

CORRESPONDENCE

Re: All-Cause Mortality in Randomized Trials of Cancer Screening

To support their argument that results of cancer screening trials based on disease-specific mortality are unreliable, Black et al. (1) compared the treatment effect measured by disease-specific mortality with the effect measured by all-cause mortality in 12 such trials. They found "major inconsistencies" between the two measures. We disagree with their interpretation and illustrate our reasons with results from the Minnesota study of fecal occult blood testing (2). After 13 years, that study found a 33% lower colorectal cancer mortality in the annually screened group than in the control group. (Note that the number of colorectal cancer deaths per 10 000 person-years in the annual group given in Table 1 of Black et al. (1) is in error. The figure should be $82/18.4160 = 4.5$, not 5.4.) Because colorectal cancer deaths constituted only 3% of deaths from all causes, the expected reduction in all-cause mortality is only 1%, i.e., 3% of 33%. The reduction in all-cause mortality actually observed in the study was 0.0 per 10 000 person-years, with a 95% confidence interval (CI) of -7.6 to 7.6 . The expected 1% reduction in the all-cause mortality rate, corresponding to a decrease of 1.8 per 10 000 person-years, is consistent with this interval. Black et al. considered the treatment effect of 1.2 for disease-specific mortality inconsistent with the 0.0 for all-cause mortality but, in fact, the result 1.2 falls well within the 95% CI of -7.6 to 7.6 for the difference in all-cause mortality. Similar consistencies can be shown for most of the studies cited by Black et al. that were designed for disease-specific outcomes and, as indicated by the large CIs, are underpowered for all-cause analysis.

We agree that, in some cancer screening trials, all-cause mortality may provide assurance against the biases that Black et al. identified. However, a problem with the design of studies using an all-cause mortality end point is the enor-

mous sample size required. For example, with all-cause mortality as the outcome, the aforementioned Minnesota trial (2) with 15 000 subjects per group would have required 20 times as many subjects, or 300 000 per group. Lung cancer trials would have to be about 10 times as large. According to the Nordic Cochrane Centre, a breast cancer screening trial would need 1.2 million women in each group (3), 25–60 times the size of some previous disease-specific studies. Given that the concerns raised may involve only certain cancers in specific populations, rejecting findings across the board based on the possibility of bias is premature.

TIMOTHY R. CHURCH
FRED EDERER
JACK S. MANDEL

REFERENCES

- (1) Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002;94:167–73.
- (2) Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N Engl J Med* 1993;328:1365–71.
- (3) Olsen O, Gotzsche PC. Systematic review of screening for breast cancer with mammography. The Nordic Cochrane Centre, Copenhagen, Oct 20, 2001. [Accessed 05/01/02.] Available from: <http://image.thelancet.com/lancet/extra/fullreport.pdf>.

NOTES

Affiliations of authors: T. Church, School of Public Health, University of Minnesota, Minneapolis; F. Ederer, School of Public Health, University of Minnesota, and The EMMES Corporation, Rockville, MD; J. Mandel, Exponent, Inc., Menlo Park, CA.

Correspondence to: Timothy R. Church, Ph.D., Division of Environmental and Occupational Health, School of Public Health, MMC 807, 420 Delaware St. SE, Minneapolis, MN 55455 (e-mail: trc@cccs.umn.edu).

We agree with Juffs and Tannock that "Screening trials are even more difficult than we thought they were" (1). We would add that the problem of slippery-linkage bias is not unique to screening trials and that adjuvant therapy trials may also be more difficult than we thought they were.

Consider, for example, the question

of the optimum duration of adjuvant androgen deprivation in the treatment of prostate cancer. The Radiation Therapy Oncology Group (RTOG) 99–10 trial is comparing a treatment of 16 weeks with a treatment of 36 weeks for total androgen suppression in men with prostate cancer who are receiving radical radiotherapy. The main end point is disease-specific survival, and the trial is powered to detect a 33% reduction in the hazard rate for death from prostate cancer, with target accrual of 1540 patients. However, there is a real possibility that, in comparison with short-term therapy, long-term androgen deprivation may be associated with an excess mortality from causes other than prostate cancer. This possibility is analogous to the slippery-linkage bias described by Juffs and Tannock and would suggest that overall survival, and not disease-specific survival, should be the main end point of this trial.

Perhaps the best data to support this possibility come from subgroup analysis of the RTOG 92–02 trial (2). This trial recruited over 1500 men with locally advanced prostate cancer who received total androgen suppression for 2 months before and 2 months during radiotherapy to the prostate and pelvis. They were randomly assigned to receive an additional 24 months of the luteinizing hormone-releasing hormone agonist goserelin or to receive no further adjuvant therapy. The initial results show a trend toward an improved 5-year disease-specific survival with adjuvant treatment (92% versus 87%, $P = .07$), with no difference in the 5-year overall survival (78% versus 79%) (2). The subgroup of patients who had a Gleason score of 8–10 (22% of all participants) showed a statistically significant benefit for adjuvant goserelin in terms of disease-specific survival (90% versus 78%, $P = .007$) and of overall survival (80% versus 69%, $P = .02$). The outcome of the remaining 78% of patients with a Gleason score of 7 or less was not presented. However, this outcome may be estimated, because the percent survival for all patients is 0.78 (percent survival for patients with a Gleason score of 7 or less) plus 0.22 (percent survival for patients with a Gleason score of 8–10).

If we take into account the possibility of rounding errors, then the 5-year disease-specific survival shows an absolute difference of 1.7%–4.3% in favor of ad-

juvant goserelin (91.9%–93.2% versus 88.9%–90.2%), but the overall survival shows an absolute detriment of 3.1%–5.6% (76.8%–78.1% versus 81.2%–82.4%). Thus, a 2-year treatment with adjuvant goserelin was associated with an absolute increased risk of 4.8%–9.9% for nonprostate cancer death. The statistical significance of this increased risk cannot be calculated from the available data, but its magnitude is too large to be explained merely by the increased number of patients at risk.

The cause of any excess mortality associated with goserelin therapy is not known, but there is some evidence to suggest an effect on cardiovascular mortality. Low testosterone levels have been associated with a range of risk factors for cardiovascular disease (3), and luteinizing hormone-releasing hormone agonist therapy leads to both increased insulin resistance and arterial stiffness (4).

A 2-year treatment with adjuvant goserelin improves overall survival for men with locally advanced prostate cancer and a Gleason score of 8–10 who undergo radical radiotherapy (2). Trials are warranted in lower risk populations comparing different durations and different methods (e.g., antiandrogen versus androgen deprivation) of adjuvant hormonal therapy. Given the possibility of an adverse effect of androgen deprivation on nonprostate cancer mortality, it is important that the main end point of such trials be overall, and not disease-specific, survival. If the RTOG 99–10 trial were designed with 90% power and a statistical significance level of .05 to detect a 10% reduction in the hazard rate for overall mortality, a total of 7400 patients would be required. Adjuvant therapy trials may be even more difficult than we thought they were.

CHRIS PARKER
DAVID DEARNALEY

REFERENCES

- (1) Juffs HG, Tannock IF. Screening trials are even more difficult than we thought they were. *J Natl Cancer Inst* 2002;94:156–7.
- (2) Hanks G, Lu J, Machtay M, Venkatesan V, Pinover W, Byhardt R, et al. Proc of Am Soc Ther Rad Oncol (ASTRO), RTOG Protocol 92–02: a phase III trial of the use of long term total androgen suppression following neoadjuvant hormonal cytotoreduction and radiotherapy in locally advanced carcinoma of the prostate

[abstract 4]. *Int J Radiat Oncol Biol Phys* 2000;48:112.

- (3) Simon D, Charles MA, Nahoul K, Orssaud G, Kremiski J, Hully V, et al. Association between plasma total testosterone and cardiovascular risk factors in healthy adult men: The Telecom Study. *J Clin Endocrinol Metab* 1997;82:682–5.
- (4) Smith JC, Bennett S, Evans LM, Kynaston HG, Parmar M, Mason MD, et al. The effects of induced hypogonadism on arterial stiffness, body composition, and metabolic parameters in males with prostate cancer. *J Clin Endocrinol Metab* 2001;86:4261–7.

NOTES

Affiliation of authors: C. Parker, D. Dearnaley, Academic Radiotherapy Department, Royal Marsden Hospital, Surrey, U.K.

Correspondence to: Chris Parker, M.D., Academic Radiotherapy Department, Royal Marsden Hospital, Downs Rd., Sutton, Surrey, U.K. SM2 5PT (e-mail: cparker@icr.ac.uk).

Misclassification biases can affect cause-specific mortality, as pointed out by Black et al. (1). One should not infer from their Table 1, however, that such biases are operating in these screening trials. There is simply too much noise to draw any inference about bias in “direction” or “magnitude,” as indicated by the confidence intervals in Table 1. The editorial (2) also makes much of the “major inconsistencies” between results for cause-specific and all-cause mortality. We, therefore, report the results of a simulation in which no biases are operating to demonstrate that these inconsistencies can be easily explained by chance effects, not by bias.

The simulations were based on the data in Table 1 from Black et al. (1) and the references therein. From the references, we obtained the numbers of deaths D_s , D_c , D_{ts} , and D_{tc} observed in each trial. These correspond to screened (D_s), control cause-specific (D_c), screened (D_{ts}), and control all-cause (D_{tc}) deaths, respectively. From the corresponding rates in the same table (1), R_s , R_c , R_{ts} , and R_{tc} , we calculated the person-years (PY) ($\times 10^4$) from the equation $PY_s = D_s/R_{ts}$ and $PY_c = D_{tc}/R_{tc}$. To eliminate all bias in our simulations, we then set $R_{ts} = R_c + R_s - R_c$. We defined the expected Poisson counts for cause-specific and other deaths as $E_s = R_s \times PY_s$, $E_c = R_c \times PY_c$, $E_{other,s} = (R_{ts} - R_s) \times PY_s$, and $E_{other,c} = (R_{tc} - R_c) \times PY_c$. Proceeding in this way for each of the 11 trials with confidence

intervals in Table 1 from Black et al. (1), we generated four independent Poisson death counts with the expectations above, computed the estimated cause-specific and all-cause rates by dividing deaths by PY_s or PY_c as appropriate, and determined how many “inconsistencies” there were in direction or magnitude by using the criteria described by Black et al. (1).

In 10000 such simulations, the average number of inconsistencies of direction was 3.61 with a standard deviation of 1.55. The average number of inconsistencies of magnitude was 1.64 with a standard deviation of 1.09. Thus, the numbers of inconsistencies of direction (five of 11) and magnitude (two of 11) reported by Black et al. are entirely consistent with chance. In fact, 26.84% of the simulated trials had five or more inconsistencies in direction, and 52.00% had two or more inconsistencies in magnitude. One does not need to invoke “sticky diagnosis” bias or “slippery linkage” bias or any other bias to explain the results in Table 1.

Estimates of a difference in all-cause mortality rates are much less precise than in cause-specific mortality rates. To get the same precision, the all-cause mortality study would need to be larger (or longer) by a factor equal to the ratio of the sum of the screened and control all-cause mortality rates to the sum of the corresponding cause-specific mortality rates. For the Swedish Two-County Study, the all-cause mortality study would need to be 37.6 times larger (or longer). Clearly, studies of all-cause mortality that are sufficiently large to have the required precision would not be feasible in many situations.

MITCHELL H. GAIL
HORMUZD A. KATKI

REFERENCES

- (1) Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002;94:167–73.
- (2) Juffs HG, Tannock IF. Screening trials are even more difficult than we thought they were. *J Natl Cancer Inst* 2002;94:156–7.

NOTES

Affiliation of authors: Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

Correspondence to: Mitchell H. Gail, M.D., Ph.D., National Cancer Institute, 6120 Executive Blvd., Rm. 8032, Bethesda, MD 20892–7244 (e-mail: gailm@exchange.nih.gov).

Black et al. (1), in their article entitled "All-Cause Mortality in Randomized Trials of Cancer Screening," argue that the only measure of benefit from a screening intervention is a reduction in "all-cause mortality." All cause mortality may be the best way to avoid bias in the determination of the causes of death, as they explain. It may be appropriate in therapy trials in which the disease being studied is a major cause of death, but the goal is impossible to achieve, for all practical purposes, for screening trials.

There are two essential differences between trials of therapies and trials of screening that make all-cause mortality appropriate in the former but impractical in the latter. Both are consequences of the fact that, in trials of therapies for a terminal disease, patients are eligible only if diseased, whereas in screening trials, both diseased and nondiseased individuals are included. As a result, the number of expected deaths per studied patient is much larger in trials of therapies. Furthermore, a much larger fraction of the deaths from all causes are caused by the specific disease in trials of therapies. Thus, required sample sizes for both disease-specific and all-cause mortality may be similar in therapy trials but very different in screening studies.

Their graphic nicely demonstrates the difference between dying from a single disease, such as breast cancer, and dying from all other causes of death. Rather than supporting the concept that screening efficacy should be judged by its effect on all causes of death, their summary provides clear evidence that basing efficacy on "all cause mortality" in a screening trial is neither feasible nor necessary. Breast cancer accounts for only a small percentage of deaths each year. Consequently, if screening reduces the number of deaths from breast cancer by approximately 25%, there will be little change in the overall per capita death rate from all causes as shown in their Fig. 1. For instance, if there are four incident cancers per 1000 women, and if screening reduces the death rate from breast cancer by 25%, and if breast cancer accounts for 10% of all deaths, then a trial would need a minimum of 1.5 million women in each arm for that reduction to show a significant reduction in all-cause mortality. Because

breast cancer accounts for far fewer than 10% of all deaths each year, a trial would have to involve many more than 3 million women.

What the authors have shown is nothing more than that the larger statistical fluctuation in all-cause mortality can mask a real benefit that would be apparent if the benefit is evaluated for breast cancer mortality alone. It is not appropriate to dismiss the fact that screening reduces the rate of death from breast cancer simply because deaths from all causes do not appear to be influenced.

Most women would prefer to not die from breast cancer. The screening trials show that the likelihood of dying from breast cancer can be reduced by earlier detection. There is no good evidence that the blinded assignment of the causes of death in the trials was biased. Even if there were cardiovascular deaths from irradiation that were not counted in the trials, these deaths would have occurred in fewer than 5% of women irradiated with the old "hockey-stick" fields (2,3) and would be virtually the same for screen-detected as well as control women with breast cancer. Decreasing breast cancer deaths will clearly translate into fewer total deaths in a given year, but proving this decrease in deaths is impossible because the trials would have to be so large that they could not be performed.

DANIEL B. KOPANS
ELKAN HALPERN

REFERENCES

- (1) Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002;94:167-73.
- (2) Cuzick J, Stewart H, Rutqvist L, Houghton J, Edwards R, Redmond C, et al. Cause-specific mortality in long-term survivors of breast cancer who participated in trials of radiotherapy. *J Clin Oncol* 1994;12:447-53.
- (3) Favorable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet* 2000;355:1757-70.

NOTES

Affiliation of authors: D. B. Kopans, E. Halpern, Massachusetts General Hospital/Harvard Medical School, Department of Radiology, Boston, MA.

Correspondence to: Daniel Kopans, M.D., Massachusetts General Hospital, Department of Radiology, Ambulatory Care Center, 15 Parkman St., 2nd Floor, Rm. 219, Boston, MA 02114 (e-mail: kopans.daniel@mgh.harvard.edu).

In their interesting and provocative review of 12 large randomized cancer screening trials, Black et al. (1) draw attention to the fact that the direction of the effect of screening on cancer-specific mortality is not consistent with the direction of the effect of screening on all-cause mortality in many of the trials. The authors point out that in "five of the 12 trials, differences in the two mortality rates went in opposite directions," and they use this seemingly disturbing observation to suggest that bias in assessing cause of death is responsible for this lack of concordance.

We do not dispute that there are ambiguities inherent in the ascertainment of cause of death, nor do we disagree that exposure to screening may potentially bias the ascertainment process. Yet the evidence presented by the authors (lack of concordance) is entirely predictable on the basis of random variation in the absence of any bias. Because these screening trials were all conducted on healthy populations, occurrences of death from causes other than the cancer under investigation vastly outnumber the cancer-specific deaths. Therefore, relatively small random fluctuations in the overall death rates easily swamp the differences in cancer-specific mortality.

We have re-analyzed the data by using the death frequencies derived from the same source publications cited by Black et al., with the exception of the Health Insurance Plan (HIP) trial, for which the numbers were extracted from the Cochrane Library (Table 1). Consider, for example, the Funen study (2) of colorectal cancer screening. In this study, there were 454 deaths attributed to colorectal cancer, and the screened group experienced 44 fewer deaths than the control group. There were 12077 deaths attributed to other causes, and these deaths should be randomly distributed between the two groups in the absence of bias. In the paradigm described by Black et al., discordance will occur if the number of deaths from causes other than colorectal cancer in the screened group exceeds by more than 44 the number observed in the control group. At the outset, this event is not unlikely. In fact, given the disparity of 44 deaths attributed to colorectal cancer, the approximate probability that discordance

Table 1. Concordance probabilities for cancer screening trials

Trial	No. of cancer-specific deaths		No. of all-cause deaths		Discordant	Probability of discordance
	Screened	Control	Screened	Control		
Breast						
HIP*	218	262	2062	2116	No	0.23
Swedish Two-County	160	167	7102	5085	Yes	0.21†
Malmo	63	66	1777	1809	No‡	0.48
Gothenburg	18	40	409	506	No	0.29†
Edinburgh	68	76	1274	1490	No	0.40†
Canadian 1	29	18	159	156	No	0.26
Canadian 2	88	90	734§	690§	Yes	0.48
Colorectal						
Minnesota	199	121	6757	3340	Yes	0.37†
Nottingham	360	420	12 642	12 515	Yes	0.35
Funen	205	249	6228	6303	No	0.34
Lung						
Czechoslovakia	80	61	465	403	No	0.24
Mayo Lung Project	337	303	2493	2445	No	0.30

*Data for this study were obtained from the Cochrane Library (mortality at 13 years follow-up). HIP = Health Insurance Plan.

†Formula was modified from text to account for unbalanced randomization: 77 080 subjects in screened group versus 55 985 in control group (Swedish Two-County); 138 402 person-years (PY) in screened versus 168 025 PY in control (Gothenburg); 157 946 PY in screened versus 147 854 PY in control (Edinburgh); 368 094 PY in screened versus 181 966 PY in control (Minnesota).

‡Although Black et al. classified this study as discordant, in fact there were both fewer cancer-specific deaths and fewer all-cause deaths in the screened group.

§All-cause deaths through 1993 versus June 1996 for cancer-specific deaths.

would occur in this trial by chance alone is 34%, as determined by the tail area of the normal score corresponding to $z = d/t^{1/2}$ (where d = the difference in the number of cancer-specific deaths and t = the total number of deaths attributed to other causes). The corresponding probabilities of discordance for the other studies range from 21% to 48%. Viewed together, the average probability of discordance in these studies was 33%, so in a sample of 12 studies, such as those in Black et al., we would expect to observe about four discordant studies by chance alone.

These results confirm the dilemma that has always faced investigators when designing cancer-screening trials. Although many methodologists would argue that all-cause mortality is ultimately the end point of interest, cancer-specific mortality has been used because of its much greater statistical power. An important measure of the utility of a surrogate end point of this nature is the probability of concordance with the real end point (3). The study by Black et al. shows the unfortunate fact that this probability is not especially high for cancer screening trials.

COLIN B. BEGG
PETER B. BACH

REFERENCES

- (1) Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002;94:167-73.
- (2) Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet* 1996;348:1467-71.
- (3) Begg CB, Leung DH. On the use of surrogate endpoints in randomized trials (with discussion). *J R Stat Soc [Ser A]* 2000;163:15-28.

NOTES

Affiliation of authors: C. B. Begg, P. B. Bach, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY.

Correspondence to: Colin B. Begg, Ph.D., Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021 (e-mail: beggc@mskcc.org).

In their provocative article on the choice of outcomes in randomized trials of cancer screening, Black et al. (1) suggest that for a given trial (or series of trials), disease-specific mortality “should only be interpreted in conjunction with all-cause mortality.” As a means of identifying possible problems in randomization or the completeness of ascertainment of outcome events, we

agree with the authors that it is appropriate to look for differences in all-cause mortality that clearly exceed the plausible impact of screening.

However, we disagree with their conclusion that “a reduction in disease-specific mortality should not be cited as strong evidence of efficacy when the all-cause mortality is the same or higher in the screened group.” Our concern stems from the fact that in most randomized screening trials, all-cause mortality will not differ to a statistically significant extent whether the screening modality does or does not lead to life-saving treatment in some persons. Moreover, the probability of “inconsistency” between the comparison based on all-cause mortality rates and that based on cause-specific mortality rates may be quite high because of chance alone. As an example, consider the Minnesota trial of screening for fecal occult blood, cited by Black et al. In that trial, the all-cause mortality rate was the same in both arms—183.6 deaths per 10 000 person-years—while the cancer mortality rates were 6.6 in the control group versus 5.4 in the screened group. Imagine that a hypothetical new trial is conducted to compare mortality from both all causes and colorectal cancer between two groups of equal size, each of which involves the same number of person-years of follow-up as in the control arm of the Minnesota trial. Moreover, suppose that the true effect of screening on colorectal cancer mortality is to reduce colon cancer mortality in the screened group by 50%, whereas screening has no effect on mortality from other causes. The expected all-cause mortality rates would be 183.6 in the control arm and $183.6 - (6.6 \times 0.5) = 180.3$ in the screened arm. According to the formula given by Rosner (2), such a trial would have only 11% power to detect such a difference in all-cause mortality by use of a two-sided test at the .05 level. Because of sampling variability, there would also be a 31% chance that all-cause mortality would actually be *greater* in the screened group.

Black et al. assert that “increasing the rigor of the death-review process might help to reduce the effects” of the biases in cause-of-death attribution that they have identified. We agree, and we suspect that, although errors may remain in assessment of cause of death after such a review, the magnitude of the bias pro-

duced by those errors will, in most instances, be relatively small. Accepting the possibility of a small bias in cause-specific mortality is, to us, preferable to relying on the presence of a difference in all-cause mortality before concluding that a screening intervention has prevented some deaths. Because it is not generally feasible to do studies that are large enough to reliably document the impact of screening on all-cause mortality, we fear that a number of truly effective cancer screening tests will incorrectly be deemed ineffective if we give undue emphasis to this parameter.

NOEL S. WEISS
THOMAS D. KOEPESELL

REFERENCES

- (1) Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002;94:167–73.
- (2) Rosner B. *Fundamentals of biostatistics*. 4th ed. Belmont (CA): Duxbury Press; 1995. p. 603.

NOTES

Affiliation of authors: N. S. Weiss, T. D. Koepsell, University of Washington, Seattle.

Correspondence to: Noel S. Weiss, M.D., Dr.P.H., University of Washington, Box 357236, Seattle, WA 98195 (e-mail: nweiss@u.washington.edu).

RESPONSE

Begg and Bach, Church et al., Gail and Katki, Kopans and Halpern, and Weiss and Koepsell point out that the inconsistencies between the disease-specific and all-cause mortality rates in Table 1 of our article could be due to chance alone. We agree (as we acknowledged in our discussion) and appreciate the estimates for the probabilities of discordance provided by Begg and Bach, Gail and Katki, and Weiss and Koepsell. Nevertheless, we provided Table 1 to show that these inconsistencies exist—regardless of cause—and to show that the deaths from the target cancers are only small proportions of all deaths (3%–16%). We suspect that many individuals contemplating screening and their referring clinicians would be surprised that these proportions are so low and that there is not even a trend toward a decrease in all-cause mortality in the screening arms (higher in six trials, lower in five, and the same in one). These facts are certainly not conveyed

in the promotional materials for cancer screening that stress saving lives.

Incidentally, Church et al. state that the death rate from colon cancer was 4.5 per 10 000 person-years in the screening arm of the Minnesota trial. However, there were actually two screening arms (and one control arm) in that trial. While the colon cancer death rate was 4.5 per 10 000 person-years in the arm screened annually, it was 6.4 per 10 000 person-years in the arm screened biennially. We did not think that inclusion of only one of the screening arms—the one with the much better outcomes—would fairly represent screening in that trial. Therefore, we combined the number of colon cancer deaths (82 and 117), all deaths (3361 and 3396), and person-years of follow-up (184, 160, and 183 934) in the two screening arms to calculate the mortality rates for screening that are shown in Table 1.

Returning to the issue of the inconsistencies between the disease-specific and all-cause mortality rates in Table 1, that they can be explained by chance alone does not mean that chance is the only, or even the most important, explanation. Clinicians and the public health community must also consider alternative explanations. We devoted much of our article to discussing the plausibility of bias in the classification of death as well as design flaws that could cause inconsistencies between the mortality rates. Although we cited some empirical evidence concerning misclassification of death, we failed to cite one particularly relevant article by Brown et al. (1). These investigators examined deaths in patients diagnosed with cancer and found that the overall noncancer death rate was 1.37 times that expected from U.S. age- and sex-specific mortality data ($P < .001$). Brown et al. also found that most of the excess noncancer deaths occurred shortly after diagnosis, and they concluded that a large proportion of these deaths were probably due to cancer treatment. Parker and Dearnaley provide further evidence that cancer treatments can be associated with excess noncancer mortality and that the potential for net harm is greatest for those with early, low-grade disease—the very individuals most likely to be identified by screening.

Weiss and Koepsell think that a statistically significant reduction in all-cause mortality is too stringent a re-

quirement for the determination of screening efficacy. We agree, and we think that a trend in the right direction along with a statistically significant reduction in disease-specific mortality may be sufficient. However, we do not think that a randomized trial showing an increase in all-cause mortality should ever be cited as “strong evidence” of efficacy, regardless of the reduction in disease-specific mortality. We recognize that the probability of an increase in all-cause mortality from chance alone can be high when the disease-specific mortality is proportionally very low. In the example described by Weiss and Koepsell, in which colon cancer causes only 3.6% of all deaths, the probability of an increase in all-cause mortality is 31% if screening reduces colon cancer mortality by 50% and causes no other deaths. However, when the disease-specific mortality is proportionally very low, it is also true that only a very slight increase in noncancer mortality is required to offset a reduction in cancer mortality. In this colon cancer example, if screening and the subsequent diagnostic evaluation and treatment increase the noncancer mortality by as little as 2%, then screening would cause more deaths than it prevents, even if it does reduce colon cancer mortality by 50%. Thus, even when there is a statistically significant reduction in disease-specific mortality, we do not think the case for screening should be closed when the all-cause mortality is higher in the screened group.

With regard to the appropriate burden of proof, a conservative statistical significance level of 5% is conventionally used in medicine to avoid the acceptance of a new treatment that is not effective. It would be ironic if this 5% significance level were reversed to avoid the rejection of a new screening test that may cause more deaths than it prevents. Furthermore, it is generally agreed that the level of evidence for effectiveness should be especially high for screening because it “converts some ostensibly healthy individuals into patients” (2). (We don’t think most individuals considering screening would be reassured by the argument that the observed increase in all-cause mortality could be dismissed as chance [$P > .05$]).

In conclusion, we stand by our recommendation that all-cause mortality should always be reported and considered

in conjunction with disease-specific mortality. Disregarding the vast majority of deaths that occur in a randomized trial of screening for the sake of statistical power simply hides an important uncertainty. Establishing the net effect of screening healthy people—only a few of whom can be helped, some of whom will be harmed, and most of whom will experience little effect—will often exceed the limits of medical science. Thus, there is all the more reason for full disclosure of both what is known and what is unknown about screening for informed decision making.

REFERENCES

- (1) Brown BW, Brauner C, Minnotte MC. Non-cancer deaths in white adult cancer patients. *J Natl Cancer Inst* 1993;85:979–87.
- (2) Prorok PC, Kramer BS, Gohagan JK. Screening theory and study design: the basics. In: Kramer BS, Gohagan JK, Prorok PC, editors. *Cancer screening: theory and practice*. New York (NY): Marcel Dekker; 1999. p. 29–53.

WILLIAM C. BLACK
DAVID A. HAGGSTROM
H. GILBERT WELCH

NOTES

Affiliations of authors: W. C. Black, Department of Radiology, Dartmouth-Hitchcock Medical Center, Lebanon, NH, and Center for the Evaluative Clinical Sciences, Department of Community and Family Medicine, Dartmouth Medical School, Hanover, NH; D. A. Haggstrom, Department of Medicine, Medical College of Wisconsin, Milwaukee; H. G. Welch, Department of Medicine, Dartmouth-Hitchcock Medical Center, Center for the Evaluative Clinical Sciences, Department of Community and Family Medicine, Dartmouth Medical School, and VA Outcomes Group, Veterans Affairs Hospital, White River Junction, VT.

Correspondence to: William C. Black, M.D., Department of Radiology, Dartmouth-Hitchcock Medical Center, 1 Medical Center Dr., Lebanon, NH 03756 (e-mail: william.black@Hitchcock.org).