BMC Medical Research Methodology

BioMed Central

Research article

Statistical issues in randomized trials of cancer screening

Stuart G Baker*[1], Barnett S Kramer[2] and Philip C Prorok[1]

Address: [1]Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, U.S.A and [2]Office of Disease Prevention, National Institutes of Health, Bethesda, MD, U.S.A

E-mail: Stuart G Baker* - sb16i@nih.gov; Barnett S Kramer - kramerb@od.nih.gov; Philip C Prorok - pp2g@nih.gov

*Corresponding author

Published: 19 September 2002

BMC Medical Research Methodology 2002, **2**:11

Received: 28 June 2002

Accepted: 19 September 2002

This article is available from: http://www.biomedcentral.com/1471-2288/2/11

Abstract

Background: The evaluation of randomized trials for cancer screening involves special statistical considerations not found in therapeutic trials. Although some of these issues have been discussed previously, we present important recent and new methodologies.

Methods: Our emphasis is on simple approaches.

Results: We make the following recommendations:

(1) *Use death from cancer as the primary endpoint, but review death records carefully and report all causes of death*

(2) *Use a simple "causal" estimate to adjust for nonattendance and contamination occurring immediately after randomization*

(3) *Use a simple adaptive estimate to adjust for dilution in follow-up after the last screen*

Conclusion: The proposed guidelines combine recent methodological work on screening endpoints and noncompliance/contamination with a new adaptive method to adjust for dilution in a study where follow-up continues after the last screen. These guidelines ensure good practice in the design and analysis of randomized trials of cancer screening.

Background

The evaluation of randomized trials of cancer screening involves special statistical considerations. Although some of these considerations have been previously discussed [1,2], our emphasis is on recent and new methodologies that are easy to implement. Throughout the article, when we refer to cancer we mean the target cancer of the screening test.

To better appreciate some of the issues, we review common biases associated with a naïve analysis of cancer screening data. These biases arise when comparing survival after cancer detection between screen-detected and clinically detected cancer cases. To better explain these biases we introduce a novel analogy.

Lead-time bias

Detection of an asymptomatic cancer by screening starts the clock at a younger age so the survival time from screen detection is longer than the survival time from clinical detection, even if screening does not change the age of death. As an analogy, imagine waiting at a bus stop C for a bus traveling north to destination D. Suppose you walk south and board the same bus at stop B prior to its arrival at C.

Although the bus ride from B to D is longer than from C to D (by the lead time from B to C), the arrival time at D is unchanged. You have simply spent more of your life on the bus. The travel time from B to D is a lead-time-biased estimate of the travel time from C to D.

Length-bias

Screening preferentially detects slower growing cancers because there is a longer period of time (hence the name length-bias) when such cancers could be found on screening. If slower growing cancers have a different prognosis than faster growing cancers, the estimated survival time after diagnosis will be subject to length-bias. Continuing with the analogy, suppose there are two types of buses: slow local buses that frequently stop at B, and fast express buses that rarely stop at B. Because a bus boarded at B is most likely local, the average time it takes to travel from C to D (thus the lead time has been subtracted) will be a length-biased estimate of the averages of the local and express travel times from C to D.

Overdiagnosis bias

Screening may detect cancers that would never surface clinically or be diagnosed in the absence of screening. Continuing with the analogy, suppose some buses stop at B but not D. Overdiagnosis bias arises when counting all buses stopping at B as going to D.

Selection bias

The type of subject who receives screening may differ from other subjects in ways that are related to survival times. Continuing with the analogy, suppose there is only one type of bus (so there is no length-bias or overdiagnosis bias), but you only board it at stop B in the morning when the traffic is heaviest. The time it takes the bus boarded at B to travel from C to D (so lead-time has been removed) is a selection-biased estimate of the average, over the entire day, of the travel times from C to D.

A randomized trial with an endpoint of death (typically measured in each trial arm as a death rate among all participants) avoids these biases. Lead-time bias is avoided by setting the time of randomization instead of the time of cancer detection as the zero time. Length-bias and overdiagnosis bias are avoided because the comparison is between randomized groups not between screen-detected and clinically detected cancer cases. The use of a mortality endpoint also avoids lead-time bias, length-bias, and overdiagnosis bias that would arise with an endpoint based on characteristics of the cancer. For example, suppose that stage were the endpoint of the trial. A screen-detected stage I cancer is likely to have a different prognosis than a clinically detected stage I cancer due to lead-time, length, and overdiagnosis biases. Therefore using stage as endpoint would bias the results.

Selection bias within the trial is avoided because randomization guarantees the same distribution of known and unknown covariates in both groups. Under randomization, imbalances can occur in the empirical distribution of baseline covariates. These imbalances are not generally a concern unless they are extremely large even after adjusting for multiple comparisons. In that case one should investigate if there were any deviation from random treatment assignment that may have affected cancer death rates. It is important that only baseline characteristics be considered in investigating imbalance. Characteristics that could be known only after randomization (e.g. number of cancers diagnosed, stage, age at diagnosis, cure rates of detected cancers) are likely to be biased because the screening could have affected these characteristics and the analysis is no longer "protected" by randomization.

Randomization does not, however, correct for another type of selection bias. Volunteers who participate in clinical trials and who consent to randomization may differ from the general population. They often have better underlying health, an effect known as "healthy volunteer bias." Although we do not discuss this bias further, it should be considered in planning trial size and in trying to generalize trial results to the population- at-large.

Methods

Our emphasis is on simple methods. Although survival analyses from time of randomization (e.g. logrank tests) are sometimes used, we focus on simple estimates based on the cumulative number of cancer deaths. Because cancer death is a rare event in asymptomatic participants in a screening trial, inference based on survival analysis and cumulative number of cancer deaths is similar [2]. We assume the yearly numbers of cancer deaths follow a Poisson distribution, which is appropriate for rare events.

Results

We make three recommendations concerning the design and analysis of a randomized trial of cancer screening.

(1) Use death from cancer as the primary endpoint, but review death records carefully and report all causes of death

The primary endpoint of most cancer screening trials is death from cancer. Recently Black [3] identified two types of biases that can affect the assessment of the cancer death endpoint. Sticky-diagnosis bias arises when deaths from an uncertain cause are more likely to be attributed to cancer if there was a previous diagnosis of cancer, especially if the diagnosis was relatively recent. If there were overdiagnosis, sticky-diagnosis would induce a higher cancer death rate in the intervention group than actually the case. Slippery linkage, the second type of bias, occurs because deaths that are caused or triggered by screening, work-up, or a subsequent therapy (e.g. perforation of the colon and

perhaps cardiovascular deaths) are not attributed to screening.

Using all deaths as an endpoint avoids these biases but leads to prohibitive sample sizes as shown in the following calculations based on a power of 80% and a one-sided type I error of 2.5%.

First consider the design of a randomized trial with a cancer death endpoint. Under the null hypothesis, the probability of cancer death in each group is p. Under the alternative hypothesis the probability of cancer death is p in the control group and p-d in the study group, where d is the probability of cancer death in the control group minus the probability of cancer death in the screened group. For computing sample size, we assume d is positive. Assuming a Poisson distribution for the number of cancer deaths, the sample size (for both groups combined) for a cancer death endpoint is

$$N_{cancer} = 2\ (1.96\ \text{Sqrt}\ [2\ v_{cancerH0}] + .84\ \text{Sqrt}\ [v_{cancerH0} + v_{cancerHA}])^2/d^2,$$

where $v_{cancerH0} = p$ and $v_{cancerHA} = p$-d are the variances for one subject under the null and alternative hypotheses respectively.

Now consider the design of a randomized trial with an all death endpoint. Let k denote the probability of death from causes unrelated to either cancer or screening. Under the null hypothesis the probability of death from all causes is $p + k$ in each group. Under the alternative hypothesis the probability of death from all causes is $p + k$ in the control group and $(p + k)$-$(d$-$e)$ in the screened group, where e is the additional probability of non-cancer deaths due to screening. Therefore d-e is the probability of death from all causes in the control group minus the probability of death from all causes in the screened group. Assuming a binomial distribution for the number of deaths from all causes, the sample size (for both groups combined) for an all death endpoint is

$$N_{all} = 2\ (1.96\ \text{Sqrt}\ [2\ v_{allH0}\ ] + .84\ \text{Sqrt}\ [v_{allH0} + v_{allHA}])^2/(d\text{-}e)^2,$$

where $v_{allH0} = (p + k)(1\text{-}p\text{-}k)$ and $v_{allHA} = (p\text{+}k\text{-}d\text{+}e)\ (1\text{-}p\text{-}k\text{+}d\text{-}e)$ are the variances for one subject under the null and alternative hypotheses respectively.

For purposes of illustration, suppose that p=. 005, k=. 15 (these values are based roughly on data from a colorectal cancer screening trial [4]), and d=. 001. To minimize N_{all}, we set $e = 0$. With these specifications, a study with a cancer death endpoint would require N_{cancer} = 150,000 participants while a study with an all death endpoint would require N_{all} = 4.1 million participants.

For practical considerations, we recommend using cancer death as an endpoint with careful review of the death records to minimize sticky-diagnosis and slippery linkage bias. We also recommend that "cancer" deaths include any non-cancer deaths attributable to screening or treatment for the cancer.

We also recommend that all deaths and their causes be reported. If, after adjusting for multiple comparisons, there is a statistically significant difference between groups in the estimated probability of a particular non-cancer cause of death, the investigators should reexamine the death records to check for potential biases. If there are no potential biases, the investigators will need to consider the possibility that screening or treatment was responsible for the difference.

(2) Use a simple "causal" estimate to adjust for nonattendance and contamination occurring immediately after randomization

Two complications in the analysis of many randomized trials for cancer screening are (a) non-attendance, whereby some subjects randomized to a screening invitation do not attend the screening, and (b) contamination, whereby some subjects randomized to no screening invitation receive screening outside the trial. The standard approach for handling these complications is to fold them into the interpretation of an intent-to-treat estimate. Let p_0 (p_1) denote the cumulative fraction of subjects in the control (intervention) group who died from cancer. The intent-to-treat estimate, $d_{ITT} = p_1\text{-}p_{0,}$ is the estimated effect of randomization to a screening invitation versus no screening invitation. However, in the presence of non-attendance and contamination, the intent-to-treat estimate is a biased estimate of the efficacy of screening, which is the effect of *receiving* screening.

If some reasonable assumptions hold (to be discussed) there is a simple, but not well-known, method for obtaining unbiased estimates of the effect of receiving screening in the presence of non-attendance and contamination. Let f_0 (f_1) denote the fraction of subjects in the control (intervention) group who receive screening, where $f_1 > f_0$. As discussed below, the "causal" estimate is

$$d_{causal} = (p_1 - p_0) / (f_1 - f_0),$$

which is the estimated effect (change in the probability of cancer death) of receiving screening among subjects who would receive screening if randomized to the intervention group but not if randomized to the control group. This estimate is not unique to screening but applies to any trial

in which nonattendance or contamination occurs soon after randomization.

Glaziou et al [5] proposed using d_{causal} to estimate the effect of receiving screening. Baker and Lindeman [6] and Angrist [7] independently proposed a "causal" model for all-or-none compliance in comparative studies that gives rise to this type estimate and sharpens the interpretation. By "causal" we mean a formulation based on potential outcomes, as for example whether or not a subject receives screening *if* randomized to a particular group. See [7] for a more precise definition. For related models applied to cancer screening see also Baker [8], Cuzick [9], and McIntosh [10]. The "causal" model relies on the following two assumptions if estimates are to be unbiased.

Assumption 1

There are three types of subjects: always-takers who would receive screening if randomized to either group, never-takers who would not receive screening if randomized to either group, and compliers who would receive screening if randomized to the intervention group but not the control group. (In other words, no subjects would receive screening if randomized to the control group but not randomized to the intervention group).

Assumption 2

For always-taker and never-takers the probability of cancer death is the same for each treatment group. (In other words, when a control subject switches to screening immediately after randomization, the screening regime is identical to that in intervention group, and when an intervention subject immediately refuses screening, the lack of screening is identical to that in the control group.)

Unfortunately neither of the assumptions is verifiable, but they are reasonable, and therefore have "face" validity. Although the analysis is not by intent-to-treat, it makes use of the randomization to avoid selection bias.

When computing f_0 and f_1, it is important to count only subjects who switch treatment immediately after randomization, so as not to violate Assumption 2. With this modification d_{causal} is unbiased even if additional subjects switch treatment later in the study, as for example, if some subjects are screened initially but refuse subsequent screenings. The effect of later switching is folded into the interpretation. Thus $d_{causal,}$ is the estimated effect of *immediately receiving* screening with the understanding that the effect is likely attenuated from later switching of treatments.

In designing a randomized trial of cancer screening one should adjust the sample size for anticipated non-attendance and contamination. Suppose the anticipated fraction receiving immediate screening is f_0 and f_1 for the control and intervention groups, respectively. As derived by Zelen [11], the adjusted sample size is the sample size if there were full attendance and no contamination divided by $(f_1 - f_0)^2$.

(3) Use a simple adaptive estimate to adjust for dilution following the last screen

In a typical randomized trial of cancer screening, screening is offered for a limited time and subjects are followed after screening has stopped. This leads to a dilution of treatment effect, as will be explained. Consider a special baseline variable B such that B = 1 if (i) the subject would not be detected with cancer *if* screened, (ii) the subject would become a cancer case after the time of the last screen, and (iii) the subject would die from the cancer during the follow-up period. Otherwise B = 0. In other words B = 1 indicates a set of cancer deaths that could not have benefited from screening. We can identify subjects with B = 1 in the screened group but not in the control group. Let D denote the number of subjects with B = 1 in the screened group. By virtue of the randomization, there will be approximately D subjects with B = 1 in the control group. As the length of follow-up after the last screening increases, the amount of dilution D increases, which increases the variance of the estimated treatment difference.

In estimating the relative risk of randomization to screening or no screening, the value of D affects the point estimate because D is added to both the numerator and denominator. But when estimating a difference in treatment effect between the groups, the value of D cancels. Nevertheless, the point estimate of a difference in treatment effect will likely change systematically during follow-up. The reason is that as follow-up increases, the point estimate includes longer-term effects of screening on cancer mortality. For example, suppose that screening reduces cancer mortality up to five years after the last screening. If one used the estimated difference in cancer mortality at the end of a 3-year follow-up period, this estimate would likely be biased relative to the true difference at 5 years. Thus, the longer the longer the follow-up period (up to some point) the less chance for bias due to excluding long-term effects of screening. But as mentioned previously, the longer the follow-up period the greater the dilution. Thus with longer follow-up, there is a variance-bias tradeoff for estimating the difference in cancer mortality.

Because of this variance-bias trade-off, the results of a randomized screening trial vary with the length of follow-up after the last screening. For example, consider data from the Health Insurance Plan of Greater New York (HIP) Study [12] in which approximately 62,000 women were randomized to either no screening or an invitation for

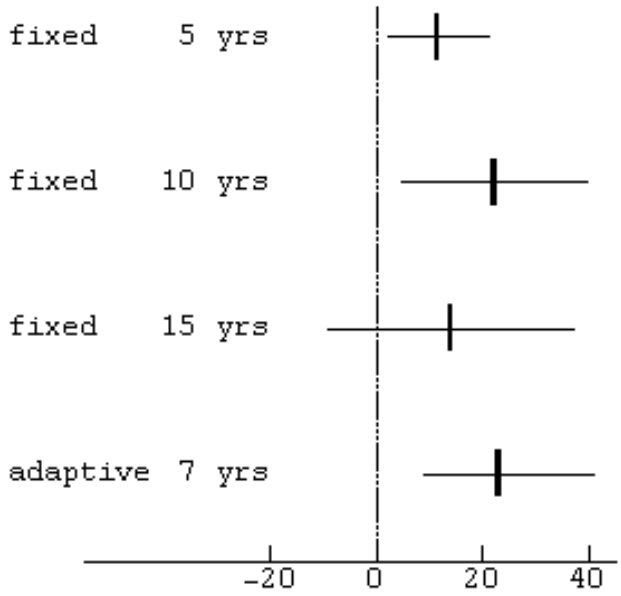

Figure 1
Effect of Follow-up on Estimated Reduction in Breast Cancer Deaths Data are from the HIP Study of breast cancer screening. The plot shows point estimates and 95% confidence intervals for estimated reduction in breast cancer deaths, per 10,000 compliers (participants who would have receive breast cancer screening if offered) due to screening. "Fixed" refers to fixing the follow-up time before examining the data. The estimated reduction is computed as negative $d_{causal}(t)$, where t is the fixed follow-up time. "Adaptive" is the proposed method that bases the follow-up time on the maximum, over time, of a Z-statistic, where confidence intervals are computed by bootstrapping. The estimated reduction is computed as negative $d_{causal}(t^*)$, where t^* is the follow-up time based on the adaptive approach.

four annual breast cancer screenings. We estimated the reduction in the probability of cancer death among compliers at years 5, 10, and 15 since randomization pretending each of these times was fixed in advance of the study (Figure 1). At 5 and 10 years after randomization, the lower bound of the 95% confidence interval was above zero; however this was not the case for 15 years after randomization. A major problem is how to best to analyze these data.

One approach is a limited mortality analysis [2] that counts cancer deaths over the entire follow-up period but only among participants with cancer up to time $t_{catch\text{-}up}$ after randomization. The time $t_{catch\text{-}up}$ is the time when the number of cases in the control group first equals or surpasses (catches-up to) the number of cases in the intervention group. The presumption is that cases surfacing after $t_{catch\text{-}up}$ only dilute the estimated effect. One problem is that $t_{catch\text{-}up}$ does not occur if there is overdiagnosis. A related problem is that $t_{catch\text{-}up}$ might not occur for a very long time, making its calculation impractical. Another problem is that equal numbers of cases in both groups do not guarantee an unbiased test [13].

A second approach is to test if screening reduces cancer mortality rates using a special weighted logrank statistic for survival data [14,15].

A third approach is to select follow-up times based on maximum power given parameter estimates from previous trials and the effect size that one would like to detect [16].

As a fourth approach, we propose a simple adaptive method to compute estimates and confidence intervals for the effect of screening when there is follow-up after the last screen. To the best of our knowledge this method is new to the screening literature. In this analysis, "adaptive" refers to using the data to select the follow-up time, with appropriate adjustment in computing confidence intervals. Let $p_0(t)$ and $p_1(t)$ denote the cumulative fraction of subjects who die from cancer up to time t in the control and intervention groups, respectively. Letting n denote the number of subjects in each group, we define

$$z(t) = (p_0(t) - p_1(t)) / (\text{Sqrt}\,[p_0(t) + p_1(t)]/n),$$

which is the difference between $p_0(t)$ and $p_1(t)$ divided by its standard error, i.e., the z-value associated with a normally distributed random variable. If screening reduces the probability of cancer death, z(t) will generally increase over the time t that screening is offered and perhaps a little longer. However at some point after screening has stopped z(t) will generally decrease over time because $p_0(t)$ and $p_1(t)$ will each increase by roughly the same amount from cases that arose after screening had stopped (i.e. the effect of dilution). See also [16] for a justification of this behavior of z(t) based on modeling natural history in breast cancer screening. We assume that screening does not cause cancer deaths; otherwise it would be possible for z(t) to decrease for reasons other than dilution. This motivates selecting as the follow-up time the time t^* that maximizes z(t) with an estimated effect of

$$d_{causal}(t^*) = (p_1(t^*) - p_0(t^*)) / (f_1 - f_0).$$

We interpret $d_{causal}(t^*)$ as the effect of receiving screening in compliers before dilution attenuates any effects. For $d_{causal}(t^*)$ to be correctly interpretable as an effect of receiving screening, we assume that after perhaps some initial fluctuations $p_1(t) - p_0(t)$ is generally increasing or constant over time until dilution reduces z(t). In other words, although there may be a brief increase in cancer

deaths due to screening soon after the start of the trial, we assume that after screening stops, screening does not start causing more cancer deaths than in the control group. Otherwise we might incorrectly attribute a small difference between $p_1(t) - p_0(t)$ to the effect of dilution when it is due to delayed harms of screening and early treatment.

Computing confidence intervals by ignoring the fact that t* was based on the data represents "cutpoint optimization" [17] and is thus inappropriate. To compute a confidence interval for $d_{causal}(t^*)$ that accounts for the adaptive choice of t*, we use the following bootstrap [18] approach.

For purposes of illustration we applied this method to data in [2] on breast cancer screening from the Health Insurance Plan of Greater of New York (HIP) Study. For each year after randomization we randomly generated a number of cancer deaths in each group based on a Poisson distribution with mean value equal to the observed number of deaths in that year and group. From these randomly generated data we computed t* and $d_{causal}(t^*)$. We repeated this calculation 10,000 times to obtain distributions for t* and $d_{causal}(t^*)$. The mean value of these distributions is the estimate and the lower 2.5 % and upper 97.5% quantiles gives the 95% confidence interval. For t* we obtained an estimate of 7.3 years with a 95% confidence interval of 4 to 13 years. For $d_{causal}(t^*)$, the estimate and 95% confidence interval are shown in Figure 1.

To compute sample size for a randomized trial with follow-up after the last screening, we propose the following approach to account for the adaptive nature of the test statistic. The first step is to create anticipated data with *m* subjects per group under the null and alternative hypotheses. The second step is to treat the anticipated data as observed data and compute bootstrap estimates of the variance. Let $v_{adpativeH0}$ and $v_{adaptiveHA}$ denote the bootstrap estimate of the variance divided by *m* under the null and alternative hypothesis, respectively. In other words $v_{adpativeH0}$ and $v_{adaptiveHA}$ are the bootstrap estimates of variance for one subject. The sample size with cancer death endpoint and adjustment for non-attendance and contamination is

$$N_{adaptive} = 2((1.96\ \mathrm{Sqrt}\,[2\, v_{adpativeH0}] + .84\ \mathrm{Sqrt}\,[v_{adpativeH0} + v_{adaptiveHA}])^2/d^2)/(f_1 - f_0)^2.$$

One other issue in design is the duration of screening. It should be sufficiently long so that any reduction in cancer mortality would be apparent before dilution has an effect.

Discussion

In cancer therapy trials, the standard statistical approach is an intent-to-treat analysis using a non-adaptive statistic with an all death endpoint. Why are we advocating a different approach for cancer screening trials? On a fundamental level, cancer-screening trials differ from therapy trials because of the high amount of "noise" relative to the "signal" of screening effect. This "noise" arises because cancer deaths are rare relative to all deaths, non-attendance and contamination immediately after randomization are common, and discontinuation of screening leads to a dilution of cancer deaths due to cases arising after screening has stopped.

With the proposed analysis, we can reduce the "noise" at the "price" of a few reasonable assumptions. In using a cancer death endpoint with careful review of death records, we assume that deaths caused by screening via unanticipated pathways, such as cardiovascular disease, are correctly attributed to screening. In using the simple "causal" model to adjust for nonattendance and contamination, we assume that (i) a subject who switches treatment immediately after randomization does in fact receive the same treatment as in the other treatment group, and (ii) no subject would receive screening outside the trial if randomized to the control group *and* refuse screening if randomized to the intervention group. In using the adaptive statistic to estimate the effect of screening in a trial with follow-up after the end of screening, we assume that screening does not increase cancer mortality after some point in time.

Even with the proposed method for reducing "noise", the sample sizes for randomized cancer-screening trials are substantial, typically requiring tens of thousands of subjects. Thus randomized screening trials should only be undertaken when there is strong preliminary evidence for a potential benefit of screening that could outweigh attendant harms. In this regard, it is important to have a well-designed strategy for selecting the most promising early detection markers for evaluation in a randomized cancer-screening trial [19–21].

Our focus has been on randomized trials for evaluating the efficacy of cancer screening and the attendant harms. However observational studies have a role particularly when investigating secondary questions involving the effect of age to begin screening, interval between screenings, or small changes in the screening modality. Case-control studies are applicable with special considerations for cancer screening [22]. Periodic Screening Evaluation (PSE) is a method for using data from subjects of various age who receive at least two regularly scheduled screenings to estimate the reduction in cancer mortality from periodic screening over a range of ages [23,24]. The main assumptions of PSE are (1) once a cancer is detectable on screening it would be detectable on later screenings (2) given age, year of birth adds no information for predicting the detection rate on the first screen, (3) no selection bias in

using refusers to estimate survival from detection in the absence of screening. The paired availability design (PAD) is a method for combining data from various before-and-after studies that adjusts for different fraction receiving the intervention in a manner similar to that for nonattendance and contamination [25]. For applications to screening, PAD requires a well-defined geographic region in which screening has been introduced, with little in- or -out- migration and no other changes over time that would affect the endpoint of cancer mortality.

We emphasized estimating the reduction (if any) in cancer deaths due to screening. For a balanced evaluation, one should also estimate the probability of an unnecessary biopsy [23] and other harms attendant to screening and interventions triggered by the screening process.

Conclusion

The proposed guidelines combine recent methodological work on screening endpoints and noncompliance/contamination with a new adaptive method to adjust for dilution in a study where follow-up continues after the last screen. They should greatly help investigators design and analyze randomized trials for the early detection of cancer. Because the assumptions are reasonable, we recommend these guidelines as one of the primary analyses.

Authors' Contributions

SGB wrote an initial draft and BSK and PCP made important improvements. All authors read and approved the final manuscript.

Competing interests

None declared.

Acknowledgements

We thank Ping Hu, Karen Kafadar, and the reviewers for helpful comments.

References

1. Etzioni RD, Connor RJ, Prorok PC, Self SG: **Design and analysis of cancer screening trials.** *Statistical Methods in Medical Research* 1995, **4:**3-17
2. Connor RJ, Prorok PC: **Issues in the mortality analysis of randomized controlled trials of cancer screening.** *Controlled Clinical Trials* 1994, **15:**81-99
3. Black WC, Haggstrom DA, Welch HG: **All-cause mortality in randomized trials of cancer screening.** *Journal of the National Cancer Institute* 2002, **94:**167-173
4. Hardcastle JD, Chamberlain JO, Robinson MH, Moss SM, Amar SS, Balfour TW, James PD, Mangham CM: **Randomised controlled trial of faecal-occult-blood screening for colorectal cancer.** *Lancet* 1996, **348:**1472-1477
5. Glasziou PP: **Meta-analysis adjusting for compliance: The example of screening for breast cancer.** *Journal of Clinical Epidemiology* 1992, 1251-1256
6. Baker SG, Lindeman KS: **The paired availability design: A proposal for evaluating epidural analgesia during labor.** *Statistics in Medicine* 1994, **13:**2269-2278
7. Angrist JD, Imbens GW, Rubin DR: **Identification of causal effects using instrumental variables.** *Journal of the American Statistical Association* 1996, **92:**444-455
8. Baker SG: **Analysis of survival data from a randomized trial with all-or-none compliance; estimating the cost-effectiveness of a cancer screening program.** *Journal of the American Statistical Association* 1998, **93:**929-934
9. Cuzick J, Edward R, Segnan N: **Adjusting for non-compliance and contamination in randomized clinical trials.** *Statistics in Medicine* 1997, **16:**1017-1029
10. McIntosh MW: **Instrumental variables when evaluating screening trials: estimating the benefit of detecting cancer by screening.** *Stat Med* 1999, **19:**2775-2794
11. Zelen M: **A new design for randomized clinical trials.** *New England Journal of Medicine* 1979, **300:**1242-1145
12. Shaprio S, Venet W, Strax P, Venet L: *Periodic Screening for Breast Cancer, The Health Insurance Plan Project and Its Sequelae, 1963–1986, Baltimore, Johns Hopkins University Press* 1988
13. Etzioni R, Self SG: **On the catch-up time method for analyzing cancer screening trials.** *Biometrics* 1995, **51:**31-43
14. Self SG, Etzioni R: **A likelihood ratio test for cancer screening trials.** *Biometrics* 1995, **51:**44-50
15. Buyske S, Fagerstrom R, Ying Z: **A class of weighted log-rank tests for survival data when the event is rare.** *Journal of the American Statistical Association* 2000, **95:**249-258
16. Hu P, Zelen M: **Planning clinical trials to evaluate early detection programmes.** *Biometrka* 1997, **84:**817-829
17. Altman DG, Lausen B, Sauerbrei W, Schumacher M: **Dangers of using "optimal" cutpoints in the evaluation of prognostic factors.** *Journal of the National Cancer Institute* 1994, **86:**829-835
18. Efron B, Gong G: **A leisurely look at the bootstrap, the jackknife, and cross-validation.** *American Statistician* 1983, **37:**36-48
19. Baker SG, Kramer BS, Srivastava S: **Markers for early detection of cancer: Statistical issues for nested case-control studies.** *BMC Medical Research Methodology* 2002, **2:**4 [http://www.biomedcentral.com/1471-2288/2/4/]
20. Baker SG, Tockman MS: **Evaluating serial observations of precancerous lesions for further study as a trigger for early intervention.** *Stat Med* 2002, **21:**2383-2390
21. Pepe MS, Etzioni R, Feng S, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y: **Phases of biomarker development for early detection of cancer.** *Journal of the National Cancer Institute* 2001, **93:**1054-1061
22. Cronin KA, Weed DL, Connor RJ, Prorok PC: **Case-control studies of cancer screening: Theory and practice.** *Journal of the National Cancer Institute* 1998, **90:**498-504
23. Baker SG: **Evaluating the age to begin periodic breast cancer screening using data from a few regularly scheduled screens.** *Biometrics* 1998, **54:**1569-1578
24. Baker SG: **Evaluating periodic cancer screening without a randomized control group: a simplified design and analysis.** *In: Quantitative Methods for the Evaluation of Cancer Screening, (Edited by: Duffy SW, Hill C, Esteve J) London: Edward Arnold Limited,* 2001, 34-41
25. Baker SG, Lindeman KL, Kramer BS: **The paired availability design for historical controls** *BMC Medical Research Methodology* 2001, **1:**9 [http://www.biomedcentral.com/1471-2288/1/9]

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/2/11/prepub