are represented by the depths of the unshaded rectangles if the point estimate is on the expected side of the null. In the 80% confidence limits shown for each year, a '+' indicates a mortality excess rather than a mortality deficit. The totals of the year-specific rates shown in the yearly columns do not necessarily match the overall numbers of breast cancer deaths in the Task Force meta-analyses, since it was not always possible for us to obtain follow-up-year-specific counts for the age-span of interest, and the totals in some of the trial reports include irrelevant follow-up years well after the influence of the last screen, or after the screening content of the two arms became similar.

Canadian Trial[19] (1980b in the Canadian meta-analysis[20]) because the year-specific mortality data are not available from the reports nor obtainable from the authors. The remaining five trials differ so greatly in the screening regimens and other important elements that we do not find it justifiable to combine the year-specific numbers of deaths. Instead, we examined the year-by-year pattern of mortality deficits in each trial separately. Thus, for each trial, we attempted to identify the "trough" or "nadir" achieved following the onset of screening.

Two authors (JH, ZL) independently extracted the year-specific numbers of breast cancer deaths in the experimental and control arms from the published articles. From the cumulative numbers of deaths reported in Table 7 in the HIP trial and Table X in the Malmö trial, we calculated the yearly numbers of deaths by successive subtractions. The reports of the other three trials contained plots of cumulative numbers of deaths over time (Figure 2, Two-County; Figure 2, Stockholm; Figure 1, Gothenburg). For each of these, we used a graph digitizer to extract the cumulative values, and then converted them into year-specific numbers of deaths, and checked the totals against the total numbers reported in the text. Disagreements between extractors were resolved by further review. In reports that did not provide sufficiently age-specific data, we used slightly wider or narrower age-at-entry bands.

There was substantial variation in the screening regimens, and the year-specific death counts in most trials were in the single digits. To reduce the statistical noise, and to avoid artifacts in estimating nadirs, we used three-year moving averages to calculate the year-specific mortality rate ratios, and their complements, the year-specific mortality deficits. Given the general lack of sufficiently sustained screening in these trials, our aim was to use the maximum annual mortality deficit in each trial to gain some idea of the sustained mortality reduction that would result if women were regularly screened (annually or biennially), from age 50 until 69, at the same participation rates as pertained in the trials.

We investigated, by simulations, whether this amount of smoothing (each deficit based on three-year moving rates) was sufficient to keep the probability of overestimating the true nadir at around 50% (i.e., whether the estimator of the nadir was median unbiased). We found that indeed, if one relied on the largest deficit in a series of moving deficits, one would tend to slightly overestimate the true nadir. But we also found that the most conservative of three adjacent such moving deficits was as likely to overestimate the true nadir as it was to underestimate it. When visually extracting a sensible nadir from Figure 2, we informally looked for an estimate of the percentage deficit that would be surpassed or equaled by the displayed moving deficit for at least three successive years. For example, the HIP study has three consecutive years with deficits of more than 40%, while the Malmö study has three with deficits of more than 45%.

RESULTS

The five trials in question are included in Figure 2 and are summarized below.

The **HIP trial**[14] employed 4 annual rounds of screening, using mammography and physical examination, with a participation rate of 65% at the initial round. The breast cancer mortality deficits begin to manifest in year 3, reaching values of 43%, 47% and 43% for the next three years, after which the effect of screening (already discontinued) again diminishes. Thus, screening is associated with a sustained deficit in annual mortality of over 40%.

Comment: The Task Force meta-analysis[20] used a 22% deficit, calculated over 14 years, including the first 2 years in which the effect of screening had not yet commenced, and the years 10-14 in which its effects had ended. Thus it clearly underestimates what a sustained program could achieve.

The **Malmö trial**[15] had the longest duration of screening: 6 rounds over 9 years, with a participation rate of more than 70%. The task force used the data for women aged 55 years and over. Probably because of its limited size (virtually all of the yearly numbers of deaths are in the single digits), breast cancer mortality deficits only begin to be expressed in year 7, reaching values of 48%, 58%, and 52% in years 8, 9 and 10, respectively, when the trial was terminated. Thus the sustained deficit in annual mortality was of the order of 50%.

Comment: The deficit in mortality used by the Task Force is an average over 18 years. Since in years 12-18 (yearly data not available), women in the control arm were invited to screening, the 18% deficit calculated by the task force would be expected to underestimate the uncontaminated impact of 6 rounds of screening. Indeed, the authors of this study recognized that "intervention at the non-invasive or early invasive stage would not influence the death rate until several years later". They estimated that after a 6-year delay and with the inclusion of preliminary data from 1987, the deficit in mortality is 42%.[14]

In the **Two-County trial,**[16] the experimental arm involved 3 rounds of screening over a span of 5 years. Women in the control arm were invited to screening from about year 8 onwards. The mortality deficits in the last three years (56%, 62%, and 58%, with an average of 59%) reflect the deficits in mortality resulting from screening in this study.

Comment: The substantial mortality deficit in this trial presumably reflects both the high participation rate (89% at the initial examination) in the experimental arm and the greater stability of the derived statistics: this trial was the largest of the five in terms of yearly numbers of deaths. Based on the average mortality over the lengths of the follow-up in the 1995 and 2002 separate-county (East and West) reports, the Task Force analysis used deficits of 19% and 47%, respectively, or 33% if one were to combine them.

The **Stockholm trial**[17] involved 2 rounds of screening over a span of 2 years. Women in the control arm were invited to screening after about year 5, thus limiting the time during which the uncontaminated effect of screening could be observed. In years 5, 6 and 7, deficits of 45%, 40% and 46%, respectively (average 44%) were observed. Over years 3-9, there is a sustained mortality deficit of approximately 40%.

Comment: In contrast, the Task Force calculated an average deficit over all 12 years of 32%.

In the **Gothenburg trial,**[18] the experimental arm involved 4 rounds of screening over a span of 6 years. Women in the control arm were invited to screening as soon as the cumulative number of breast cancer deaths in the experimental arm was statistically significantly lower than that in the control arm (thereby preventing the full expression of the effect of screening). The 3 rounds of screening appear to have resulted in mortality deficits of 45% and 29% in the two years before the trial was effectively terminated by introducing screening to the control group. Thereafter the time-

pattern of the mortality deficits becomes erratic. A very approximate estimate of the effect of screening would be the average of the two years in which it was observed, i.e., 38%.

Comment: Not surprisingly, given the similarity of the intervention in the two arms from year 5 onwards, there is no evidence of the impact of screening beyond year 13. The 21% average over all 14 years used by the Task Force reflects both this attenuation and the inclusion of the initial years in which no effect could have been seen.

Estimated mortality reduction of a program that screens regularly for a 20-year age span

From observation of the deficits in mortality associated with screening in each trial (Figure 2), it is apparent that (except for the Malmö trial) screening was not maintained sufficiently long to achieve its full effect. However, some idea of the magnitude of the reduction in mortality that would have been achieved if screening were continued for 20 years can be estimated from the pattern of deficits. Despite the variability, expected with such small numbers, the trials consistently suggest that 20 years of offering screening to women from age 50 to 69 would be followed by 20 years (approximately ages 55-74) in which the breast cancer mortality reductions would be at least 40%. Moreover, since the maximal deficits were achieved with participation rates that were well below 100%, they in turn underestimate the probability of benefit for women who would participate more fully than the "average" in the trials.

DISCUSSION

The decision to initiate and/or sustain a program of breast cancer screening will always require up-to-date and accurate estimates of the harms and benefits that it will cause. Since the time when the studies cited above were carried out, screening techniques have become more sensitive (and less specific) and cancer therapies have become more effective. However, if they are to be used for the formulation of policy, they must be correctly interpreted. Without engaging in the debate on the overall value of screening, we believe that the reduction in mortality estimated by the Task Force on the basis of these studies is a considerable underestimation.

What we need to know for such a decision is the yearly reduction in mortality that will result from screening (say annually or biennially) of women of a given age at entry (say 50 years) over a prolonged (say 20 years) time, compared with the mortality in women who do not take part in screening. This we must attempt to derive from data reflecting much shorter periods of screening (usually terminated before the full effect can be seen) of women invited to screening, compared to control groups in which substantial proportions undergo "external" screening. Furthermore, we need to know the reduction in annual mortality rate produced by the screening rather than the reduction over the overall length of the follow-up, a figure that will be unduly low due to inclusion of mortality data at times when the intervention can only have zero or reduced effects. Even without correction for rates of external screening, the deficits shown in Figure 2 indicate that, in contrast with the 21% calculated by the Canadian Task Force, the estimated reduction lies closer to 40%. The mortality reduction in women screened, as distinct from invited, would be greater and would be further increased when compared to women who are not screened.

To appreciate the numbers involved, one might wish to apply these different percentage reductions, and the amount of screening that would be involved, to the current population of Canadian women. At present, approximately 4 million Canadian women are between the ages of 50 and 69. Each year, more or less uniformly distributed over the age range 50 to 85, there are approximately 5,000 breast cancer deaths. If screening from age 50 to 69 resulted in a 20% reduction in the breast cancer mortality rates in the age ranges 55-75, with smaller reductions in younger and older ages, approximately 650 breast cancer deaths would be averted each year; if it resulted in a 40% reduction, 1,300 would be.

We did not attempt to calculate what the reductions would be with other or full participation rates. We merely show that despite participation rates that are well below those seen in therapeutic trials, and despite the fact that the regimens used in the trials were much shorter than those that would be used in a screening program, the deficits achieved were still considerably larger than the reductions estimated in the Task Force report.

An implicit but clearly inappropriate assumption in the meta-analysis underpinning the Task Force report is statistical exchangeability of deaths in different person years, no matter whether they occur in year 1, 11 or 24. Unlike the practice in other "latency" contexts,[21] most data analysts ignore the non-proportional hazards[5,22] that characterize mortality patterns in cancer screening trials. We suggest they adopt a time-specific approach such as that in Figures 1 and 2, and dispense with single (aggregated over all follow-up time) numbers.

Ideally, i.e., if they were sufficiently numerous, the data in each separate trial we examined would coherently "speak for themselves" as to the time windows in which one should and should not expect mortality deficits. However, in many of the trials, and despite our attempts to reduce the noise, the numbers of screenings and the numbers of breast cancer deaths were almost too low to interpret. The Malmö trial is the only one with a sufficiently sustained screening regimen to generate a genuine asymptote. And indeed, when the time-specific data from this trial were reconsidered in detail,[4] and allowance was made for the expected lag, they suggested that large mortality reductions (>50%) are possible with sustained screening.

Likewise, the long-term (25-30 year) follow-up of cancer screening trials with limited screening, and the use of (one-number) reduction measures based on all deaths in the follow-up window, in subjects whose last screening examination was carried out decades earlier,[19,23] will not be informative. In such analyses, the inclusion of the time window before any deficits would be expected will already dilute the effect; but the inclusion of the very long post-last-screen time window – when deficits will long since have disappeared – will dilute it even more,[4,5,22] and make the resulting number meaningless as a measure of what a screening program that involves 20 years of screening would accomplish.

The duration of screening in a trial is typically shorter than that in a program and the deficits last for fewer years. The Canadian Task Force failed to distinguish trials from programs, as is evident in their statement "Screening women aged 50-69 years … for about 11 years" and in their calculations based on this arbitrary time-horizon. If numbers needed to screen are to be meaningful, they should refer to the full length of a program, in which women would undergo 20 years of screening (10-20 examinations say), starting at

age 50, rather than the limited number (typically 3-4) of examinations and an average of 11 years of follow-up in the trials the Task Force used. Likewise, mortality deficits should be tallied in a 30-year follow-up window extending from 50 to 80 years of age.

Finally, it should be noted that the full effect of an earlier detection program will always be underestimated by the focus on statistical hypothesis-testing and the practice of announcing results when the accumulated deficits first become "statistically" significantly different from zero. When used in the context of policy-making, the "key question" targeted by the Canadian Task Force "*Does* screening… decrease breast cancer mortality for women of all ages?" is seriously incomplete. Decision makers need to know how *great* the benefits might be.

SUMMARY

To estimate the magnitude of the impact on breast cancer mortality in a screening program using data from trials, one must recognize the critical roles of the screening regimen, and the time-window in which the delayed deficits are seen. These issues were ignored in the recent Canadian, US, and UK Task Force reports. Reanalysis of data from the same trials, paying attention to the timing of the deaths in relation to the timing of the screening, indicates that yearly breast cancer mortality reductions under a screening program would be at least 40% – double the Task Force's estimate.

REFERENCES

1. US Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med* 2009;151:716-26.
2. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: An independent review. *Lancet* 2012;380(9855):1778-86.
3. Canadian Task Force on Preventive Health Care, Tonelli M, Connor Gorber S, Joffres M, Dickinson J, Singh H, et al. Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *CMAJ* 2011;183(17):1991-2001.
4. Miettinen OS, Henschke CI, Pasmantier MW, Smith JP, Libby DM, Yankelevitz DF. Mammographic screening: No reliable supporting evidence? *Lancet* 2002;359(9304):404-5.
5. Miettinen OS, Karp I. *Epidemiological Research: An Introduction*. New York, NY: Springer, 2012;81.
6. Morrison AS. *Screening in Chronic Disease,* First Edition. New York: Oxford University Press, 1985.
7. Caro J. Screening for breast cancer in Québec: Estimates of health effects and of costs. Montreal: CÉTS, 1990;24. Available at: http://www.aetmis.gouv.qc.ca/site/en_publications_liste.phtml (Accessed January 7, 2012).
8. Hu P, Zelen M. Planning clinical trials to evaluate early detection programs. *Biometrika* 1997;84:817-29.
9. Hu P, Zelen M. Planning of randomized early detection trials. *Stat Methods Med Res* 2004;13(6):491-506.
10. Hanley JA. Analysis of mortality data from cancer screening studies: Looking in the right window. *Epidemiology* 2005;16:786-90.
11. Baker SG, Kramer BS, Prorok PC. Early reporting for cancer screening trials. *J Med Screen* 2008;15:122-29.
12. Hanley JA. Mortality reductions produced by sustained prostate cancer screening have been underestimated. *J Med Screen* 2010;17(3):147-51.
13. Hanley JA. Measuring mortality reductions in cancer screening trials. *Epidemiol Rev* 2011;33(1):36-45.
14. Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* 1977;39(6 Suppl):2772-82.
15. Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: The Malmö mammographic screening trial. *BMJ* 1988;297(6654):943-48.
16. Tabár L, Fagerberg CJ, Gad A, Baldetorp L, Holmberg LH, Gröntoft 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* 1985;1(8433):829-32.
17. Frisell J, Lidbrink E, Hellström 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.
18. Bjurstam N, Björneld L, Warwick J, Sala E, Duffy SW, Nyström L, et al. The Gothenburg Breast Screening Trial. *Cancer* 2003;97:2387-96.
19. Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. *J Natl Cancer Inst* 2000;92(18):1490-99.
20. Fitzpatrick-Lewis D, Hodgson N, Ciliska D, Peirson L, Gauld M, Yun Liu Y. Breast cancer screening. Available at: http://www.ephpp.ca/pdf/breast_cancer_2011_systematic_review_ENG.pdf (Accessed July 26, 2012).
21. Breslow NE, Day NE. Statistical Methods in Cancer Research. Volume II - The Design and Analysis of Cohort Studies. Lyons, France: IARC Scientific Publications No. 82., 1987.
22. Liu Z, Hanley JA, Strumpf EC. Projecting the yearly mortality reductions due to a cancer screening programme. *J Med Screen* [2013 Sep 18. Epub ahead of print].
23. Marcus PM, Bergstralh EJ, Fagerstrom RM, Williams DE, Fontana R, Taylor WF, Prorok PC. Lung cancer mortality in the Mayo Lung Project: Impact of extended follow-up. *J Natl Cancer Inst* 2000;92(16):1308-16.

Received: June 19, 2013
Accepted: September 19, 2013

RÉSUMÉ

OBJECTIFS : i) Estimer de combien baisserait la mortalité si l'on proposait aux femmes un dépistage du cancer du sein dès 50 ans et jusqu'à 69 ans; ii) procéder en utilisant les mêmes essais et les mêmes taux de participation que ceux examinés par le Groupe d'étude canadien; iii) mais dans notre analyse, nous guider sur les différences essentielles entre le dépistage et les traitements du cancer, sur l'enchaînement chronologique qui caractérise les baisses de mortalité produites par un nombre limité de dépistages, et sur les données de mortalité annuelles dans le segment de suivi approprié à l'intérieur de chaque essai; et donc iv) éviter les sous-estimations graves qui découlent de l'inclusion de segments de suivi inappropriés, c.-à-d. trop tôt après l'entrée dans l'étude et trop tard après l'abandon du dépistage.

MÉTHODE : Nous nous sommes concentrés sur les ratios annuels des taux de mortalité dans les années de suivi où, d'après le régime de dépistage employé, on pourrait s'attendre à des déficits de mortalité. Comme les régimes diffèrent d'un essai à l'autre, nous n'avons pas groupé les données annuelles de chaque essai. Pour éviter les valeurs statistiques extrêmes dues au petit nombre de décès annuels dans chaque essai, nous avons calculé les ratios des taux selon des fenêtres mobiles de trois ans.

RÉSULTATS : Nous avons pu extraire des données annuelles dans les rapports de cinq essais. Les données sont limitées pour la plupart par le petit nombre de cycles de dépistage. Néanmoins, elles donnent à penser que le dépistage de 50 à 69 ans résulterait, à chaque âge entre 55 et 74 ans, en une baisse de la mortalité par cancer du sein beaucoup plus importante que l'estimation de 21 % sur laquelle se fonde le rapport du Groupe d'étude canadien.

ANALYSE : En ne tenant pas compte de certaines caractéristiques clés du dépistage du cancer, plusieurs analyses contemporaines sous-estiment gravement l'impact attendu d'un tel programme de dépistage du cancer du sein.

MOTS CLÉS : dépistage du cancer; diagnostic précoce; essais cliniques randomisés; mortalité

Screening for Breast Cancer: What Truly Is the Benefit?

O.S. Miettinen, MD, MPH, PhD

In an article in this issue of the CJPH, Hanley et al.[1] respond to the most recent review by the Canadian Task Force on Preventive Health Care aimed at quantification of the benefit from screening for breast cancer.[2]

These authors have much in common with the Task Force. They, like the TF: are concerned to quantify the aggregate benefit from the screening for some defined type(s) of population for which public policy is, or might be, in favour of the screening; think of this benefit for any such population in terms of the proportional reduction in mortality from the cancer; think that a measure of this reduction can be derived from the studies that have addressed the counterpart of this reduction in experimental cohorts; are not concerned to judge whether the diagnostic work-ups and treatments in a given trial, even dating from a half-century ago, have relevance to practices at present and beyond; do not view it relevant to take note of the protocols for and practices of the diagnostic work-ups and treatments in those studies; and finally, do not think it necessary to judge the validity of the trials used for quantification of the mortality reduction, even if major problems of validity in them are well known.

In this aggregate of principles, the inattention to the well-known problems of validity in the trials is particularly surprising. Suffice it to note that while much emphasis is placed on the trials being randomized, even this feature of them has been prone to be invalid. Thus, S. Mukherjee, in his "biography of cancer",[3] points out and explains how the HIP trial was "instantly a logistical nightmare" (p. 295); and how, in this trial, "The unscreened group had been mistakenly *overloaded* with patients with prior breast cancer" (p. 297; italics in the original). He also points out and explains how "The [Canadian National Breast Screening Study] faltered, …, by succumbing to the opposite sin: by selectively *enriching* the mammography group with high-risk women" (p. 299; italics in the original).

On the other hand, Hanley et al. diverge, profoundly, from the Task Force in the way they think of, and derive, the results from the trials; and in this, these authors take major exception also to the culture that the TF in this respect shares with the trialists themselves.

Different from the TF, Hanley et al. appreciate a fundamental truism that has been obvious to many others before them: that in trials on screening for a cancer, the proportional reduction in mortality from the cancer – in their screened subcohorts, insofar as the reduction occurs at all – cannot be constant over successive intervals of time after the screening's initiation; that it is initially nil, then increases and later declines, and ultimately totally vanishes. Despite these understandings by others, the TF, like trialists themselves, draws from any given trial a single-value result for the proportional reduction in mortality from the cancer; and this is done with no regard for the arbitrarily set durations for the screening and the follow-up for deaths from the cancer, the follow-up starting from the initiation of the screening.

These authors appreciate, also, that the proportional reduction in mortality from the cancer in a screened cohort reaches, under certain conditions, its maximal, asymptotic level, which prevails for a certain duration even after the screening's discontinuation. They do not, however, elaborate on this asymptote, nor do they specify where this has been done (which is their reference no. 4).

Hanley et al. take a keen interest in this experimental asymptote. They take this asymptote to represent what they, like the TF, want to know; that is (to say it again), the proportional reduction in mortality from the cancer resulting from the screening's introduction, after the experimentations, as an available service, to a population (dynamic rather than cohort-type).

They therefore set out to assess this asymptotic level of the reduction on the basis of five of the six trials that the TF made use of, the five from which they could derive the mortality ratios specific to particular subintervals of time since the screening's inception. Their core "finding" was that the asymptotic reduction in those experimentally screened cohorts was at least 40%, and they took this to be the reduction also for a population to which the screening has been, or might be, introduced as a service – thus estimating the proportional reduction in deaths from the cancer in such a population to be much higher than what the TF inferred from the same trials.

Sadly, Hanley et al., like the TF, were mistaken in their goal, i.e., in what they set out to quantify. There is no need for "measuring the mortality impact of breast cancer screening" for the entire population for which it has been, or might be, made available.

There is no population-level benefit from the availability of screening for breast cancer in any meaning other than the sum of the *individual benefits* from the availability of this service to the women constituting the population at issue. These individuals are not concerned with the epidemiological topic of the rate of mortality from the cancer in that population, nor with its subordinate, equally esoteric topic of proportional reduction in this mortality consequent to the screening having become available. The benefit to these individuals, if any, is an instance in which their undergoing a round of the screening leads to detection of a (latent) case of the cancer and the ensuing treatment results in cure of the cancer while otherwise – upon diagnosis prompted by the cancer's clinical manifestations – the disease would no longer be curable. The *value*

Author's Affiliation

Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC

of this individual benefit they understand to be quite individualistic, dependent on their age at the time, among other factors.

The proportional reduction in mortality from the cancer in a screening-eligible population, while thus irrelevant to the members of the population, is also unrealistic to try to quantify, whether in terms of the Hanley et al. approach or that of the TF. One reason for this is the unjustifiability of the premise in these attempts that once the screening is available, the eligible women avail themselves of it, for decades, at the same rate as those in the trials did, for a few years.

The interest that Hanley et al. took in the asymptote of the proportional reduction in mortality from the cancer, as it on certain conditions (which they did not appreciate) is estimable from screened cohorts, would have been justifiable, and highly so, for a reason very different from that which motivated them.

The asymptotic level of the proportional reduction in mortality from the cancer in screening experiments equals something that is critically important to individual women in the population at issue. It equals the proportional reduction in the cancer's rate/probability of incurability, or in its case-fatality rate, attendant to its detection under the screening, when not considering whether the diagnosis is due to the screening or to symptoms emerging between two successively scheduled rounds of the screening. (This is explained in the authors' reference no. 4.)

In this individual-centered, clinical-type framework of thought, the population-level benefit – the sum of the individual benefits (cf. above) – from the screening's availability in a given span of calendar time (e.g., the first year of its availability) is, in plain numerical terms (when not accounting for the valuations of the cures), the total number of otherwise incurable cases that, in the population in that period, were cured by screening-afforded early treatments. This is the period-specific number of detections of the cancer consequent to the screening multiplied by three probabilities: the probability of the case being one of a genuine, life-threatening cancer (rather than overdiagnosed as such); the probability of a screen-diagnosed genuine case of the cancer being incurable by treatment delayed to the time when the cancer already would be clinically manifest; and the probability of undelayed treatment upon screen-diagnosis being curative of such an otherwise incurable case (i.e., the proportional reduction in incurability addressed above, though adjusted for it to be specific to screen-diagnosed cases).

All of the clinical-type probabilities in this calculation of the population-level benefit from the screening – should the latter be of interest – are relevant in themselves: they are germaine to *knowledge-based screening* for breast cancer and for women's decisions to avail themselves of it (while the proportional reduction in the rate of mortality from the cancer, in the population at issue, is not; cf. above).

Laudably, Hanley et al. set out to help correct a major flaw in the still-common way of thinking about, and estimating, the magnitude of the benefit from screening for breast cancer, the benefit as it takes place in trials on the screening, this flaw being a major reason why the extensive research on this topic has resulted in a very high degree of confusion and controversy about the extent of the benefit in those trials, and secondarily about it in actual practice of the screening. These authors make a compelling case for the need to correct this flaw, even though this point of theirs is not new but only routinely ignored. They also call attention to, and illustrate, an alternative measure of the benefit in those trials but, regrettably, fail to grasp the true meaning of this measure, as they too do not proceed from tenable, genuinely first principles.

I remain almost as pessimistic about progress in the research as I have been before[4] – except if the CJPH should now proceed to arrange for public discourse aimed at correcting the various prevailing, ingrained anomalies of the research culture surrounding screening for breast cancer, among other types of cancer.

REFERENCES

1. Hanley JA, McGregor M, Liu Z, Liu Z, Strumpf EC, Dendukuri N. Measuring the mortality impact of breast cancer screening. *Can J Public Health* 2013;104(7):e437-e442.
2. The Canadian Task Force on Preventive Health Care. Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *Can Med Assoc J* 2011. DOI: 10.1503/cmaj. 110334.
3. Mukherjee S. *The Emperor of All Maladies. A Biography of Cancer.* New York, NY: Scribner, 2010.
4. Miettinen OS. Screening for a cancer: A sad chapter in epidemiology. *Eur J Epidemiol* 2008;23:647-53.