Epidemiologic Reviews
Vol. 33, 2011
DOI: 10.1093/epirev/mxq021
Advance Access publication:
May 30, 2011

Measuring Mortality Reductions in Cancer Screening Trials

James A. Hanley*

* Correspondence to Dr. James A. Hanley, Department of Epidemiology, Biostatistics and Occupational Health, Purvis Hall, 1020 Pine Avenue West, McGill University, Montreal, Québec, H3A 1A2, Canada (e-mail: james.hanley@mcgill.ca).

Accepted for publication December 9, 2010.

Randomized trials involving large numbers of people and long follow-up have helped measure the mortality reductions achievable by screening for cancer. However, in many of these trials, the reported reductions have been modest. Part of the reason is the inappropriate way the reductions have been calculated. Analyses have largely ignored the fact that there is a time window in the first several years after screening begins in which there cannot be a sizable mortality reduction, followed by one in which the reductions become evident, and—unless screening is continued—a third window in which mortality rates in the screened group revert to those in the unscreened group. This review uses time-specific mortality ratios to address the timing and extent of the reductions achieved in trials of screening for prostate, breast, and colorectal cancer. The author finds that the mortality reductions reported in the literature have substantially underestimated what might be accomplished with continued screening. The natural history of the disease, the frequency of screening, and the duration of follow-up determine the time patterns in the reductions observed in trials. Without appropriate analyses, results from cancer screening trials will be distorted.

mortality; neoplasms; proportional hazards models; randomized controlled trials as topic

Abbreviation: ERSPC, European Randomized Study of Screening for Prostate Cancer.

> It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once.
>
> Francis Galton, 1889

INTRODUCTION

Before implementing an expensive organized program for earlier detection and treatment of a cancer, funders need good estimates of the mortality reductions and other savings that will ensue so that they can weigh them against the costs. Individuals contemplating being screened must also consider this trade-off.

Randomized trials involving very large numbers of people and long follow-up have provided estimates of the mortality reductions achievable by screening for cancer. However, as noted in the recent European Randomized Study of Screening for Prostate Cancer (ERSPC) (1), the *reported* reductions from many such trials have been modest.

In some trials, the modest/absent reductions are not surprising given the weak nature of the screening tools. However, there are also methodological reasons for many of them. Some screening studies have used just one round of screening. In some circumstances, this regimen emulated what would be implemented in practice and produced a detectable signal, whereas in others it did not. In others, follow-up has been too short to enable the full mortality reductions to be expressed, or reliably measured.

This review focuses on a critical aspect of data analysis and reporting, even when follow-up has been sufficient. Virtually all reports have effectively averaged 1) the (expected) nonreductions early in follow-up and 2) the mortality reductions that emerge later, and they have presented this average as a one-number summary measure. This measure systematically dilutes the estimate of the mortality reductions produced by screening. In the case of prostate cancer, with its long sojourn times in the various preclinical and postdiagnosis states, the underestimation is considerable. In a few instances (2, 3), the estimate has been diluted further by including an excessive amount of follow-up time in the calculation, that is, by averaging not just quantities 1) and 2) but 1), 2), and 3) the further (expected) nonreductions seen in the years long after the last round of screening could have had any effect.

Such a one-number summary measure may be adequate when studying the results of interventions with virtually immediate and long-lasting effects, such as some vaccinations (4), many medications (5), and screening for abdominal aortic aneurysms (6). For these, it is logical to cumulate the outcome events from the time the intervention commenced and to report a single (proportional hazards) rate ratio. Such a summary is clearly not appropriate for measuring the mortality reductions produced by cancer screening. The appropriate principles for measuring them were set out a generation ago in Morrison's classic textbook on screening (7). This review reveals that an inappropriate summary measure that has become predominant over the past 20 years has led to considerable underestimates, and it illustrates how data from cancer screening studies can be appropriately analyzed by attention to time specificity.

The review begins with an orientation that focuses on the numbers of cancer deaths if a population is subjected to a screening program versus if it is not and on how trial data, combined with reasonable assumptions, can be used to project the reduction expected from implementing such a program. It uses the European trial of prostate-specific antigen–based screening for prostate cancer to illustrate how the data from screening trials should be analyzed, objectively and time specifically. It finds that the prevailing data analysis practices have also led to serious underestimates in trials of screening for breast and colorectal cancer. The review also has implications for the design and analysis of future screening trials.

ORIENTATION AND DATA ANALYSIS PRINCIPLES

Basic counterfactual comparison

The extent to which, say, a prostate cancer screening program reduces cancer mortality is depicted in Web Figure 1 (the first of 7 supplementary figures, each of which is posted on the *Epidemiologic Reviews* Web site (http://epirev.oxfordjournals.org/)), which shows what would occur, over some time scale, in the absence of a screening program and if a program had been in place for some portion of that time. The impact, measured by the difference in the areas of the 2 smallest circles (these circles represent the numbers of fatal cancers under the ''screening absent'' and ''screening present'' scenarios, respectively), occurs among the ''otherwise fatal'' cancers, that is, those that would have proved fatal despite treatment at the time they would have presented clinically.

Applying results of trials of screening to screening in practice

In Web Figure 1, the ''time scale'' and population were deliberately left vague, as were the screening frequency, intensity, and uptake. In actual trials, such as the ERSPC, previously unscreened men of different ages are enrolled and followed in a screening or control arm; a (ideally, high) percentage of men in the screening arm and a (ideally, low) percentage of men in the control arm are screened at each round, screening is terminated after a number of rounds, and data are analyzed some number of years after the first men were invited to participate.

In practice, screening would be carried out differently: men would be invited to be screened once they become, say, age 50 years and repeatedly until, say, age 70 years. Of interest is how many fewer cancer deaths there would be annually, under steady-state conditions, in the age range of, say, 55–80 years under this program. This comparison is shown in Web Figure 2. The focus of this review is the rate ratio curve shown in Web Figure 2(b), modeled after Figures 2–5 (b) in the work of Morrison (7).

This time curve is central for 2 reasons. First, it reminds us that the mortality reductions produced by screening and earlier treatment of cancer do not, and cannot, become apparent immediately after screening commences: if the cancer, ''cured'' today because of earlier (screen) detection, had not been treated in time, it would have proved fatal several years in the future. For some cancers, the delay is considerable. Despite this, most analyses of data from both trials and nonexperimental (cohort-type) studies merge the deaths in this ''early no-reduction'' window with those in the time or age window during which reductions do become apparent. (Interestingly, case-control studies consider the latency between exposure to a disease-causing agent and occurrence of a disease, and between the earlier treatment prompted by screening and the time when the cancer would otherwise prove fatal (2).) Nor can mortality reductions produced by screening and earlier treatment of cancer persist indefinitely after screening is discontinued; nevertheless, some analyses of data from screening trials have also merged the deaths in this ''late no-reduction'' window with those from the time or age window during which reductions would have been apparent (2, 3).

Second, the curve depicted in Web Figure 2(b) can be estimated from data from screening trials. This rate ratio curve can then be applied to the ''no-screening'' curve of numbers of deaths in Web Figure 2(a) to produce an estimate of the ''screening'' curve of numbers of deaths in Web Figure 2(a). The differences between the 2 curves of numbers of deaths can then be used to calculate the absolute impact of different versions of the screening program that might be offered in practice. Going from (b) to (a) requires some assumptions. In the curve from a typical trial, ''t_0''—the time of the first screen—refers to a different age at first screen for different participants: in the ERSPC, the ''core'' age range at intake was 55–69 years. In the ''in practice'' curves shown in Web Figure 2(a), there is a common t_0: age 50 years at first screen.

Before considering the ERSPC data, we use additional hypothetical data to illustrate how—before we examine the data—we should analyze them.

Expected pattern with 20 years of follow-up in a trial such as ERSPC

Web Figure 3 uses hypothetical, but realistic data to show how, unlike an overall ratio, *time-specific* mortality rate ratios directly measure the timing and magnitude of the reductions produced by different screening regimens and are not affected by the length of follow-up. Construction of Web Figure 3 is described in the Appendix.

The impact of a particular screening regimen (row) can be seen in Web Figure 3(a) by examining the shaded portions of the squares (i.e., the numbers of prostate cancer deaths) for the row in question: the ''missing'' portions of the different squares show the extent to which the yearly numbers of deaths under the no-screening scenario (top row) are reduced by screening. As expected, the shortest impact is with a one-time-only screen: the reduction is limited to a window beginning at about 6 years and (practically) ending at about 10–12 years. The greatest and most obvious impact is with a ''0, 4, 8, and 12 years'' regimen (last row). In this scenario, mortality is considerably lessened from about year 6–7 onward and remains diminished at least until year 20. There is virtually no visible impact for the first 5 years in any regimen: just about all of the otherwise-fatal cancers in these years would have been incurable when screen detected, and the otherwise-fatal cancers that are ''missing'' from follow-up years 6–7 and beyond are absent because of their timely detection and treatment years earlier.

The effects of different screening regimens would have been somewhat blurred if, as in the prevailing approach, the time element were ignored, for example, if we were merely shown, say, the *total* number of prostate cancer deaths over the 20 years in row *i* (e.g., 1,055 if [$i =$] 1 screen or 895 if [$i =$] 2 screens) and the corresponding number of deaths in row 0 (1,177 in the 0-screens arm), *without their time location*. The ''percentage reductions'' in the second-to-last column of Web Figure 3(a) are calculated after a full 20 years of follow-up. Few reports would wait this long; nevertheless, the column illustrates why, even under these ideal circumstances, the prevailing measure, which ignores the highly time-specific nature of these reductions and effectively assumes a constant-over-time hazard ratio, would underestimate the impact of a screening program.

Since the first cancer screening trials, investigators have tried to avoid the dilution caused by including cancer deaths that occur after the window in which the benefits of the regimen become manifest (7–9). The recognition that the dilution caused by including those that occur before this window is more recent (2, 10, 11), but the message to avoid it has been largely unheeded. Moreover, analysts have overlooked an additional dilution inherent in the prevailing ''single number summary'' measure: this attenuation is highlighted by the entries in the ''over 9 years'' column of Web Figure 3(a). Whereas the numbers of cancer deaths in a traditional therapeutic trial are larger proximal to the time that patients enter the trial, the corresponding numbers in a screening trial are smaller at this end of the time scale and larger at the other end. Thus, the shorter and more inadequate the follow-up, the more heavily is the overall percentage reduction over this period (incorrectly) weighted toward the lesser reductions in these early follow-up years.

This time insensitivity in the analysis of actual screening trials is puzzling: other trialists test whether their data justify use of a proportional hazards model, that is, of a single hazard ratio. Biologic principles suggest that this ''constant-over-time right-from-the-outset reduction'' assumption may govern the time pattern of the effects of screening for some conditions, such as abdominal aortic aneurysms, but not cancers.

The next section documents that inappropriate data-analysis practice, and thus underestimation, are widespread in actual trials of cancer screening.

SCREENING TRIALS: PROSTATE CANCER

The ERSPC

The main features of the prostate cancer screening trials that have been reported have been discussed elsewhere (12) and are shown in Web Table 1 (http://epirev.oxfordjournals.org/) (1, 13–18). As in that article, the focus here is on the ERSPC because its larger sample size and the substantial difference in the participation rates in the 2 arms meant that only it has any substantial ''resolving power.'' In the ERSPC report, the effect of screening on prostate cancer mortality was expressed as one number, derived from the numbers of prostate cancer deaths over the entire period of observation available for each man (range, 3–15 years; average, 9). Over this period, 214 prostate cancer deaths occurred during 643,401 man-years of observation in the screening arm and 326 during 785,585 man-years in the control arm. These are the basis for the reported rate ratio of 0.80 and the conclusion that ''PSA-based screening reduced the rate of death from prostate cancer by 20%'' (1, p. 1320) (95% confidence interval: 2, 35).

The 2 curves from the graph in the ERSPC report are redrawn in Web Figure 4(a). On the basis of these curves, the authors did note that ''the rates of [prostate cancer] death in the two study groups began to diverge after 7 to 8 years and continued to diverge further over time'' (1, p. 1325). My reanalysis (12) quantified this divergence, thereby providing a more appropriate measure of the reduction in mortality produced by screening than the reported 20% figure. It examined the prostate cancer mortality in follow-up intervals, thus allowing both the timing and extent of the reduction to become clear.

The reanalysis followed the approach used to extract the signal from the data from a breast cancer screening trial (10, 11). A formal curve-fitting approach was also used to further reduce statistical noise, to objectively measure the steady-state reduction in mortality, and to identify when it reached this level, as precisely as the data allow. Candidate curves with the same general form as curve 4 in Web Figure 3(b) were considered, since a repeated 4-year screening interval was used in the countries that contributed more than 80% of the men, and the mortality in the screening arm would not be expected to have begun to revert upward toward that in the control arm until after the end of year 12.

Web Figure 4(b) shows the yearly numbers of prostate cancer deaths in each arm, along with the mortality rate ratios for the intervals centered on years 2, 3, . . ., 12. They indicate that, after an expected delay (which the data indicate is approximately 7 years), the prostate mortality reductions that become evident in years 9 and beyond are statistically significant and considerably greater than the reported 20% reduction in the rate of prostate cancer deaths. The best estimate was that, although the rate ratio became nonnull starting at 6–7 years, the steady-state reduction has not yet been reached: the point estimate so far is a sustained 67% reduction (80% confidence interval: 30, 89) beginning at year 12. However, the numbers of deaths are not sufficient to precisely measure the signal in the very follow-up time window where it is probably strongest. Fortunately, this potential can be achieved by collecting additional follow-up data (although the report (1) was published in 2009, follow-up ended in December 2006, just as the pattern began to emerge).

The reanalysis respects the intention-to-treat principle, using time-specific rates to reveal the expected nonproportional-hazards pattern. The objective curve-fitting approach avoids having to ''prespecify'' when the reduction reaches steady state; it specifies the smooth form of the rate-ratio curve but allows the data to inform us about the 2 essential parameters: the timing and extent of the cancer mortality reduction enabled by screening.

How others have used time specificity to analyze data on screening

Most of the above principles used in the reanalysis were addressed in a classic textbook devoted to scientific evaluation of screening programs (7). It used breast cancer as an illustration and specifically targeted the ''late no-reduction'' window, but it did not dwell on the ''early no-reduction'' window. Screening trials exclude people with already diagnosed cancer, so the numbers of cancer deaths in this ''early'' window may not be substantial. In prostate cancer, however, with its longer lead time, these deaths can make up a sizable portion (Web Figures 3 and 4), particularly if the follow-up window in which reductions become apparent is short relative to this (2). Thus, including deaths in the ''early no-reduction'' window in one overall rate ratio can produce considerable underestimation.

Chapter 4 of the Morrison textbook (7) was devoted to assessing the value of early treatment by using experimental studies. The author recommended that ''the length of the planned observation period should reflect the natural history of the disease under study'' (7, p. 89) and thus ''be long enough to permit observation of most or all of the changes in mortality rate from early treatment. In other words, [it] should at least encompass the times at which most screen-detected cases would die if they were not treated early; this period is equivalent to what the length of the clinical phase would be without early detection and treatment, plus the lead time'' (7, p. 89). However, the period should not be too long, ''since cases that arise after screening is over cannot possibly be benefited by early treatment; including them obscures its value'' (7, p. 36). Baker et al. also addressed this dilution, termed it ''post screening noise,'' and proposed ''early reporting at the time during the follow-up period when post-screening noise first starts to overwhelm the estimated effect of screening as measured by a z-statistic'' (9, p. 122). In the ERSPC data, the ''*front* end'' contains the noise.

Implications for the design and analysis of prostate cancer screening trials

The editorial (19) accompanying the ERSPC report (1) looked forward to the results from the Prostate Testing for Cancer and Treatment (ProtecT) trial. However, the pattern of results in Web Figure 3 suggests that it may be difficult to show the benefit of one-off screening that (currently) is being tested in that trial. Another editorial also noted that the ERSPC study has ''not fully matured, and it is essential to continue the follow-up in each group'' (20, p. 222). It added, ''unfortunately, the authors of the ERSPC have already performed 3 interim analyses. The criteria for statistical significance in subsequent analyses have become much more rigorous as the number of interim analyses has increased. The ERSPC has 'eroded its alpha,' meaning it may have difficulty conducting future statistically valid analyses. It may be impossible for future analyses of the ERSPC to have a statistically significant finding that screening is beneficial'' (20, p. 222). The editorial might have added that most of the alpha was spent looking in the wrong time window (21).

SCREENING TRIALS: BREAST CANCER

Overview

The 8 known randomized trials of mammography are extensively reviewed in an American College of Physicians monograph (22). The numbers of rounds of screening varied from 2 to 9 (median, 4–5) at screening intervals of 1 to approximately 2.5 years (median, 1.5), and mean follow-up ranged from 13 to 17 years. The main results were presented as 8 ''relative risks,'' each one being the ratio of the breast cancer death rates in the overall period of observation in the screened and control groups. These relative risks ranged from 1.02 to 0.68, and, in a meta-analysis of the 7 trials rated of fair quality or better, the summary relative risk was 0.84, that is, a 16% reduction in breast cancer mortality during the overall period of observation. An Appendix in the American College of Physicians report gives more detailed results for each study, including relative risks taken from earlier reports with shorter mean follow-up, and the figure in the article shows a pattern in the successive reports that ''suggests a gradual decrease in benefit with longer observation time'' (22, p. 85). In the analysis restricted to women aged 50–74 years at entry, the fitted relative risk function varied from approximately 0.62 during an average follow-up of 5 years (38% reduction in the rates for that period) to approximately 0.87 during an average follow-up of 18 years (13% reduction).

The HIP study

The first completed of the 8 trials, the Health Insurance Plan of Greater New York (HIP) study (8) is a good example of the uninformative window following the first screen, and likewise the window located some years after the last screen, and the extent to which their inclusion in an overall average attenuates the effects of screening. Morrison is the earliest we know of to have directly examined—via yearly numbers of deaths in the screened and control groups—the detailed ''temporal relation between screening activities in the [first four years] of the study and [subsequent] breast cancer mortality'' (7, p. 76). He uses as an illustration the results from this ''landmark'' HIP study. He presents both cumulative and yearly numbers in his Figure 4-1 and Table 4-1 and pays particular attention to the latter: ''Over the first three years of observation, the numbers ... in the two groups were similar. During the fourth through seventh years, the breast cancer mortality rate was lower in the screened group ... After the seventh year, there appears to have been no systematic or substantial difference: the breast cancer mortality rates in the two groups again were similar'' (7, p. 77).

The absence of yearly denominators precludes the use of rate differences, but since the screened and control groups were the same size, we can calculate yearly (what Morrison (7) calls *interval-specific*) mortality rate ratios. These are replotted in Web Figure 5. Morrison summarized the effect based on the cumulative numbers of deaths since entry, but only up to the end of year 7. He expressed the observed difference in deaths after 7 years ($n = 38$) as 35% of the number in the control group up to that time ($n = 108$) and not as a 29% reduction based on all deaths up to the end of year 9. Notably, Morrison let the year-by-year data reveal the pattern and did not prespecify the location of the evidence-containing window, although its general location was known a priori.

The author of the source report (8) for Morrison's data (7) was also keenly aware of the dangers of diluting the estimate of the effect of 4 yearly screening examinations by using too long an observation period. He used data-analytic strategies to ''reduce 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 with respect to both the annual rate of detection through regular medical care and the stage of the disease at time of detection'' (8, p. 2776).

The Swedish trials

Just as including too long an observation period after screening ceases in a single ratio of average rates will attenuate the estimates of the yearly reductions that continued screening would still (after some lag) achieve, so too will including in that single ratio the deaths occurring in the early years after the first screen. Such deaths could not have been affected by screening: investments in screening today produce returns in only some future years, just as the full, steady-state individual- or population-level cholesterol reductions following initiation of statins are not apparent for some time (the time curves (23, 24) have intensity-dependent patterns very similar to those in Web Figures 2(b) and 3(b), with the time scale in days rather than years).

The authors of a 1993 cost-effectiveness analysis (25) were keenly aware of this time lag. So too were the authors of a 2002 reanalysis (10, 11) of the data from another screening study in the American College of Physicians report, the Malmö trial. This reanalysis was prompted by a meta-analysis (26) that focused on the 2 trials (Malmö, Sweden, and Canada) that the meta-analysts deemed to be methodologically sound. Unfortunately, the meta-analysts neglected the critical issues of the screening intensities and durations in these 2 trials and the location in time of the breast cancer deaths. Since the meta-analysis merely used the cumulated deaths over the entire period of observation in each trial, it is not surprising that the impact of screening on mortality was obscured. The impact becomes evident only if there is sufficient screening and follow-up and if time-specific data are used (10, 11). The Malmö study is of particular note because it used sustained screening and thus allowed the steady-state mortality reduction to be measured.

The task of analyzing the later follow-up in another Swedish study—the 2-county trial—shows just how important time specificity is. When a significant reduction in mortality from breast cancer was achieved during active screening in 1985, the passive study population was invited to screening, on average 7–8 years after randomization (27) (unlike Malmö, other Swedish trials did likewise). Use of the subsequent data, or portions of them, has been controversial. However, what is not in dispute, but has been overlooked, is that over the first 7 years, that is, before any expected convergence of the yearly mortality rates, there was no reduction for years 1–4 but a reduction of approximately 45% for years 5–7 (refer to Figure 2 in the 1985 report (28)). The 1985 report noted a reduction of 31%.

The similar lack of time specificity inherent in the cumulative mortality analyses in the other 5 breast cancer studies reports (and in the American College of Physicians report generally), with its neglect of both the timing and frequency of screens and its ''mixing of irrelevant experience with the relevant experience'' (10, p. 3), means that these analyses have underestimated the impact of earlier detection and treatment. A more complete analysis of these 8 studies focusing on time-specific mortality reductions would help decision makers who invest in yearly breast cancer screening today to quantify what gains they can expect to realize starting some (it seems to be 4–5) years from now.

This neglect of time specificity has not been limited to randomized trials of breast cancer screening. The same dilution, caused by ''averaging the relevant with the irrelevant'' time windows, has also occurred in nonexperimental studies of the effect of introducing national cancer screening programs. For example, the Norwegian breast cancer screening program, recently reported on (29), had not been in operation long enough for the full effect on mortality from the disease to have taken place by the end of the reported follow-up. By then, the various county-specific programs were only in their 2nd, 6th, and 10th years of operation. Given how recently they had been introduced, the reported measure of their impact on mortality is severely downward-biased. Screening for a cancer is aimed at early detection and treatment in cases that otherwise would prove fatal. Proper measurement of the extent of the reduction in mortality, which follows several years later, requires specificity to the duration of screening and the time since it was begun. The time lags in the consequences of 1–5 rounds of biennial screening are illustrated schematically in Web Figure 6. From this figure, it is evident that, by merging years 1–10 in the first region to begin screening with the fewer years (1–6) in the other counties, an already diluted measure (in the first counties) was diluted even more.

Furthermore, by not following women screened in their mid- and late 60s until the ages at which the benefits for these women would be realized (i.e., their 70s), the data analysis missed a further portion of the (as-yet, probably small) mortality impact of the screening program. Breast cancer deaths in their early and mid-70s of women who had been screened in their mid- and late 60s were conflated with breast cancer deaths in their 70s and early 80s of women who had not been screened in their 60s. Just as there were deaths of women in their 30s and 40s, all deaths of women in their 70s were treated as though none of them could have been affected by screening of women aged 50–69 years.

SCREENING TRIALS: COLORECTAL CANCER

The American College of Physicians monograph (30) listed 3 completed randomized studies of screening using fecal occult blood testing. The United Kingdom and Danish studies randomly invited members of the eligible population to no screening or to biennial (i.e., every 2 years) screening with a nonrehydrated test and followed them for an average of 8 and 10 years, respectively, when the corresponding reductions in cumulative colorectal cancer mortality were 15% and 18%. The US trial randomized enrolled volunteers in equal numbers to no, biennial, or annual screening with a rehydrated test and reported the results after 13 years and again after 18 years of follow-up, using the percentage reduction in cumulative colorectal cancer mortality as the measure of effect. A separate report at 18 years examined reductions in the cumulative incidence of new colorectal cancers. A reanalysis of these 3 reports using year-by-year event rates, and of the temporal relation between the actual screening activity schedules and subsequent cancer mortality, has important lessons for the analysis of other cancer screening studies.

One important feature of the US trial (31), generally ignored, is its screening schedule. The authors reported that ‘‘the study was initially designed with a 5-year screening period during which five annual screens and three biennial screens were to be completed. The screening began in February 1976 and was completed by December 1982 (phase I). During the subsequent follow-up phase, it became apparent that the death rate in the control group was considerably lower than expected due to a large healthy volunteer effect. This led to a recommendation from the study's Policy and Data Monitoring Group to reinstitute screening. A second phase of screening began in February 1986 and concluded in February 1992 (phase II). Because of the hiatus, screen group participants had an interval ranging from 3 to 5 years between the end of the last screening interval of phase I and the first screen attempt of phase II'' (31, p. 434).

In a report on colorectal cancer incidence (32), the reduction in cumulative incidence in the screened groups (due to the removal of precancerous lesions) was just under 20%. My reanalysis (2) showed that, whereas the cumulative incidence curves masked it, the yearly incidence ratios showed a clear and expected input-output pattern: a lag from the phase I screening to impact, a lag from the discontinuation of screening to the loss of impact, and (after screening resumed) the lag from phase II screening to impact. Moreover, I estimated that had screening not been interrupted, there might have been a 40% sustained reduction in incidence.

The 13-year report (33) showed reductions of 33% and 6% in cumulative colorectal cancer mortality in the annual and biennial groups, respectively; the 18-year report (31) showed respective reductions of 33% and 21%. I have since applied a time-specific reanalysis to these data. The results, shown in Web Figure 7, offer a more cogent reason for the small and statistically not significant 6% reduction seen in the biennial group at year 13 than the play of chance on small numbers of deaths. The less intensive screening schedule, and the cessation of screening 9 years before, meant that the yearly mortality rates in the biennial group had returned to the same levels as those in the unscreened group. A single mortality rate for the entire 13 years is an average of the (expected) nonreductions in later and earlier years, and the reductions for a few years in between. The ‘‘cures'' effected by the biennial screening in phase II become evident in the mortality rate ratio statistics for years 15–18.

This example emphasizes the importance of a data display that reveals rather than conceals, that reflects prior knowledge of the natural history of cancer and the schedule and other particulars of screening. It displays the effect of the interruption in screening while the researchers secured additional funding. As in Web Figure 1, it also helps those in nonresearch settings visualize the reductions that would be realized each year once a program reached steady state.

A 2008 meta-analysis (34) included one additional fecal occult blood screening trial from Sweden and the 18-year results of the United Kingdom trial. Given the different recruitment methods, compliance levels, schedules and durations of screening, and lengths of follow-up (refer to Web Table 1 of the meta-analysis), a display of the 4 rate ratio time curves would be more informative—and would serve as an informal meta-regression. The time-specific ratios calculable from the data in the United Kingdom study show clear nonproportionality: in the 3-year moving windows centered on years 2–15, where the numbers of deaths are reasonably large, the percentage mortality reductions are 5, 17, 15, 23, 17, 23, 23, 16, 15, 6, 4, −2, 1, and 2. The reported percentage reduction in cancer mortality over the entire follow-up period (median, 12; range, 8–18) was 13; it was given a weight of 40% in the meta-analysis.

Detailed mortality data have not yet been published for the Swedish study; the information for the meta-analysis was obtained via a 2005 personal communication between the authors of the meta-analysis and the authors of the Swedish study, reporting a 16% reduction over a follow-up period of 15.5 years. In view of the limited screening used (a fecal occult blood test repeated just once an average of 20 months after the first), time-specific rate ratios are likely to be more informative than an overall reduction that is likely to be further diluted by even more follow-up. The effect of screening for cancer of the colon may indeed appear in 2 separate waves, resulting from the removal of already cancerous and precancerous lesions; but, since the last screening was at 20 months, can one expect that this second reduction will be seen in years 15–20?

The recent report on once-only flexible sigmoidoscopy screening to prevent colorectal cancer (35) explicitly discussed the general nonapplicability of the proportional hazards model for analyzing data from cancer screening studies. The report also has several other notable features. It used time-specific hazards when addressing cancer incidence; curiously, it still used cumulative displays for mortality. Over the entire follow-up of more than 11 years (median; interquartile range, 10.7–11.9), the mortality reduction was 31%. However, even in the cumulative graph, a quite-early-on reduction, one that gradually increased and persisted for the full 10 years of follow-up, is obvious and striking.

In a more detailed, time-specific reanalysis, the percentage reductions were in the teens in the early follow-up years, in the twenties in the middle, and in the thirties late in follow-up. The trial used a narrow age-at-intake window, 55–64 years, ‘‘based on observations suggesting that most people who develop a distal colon cancer will have developed an adenoma by 60 years of age, and that removal of adenomas by sigmoidoscopy provides long-term protection against the development of distal colorectal cancer’’ (35, p. 1624). Cancers of the rectum and sigmoid colon, and the proximal colon, were analyzed separately for their incidence (with no effect of sigmoidoscopy on the latter) but not for mortality. Case-control studies have found that sigmoidoscopy screening considerably reduces mortality from cancer of the rectum or distal colon only, but it does not reduce fatal colon cancer above the reach of the sigmoidoscope (36).

In Galton’s example (given at the beginning of this review), the contrast in scenery was obliterated by the conflating of places. The effect of sigmoidoscopy screening on overall colorectal mortality was diluted by the anatomic conflating of distal and proximal cancers. This review has shown that the effect of screening on cancer mortality has repeatedly been diluted by the temporal conflation of the follow-up time segments before, during, and after the numbers of cures produced by screening-initiated treatment become conspicuous (by their absence from the mortality statistics).

DISCUSSION

As is the study of history, clinical trials are a guide to the future. In cancer screening, we need to calculate the start-up and ongoing costs of, and unavoidable harms (some beginning in year 1) caused by, screening programs that lead to earlier detection and treatment of a cancer. To decide whether the benefits are worth these costs, we need a good estimate of the time patterns of the future mortality reductions (and other savings) that will ensue when/if a program is implemented—and not necessarily with the limited screening schedule or time horizon, as in the trials. This review has shown that the one-number summary measures in the reports of trials underestimate the steady-state mortality reductions that would be realized with a sustained screening program, or the more transient reductions achievable with a one-time screen. The complex shape of the ‘‘invest now, realize gains later’’ time function means that, to adequately describe what might be achieved with a particular regimen of cancer screening, a more comprehensive set of time-specific measures needs to be reported.

When evaluating screening for cancer, trialists must endure several years of nonreductions before they learn anything about the possible benefits of screening. Similar time principles apply regarding several other preventive measures in public health, and data analysts are careful not to routinely extract a single ‘‘constant over time’’ measure using an inappropriate proportional hazards model. Two examples of the critical role of ‘‘time specificity’’ are the transient (beneficial and adverse) effects of some medications, and the smaller reduction in the risk of becoming infected with human immunodeficiency virus in the first 6 months following adult circumcision (37) but then the larger (and presumably lifelong) risk reduction once the wound has healed. The clearest example of an immediate benefit of screening is in the case of abdominal aneurysms; because of their ‘‘ticking time bomb’’ nature, their detection and repair results in an immediate and constant reduction in mortality from rupture (6). The yearly reductions are approximately 50% each year from year 2 onward.

Those who evaluate screening programs for cancer cannot expect such nearly instant gratification, but they can take an important time lesson from these findings: the benefits that await them are usually revealed more clearly and more quickly by using a time-specific analysis rather than a non-time-specific, one-number measure. Recognition of the difference between interventions with immediate and delayed effects should prompt similar reanalyses of the data from trials of screening for other cancers and similar analyses in yet-to-be reported cancer screening trials.

Those who perform meta-analyses of data from cancer screening programs must also address this time specificity. As is clear from Web Figure 3, the tables in the American College of Physicians reports (22, 30), and the study-specific data in meta-analysis of fecal occult blood testing (34), different screening regimens, different degrees of compliance, and different methods of subject selection—and even different funding patterns—produce different response curves. A traditional (implicitly proportional hazards) meta-analysis that uses the overall arm-specific numbers of cancer deaths and persons (the typical input to Cochrane software (38)) will therefore produce a rate ratio that conceals more than it reveals and yield the type of meaningless average that Francis Galton humorously cautioned against.

The meta-analysis of the 6 randomized controlled trials of screening for prostate cancer (largely overlapping those in Web Table 1) is an example of such an average. It found ‘‘no significant effect of screening on death from prostate cancer (Risk ratio = 0.88, 95% confidence interval: 0.71, 1.09)’’ (39, p. 1). It did so by merging the results from 5 of the trials, even though, as is evident from Web Table 1, they differed greatly in the frequency and duration of screening, in participation rates, and in the length of follow-up. The report noted that ''all trials had one or more substantial methodological limitations'' (39, p. 1), but, interestingly—as though they were of no consequence—the frequency and duration of screening, and the location of the prostate cancer deaths in the time domain, were entirely ignored.

Law (40) carried out a similarly time-non-specific meta-analysis when he combined the results of the American and European trials (an 11% increase and a 20% reduction, respectively) without considering these time locations. However, he then correctly dismissed the summary number (an 8% reduction) as meaningless: even if one restricted attention to the appropriate time window, the small difference in screening rates in the 2 arms of the American study precluded any meaningful contrast in this window. Moreover, he noted the ''delay in benefit'' in the European trial:

> The prostate cancer death rate in the screened group in the European trial diverged from that in the control group only seven years after screening was initiated (see Figure 2). . .. Perhaps a better summary of the European trial result is not the 20% overall reduction in prostate cancer mortality, but the combination of no reduction in the first seven or so years and a reduction of about 50% after 10 years (40, p. 109).

A particular cancer screening regimen in a particular setting produces as its ''output'' a particularistic hazard ratio time curve, one driven by (a function of) all of the above factors; moreover, if the screening is at all effective, it cannot exhibit a ''proportional hazards'' pattern until it reaches steady state. In the same way, the reductions in cholesterol after beginning to use statins (23, 24) are a function of species, dose, and (especially) time. Clearly, it is not possible to meta-analyze the 6 cholesterol time curves (all short-term) in these reports into a single meaningful number. Likewise, unless it limits itself to the steady state that follows sustained screening, any single number from a ''meta-analysis'' of the (dose-specific and time-specific) results of the trials of screening for a cancer cannot capture one of its most important elements, namely, its time specificity. However, several ''meta-analyses'' of screening trials have been carried out, not just in relation to cancer of the prostate but also of the lung, colon, and breast.

If one were to perform a meta-analysis, it should target some common and relevant feature (the nadir, or preferably the sustained asymptote) of the various ''response curves.'' Just as when confronted with the 6 cholesterol reduction curves, it would make more sense to use a regression analysis to relate the depth of the nadir (or the asymptote) to the dose (compliance) and to other measures of the intensity of the activity in question.

Some (18) have attempted to deattenuate the dilution caused by the often-substantial ''noncompliance'' and ''cross-contamination'' in cancer screening and other (41) trials. Their target is the expected benefit that would accrue to those who actually avail themselves of screening, rather than that from an intention-to-screen analysis. Such analyses, involving certain assumptions, are controversial, although in many instances one might make a case why participation is blind (orthogonal) to cancer risk or the likelihood of cure in the absence of screening. The use of time-specific mortality ratios respects the randomization and relies on the full, actual, time-specific data themselves to teach us the timing and full extent of the mortality reductions produced by offering people the opportunity for earlier detection and treatment. It might be possible to undo both conflations. However, before trying to bring out the fullest signal using mathematical modeling and instrumental variables, we should first address time specificity. Even in a cancer screening trial with perfect adherence to randomization, the (necessarily delayed) signal will be diluted—and even completely missed—if time intervals are conflated.

ACKNOWLEDGMENTS

Author affiliation: Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Québec, Canada.

This work was supported by the Natural Sciences and Engineering Research Council of Canada and Le Fonds Québécois de la recherche sur la nature et les technologies. These 2 funding agencies had no role in the study.

The author thanks C. Begg, S. Hanley, J. Kaufman, M. McGregor, O. Miettinen, G. Paradis, and I. Shrier for their input.

Conflict of interest: none declared.

REFERENCES

1. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med*. 2009;360(13):1320–1328.
2. Hanley JA. Analysis of mortality data from cancer screening studies: looking in the right window. *Epidemiology*. 2005; 16(6):786–790.
3. Marcus PM, Bergstralh EJ, Fagerstrom RM, et al. Lung cancer mortality in the Mayo Lung Project: impact of extended follow-up. *J Natl Cancer Inst*. 2000;92(16): 1308–1316.
4. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. MOPH-TAVEG Investigators. *N Engl J Med*. 2009;361(23):2209–2220.
5. Ridker PM, Danielson E, Fonseca FA, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. JUPITER Study Group. *N Engl J Med*. 2008;359(21):2195–2207.
6. Thompson SG, Ashton HA, Gao L, et al. Screening men for abdominal aortic aneurysm: 10 year mortality and cost effectiveness results from the randomised Multicentre Aneurysm Screening Study. Multicentre Aneurysm Screening Study Group. *BMJ*. 2009;338:b2307. doi:10.1136/bmj.b2307.
7. Morrison AS. *Screening in Chronic Disease*. 1st ed. New York, NY: Oxford University Press; 1985.

8. Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer*. 1977;39(suppl 6):2772–2782.
9. Baker SG, Kramer BS, Prorok PC. Early reporting for cancer screening trials. *J Med Screen*. 2008;15(3): 122–129.
10. Miettinen OS, Henschke CI, Pasmantier MW, et al. Mammographic screening: no reliable supporting evidence? (http://image.thelancet.com/extras/1093web.pdf). (Accessed July 6, 2005).
11. Miettinen OS, Henschke CI, Pasmantier MW, et al. Mammographic screening: no reliable supporting evidence? *Lancet*. 2002;359(9304):404–405.
12. Hanley JA. Mortality reductions produced by sustained prostate cancer screening have been underestimated. *J Med Screen*. 2010;17(3):147–151.
13. Labrie F, Candas B, Cusan L, et al. Screening decreases prostate cancer mortality: 11-year follow-up of the 1988 Quebec prospective randomized controlled trial. *Prostate*. 2004;59(3):311–318.
14. Sandblom G, Varenhorst E, Löfman O, et al. Clinical consequences of screening for prostate cancer: 15 years follow-up of a randomised controlled trial in Sweden. *Eur Urol*. 2004;46(6):717–723; discussion 724.
15. Kjellman A, Akre O, Norming U, et al. 15-year followup of a population based prostate cancer screening study. *J Urol*. 2009;181(4):1615–1621; discussion 1621.
16. Andriole GL, Crawford ED, Grubb RL III, et al. Mortality results from a randomized prostate-cancer screening trial. *N Engl J Med*. 2009;360(13):1310–1319.
17. Hugosson J, Carlsson S, Aus G, et al. Mortality results from the Göteborg randomised population-based prostate-cancer screening trial. *Lancet Oncol*. 2010;11(8):725–732.
18. Kerkhof M, Roobol MJ, Cuzick J, et al. Effect of the correction for noncompliance and contamination on the estimated reduction of metastatic prostate cancer within a randomized screening trial (ERSPC section Rotterdam). *Int J Cancer*. 2010;127(11):2639–2644.
19. Barry MJ. Screening for prostate cancer—the controversy that refuses to die [editorial]. *N Engl J Med*. 2009;360(13): 1351–1354.
20. Boyle P, Brawley OW. Prostate cancer: current evidence weighs against population screening. *CA Cancer J Clin*. 2009;59(4):220–224.
21. Salinger MA. Looking for the keys under the lamppost: insights from economics into standards for unilateral conduct. 2006. Washington, DC: Federal Trade Commission; 2006. (http://www.ftc.gov/speeches/salinger/060731lookingforthekeys.pdf). (Accessed June 20, 2010).
22. Humphrey LL, Helfand M, Chan BKS, et al. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. In: Vincenza Snow, ed. *Screening for Diseases: Prevention in Primary Care*. Philadelphia, PA: American College of Physicians; 2004:77–99.
23. Endo A. The discovery and development of HMG-CoA reductase inhibitors. *J Lipid Res*. 1992;33(11):1569–1582.
24. Tsujita Y, Kuroda M, Tanzawa K, et al. Hypolipidemic effects in dogs of ML-236B, a competitive inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Atherosclerosis*. 1979;32(3):307–313.
25. Caro J, McGregor M. Screening for breast cancer in women aged 40–49 years. Montreal, Quebec, Canada: Agence d'évaluation des technologies et des modes d'intervention en santé [AETMIS, the Québec government agency responsible for health services and technology assessment]; 1993:91. (CETS report no. 22). (http://www.aetmis.gouv.qc.ca/site/download.php?f=503b634ef04a597215ff3dc734d8d84e). (Accessed July 6, 2005).
26. Gøtzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet*. 2000;355(9198): 129–134.
27. Tabár L, Vitak B, Chen HH, et al. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am*. 2000;38(4):625–651.
28. Tabár L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet*. 1985;1(8433):829–832.
29. Kalager M, Zelen M, Langmark F, et al. Effect of screening mammography on breast-cancer mortality in Norway. *N Engl J Med*. 2010;363(13):1203–1210.
30. Pignone M, Rich M, Teutsch SM, et al. Screening for colorectal cancer in adults at average risk: a summary of the evidence for the U.S. Preventive Services Task Force. In: Snow V, ed. *Screening for Diseases: Prevention in Primary Care*. Philadelphia, PA: American College of Physicians; 2004:101–111.
31. Mandel JS, Church TR, Ederer F, et al. Colorectal cancer mortality: effectiveness of biennial screening for fecal occult blood. *J Natl Cancer Inst*. 1999;91(5):434–437.
32. Mandel JS, Church TR, Bond JH, et al. The effect of fecal occult-blood screening on the incidence of colorectal cancer. *N Engl J Med*. 2000;343:1603–1607.
33. Mandel JS, Bond JH, Church TR, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. *N Engl J Med*. 1993;328(19):1365–1371.
34. Hewitson P, Glasziou P, Watson E, et al. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. *Am J Gastroenterol*. 2008;103(6):1541–1549.
35. Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet*. 2010; 375(9726):1624–1633.
36. Selby JV, Friedman GD, Quesenberry CP Jr, et al. A case-control study of screening sigmoidoscopy and mortality from colorectal cancer. *N Engl J Med*. 1992;326(10): 653–657.
37. Gray RH, Kigozi G, Serwadda D, et al. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. *Lancet*. 2007;369(9562):657–666.
38. Review Manager (RevMan) [computer program]. Version 5.1. Copenhagen, Denmark: The Nordic Cochrane Centre, The Cochrane Collaboration; 2011.
39. Djulbegovic M, Beyth RJ, Neuberger MM, et al. Screening for prostate cancer: systematic review and meta-analysis of randomised controlled trials. *BMJ*. 2010;341:c4543. doi:10.1136/bmj.c4543.
40. Law M. What now on screening for prostate cancer? *J Med Screen*. 2009;16(3):109–111.
41. Sussman JB, Hayward RA. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *BMJ*. 2010;340:c2073. doi:10.1136/bmj.c2073.
42. Donovan J, Hamdy F, Neal D, et al. Prostate Testing for Cancer and Treatment (ProtecT) feasibility study. *Health Technol Assess*. 2003;7(14):1–88.
43. Hutchison GB. Evaluation of preventive services. *J Chronic Dis*. 1960;11:497–508.

APPENDIX

Construction of Web Figure 3

For Web Figure 3, hypothetical, but realistic data were used to emphasize the need to examine time-specific cancer mortality rate ratios in order to measure the timing and magnitude of the reductions produced by screening, and to enable readers to appreciate the critical impact of length of follow-up. To generate this Web figure, I first calculated how many men would be alive at each follow-up year if 100,000 men were followed, without screening, for a full 20 years or until death from another cause. The mix of entry ages—ranging from 55 to 69 years—in this hypothetical cohort was chosen to mimic the mean (age 61 years) in the ERSCP. The numbers at risk (i.e., alive) at each year of follow-up were based on the actual mortality experience of the Quebec male population, scaled so that the proportions alive in the first 7 years of follow-up in the hypothetical cohort would match the observed proportions under follow-up in the first 7 years in the control arm of the ERSPC. Of note is that some 66,000 of the 100,000 men would die, most of them from competing causes, within the 20 years.

The first row of squares in Web Figure 3(a) shows the yearly expected numbers of prostate cancer deaths (1,177 in total over the 20 years) of the 66,000 deaths from all causes over this 20-year period of follow-up. As with the numbers of deaths from other causes, these numbers also were based on the population experience in the entire province of Quebec in the early 1990s, with rates scaled to match the actual prostate cancer mortality to date in the control arm of the ERSPC (326 deaths during 785,585 man-years of follow-up of 89,435 men in the ERSPC; 326 × (100,000/89,435) = 364 deaths over 9 years for the hypothetical 100,000 men). Despite the decreasing numbers of men alive at successive follow-up years, the increasing-with-age prostate cancer death rates mean that the numbers of prostate cancer deaths in the cohort would increase each year. Thus, if entry to the cohort was staggered over several calendar years, and the experience was analyzed after an average of 9 years of follow-up, as was the case in the ERSPC, the number of prostate cancer deaths in such a ''nonscreening'' arm ($n = 364$) at the time of this analysis would be substantially less than 9/20ths of the total number ($n = 1,177$) of prostate cancer deaths that would be observed over a full 20 years of follow-up.

The remaining rows of Web Figure 3(a) show the yearly expected numbers of prostate cancer deaths of these 100,000 men under various screening regimens: none; once (as in the ongoing ProtecT trial (42) and in the completed Stockholm trial (15)); and 2, 3, and 4 times, 4 years apart—the last being the target interval in the countries that contributed the majority of ERSPC participants. These hypothetical results are based on a simple conceptual screening model. The model provides hypothetical, but reasonably realistic ''observations'' that enable us to examine mortality data in a screening trial in a time-specific way; the exact form of this screening model does not affect the general message (42).