STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:569–581
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2037

Computational methods in medical decision making: to screen or not to screen?

Karen Kafadar[1,*,†] and Philip C. Prorok[2]

[1]*Department of Mathematics, University of Colorado-Denver, Denver, CO 80217-3364, U.S.A.*
[2]*Biometry Research Group, National Cancer Institute, Bethesda, MD 20982-7354, U.S.A.*

SUMMARY

Screening for a disease such as cancer is often regarded as a beneficial and successful strategy for reducing mortality. However, as with any clinical treatment or intervention, benefit cannot be assumed, and screening can entail both costs and harms, so the screening as a 'treatment' must undergo evaluation. An evaluation requires a definition of the treatment 'benefit', design of studies to measure that benefit with as little bias and variance as possible, and the development of methods for estimating the potential benefit. In screening studies, the factors most central to the evaluation are unobservable (e.g. earliest point in time at which disease becomes detectable, or 'preclinical'; time at which disease might have been detected in the absence of screening; test sensitivity). Thus, screening programs should be evaluated on scenarios in which these factors are varied, to ensure the robustness of the estimated benefit under a variety of circumstances. This article describes the importance of computational methods and simulations to assess the benefit of screening programs, particularly for cancer, based on randomized screening trials, with special attention to benefit time, lead time, and bias due to length-biased sampling. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: randomized screening trial; disease progression model; lead time; length-biased sampling; preclinical duration; sojourn time; survival function; discrete event simulation

1. INTRODUCTION

Screening tests are used to detect the presence of preclinical disease in asymptomatic individuals. The underlying assumption in administering these tests is that early diagnosis offers some benefit to the individual, such as better prognosis through earlier treatment, less invasive or aggressive treatment, or reduced chances of mortality. Like any intervention, the administration of screening must be evaluated. The specific nature of the 'benefit' must be quantified, studies must be designed and conducted to measure that benefit with as little bias and variance as possible, and methods for estimating the potential benefit must be developed.

*Correspondence to: Karen Kafadar, Department of Mathematics, University of Colorado-Denver, P.O. Box 173364, Campus Box 170, Denver, CO 80217-3364, U.S.A.
†E-mail: kk@math.cudenver.edu

 Accepted July 2004

Many of these issues have arisen and been discussed in connection with the analysis of randomized treatment clinical trials. In either a screening trial or treatment trial, two conventional endpoints are survival indicator (lived/died) and survival time (time from beginning of trial to end of life or duration of disease). The analysis of screening studies, however, entails additional challenges [1]. The factors most central to the evaluation are unobservable: duration of preclinical disease in the absence of screening (sometimes called 'sojourn time'); time of diagnosis of disease for a screen-detected case had the subject not been screen detected; correlation between sojourn time and clinical duration (time between clinical diagnosis and endpoint); and the possible dependence of screening test sensitivity on age of patient, severity of disease, or number of previous screens. These factors cannot be measured on an individual, so mathematical models of the screening process are often created with numerous unverifiable assumptions [2]. Consequently, the randomized screening trial is considered the most objective and unbiased method for assessing the benefit of screening, most commonly defined in terms of a reduction in population mortality from the disease targeted by screening. Due to these unobservable influences, however, a variety of scenarios should be constructed and evaluated, to ensure the robustness of the estimated benefit under different circumstances.

Screening is the widespread testing of apparently healthy individuals with no clinically apparent symptoms of disease. The encounter is initiated not by the individual, but rather by those who administer the test. The goal of the screening process is to separate the population into two groups: those with a high versus low probability of the given disorder. *The implicit assumption in promoting and conducting a screening program is the potential benefit for early diagnosis*. This benefit may take the form of better prognosis, safer treatment, less invasive procedures, higher 'quality of life', etc. A screening program might be indicated under the following conditions:

- The disease to be identified by the screening test represents a serious public health condition (e.g. large numbers of persons suffering from, or seeking medical treatment for, the condition; e.g. obesity; breast cancer; prostate cancer; colorectal cancer).
- The population targeted by the screening program is well defined (e.g. teenagers; women over 40; men over 60; persons over 55).
- The disease has a clearly recognizable asymptomatic phase which ideally (but not always) correlates with disease (e.g. upper 90 per cent of weight for given age–gender class; unidentifiable mass in breast; elevated levels of prostate specific antigen; polyp in colon).
- The results from a proposed screening test can be measured with high reliability (e.g. body mass index; mammogram or clinical breast exam; blood test measuring prostate specific antigen; colonoscopy).
- Treatment of the disease in its pre-symptomatic stage has a demonstrated benefit (e.g. exercise and/or weight-loss program; radiation, lumpectomy, or mastectomy; prostatectomy; polyp removal).

Under conditions such as those listed above, the potential advantages of screening include:

- improved prognosis;
- higher quality of life due to the need for less radical or invasive treatment;
- reassurance to those with negative screening test results;
- potential health-cost savings due to less aggressive treatment.

However, screening may involve potential disadvantages also:

- cost of screening test;
- longer morbidity (i.e. subject lives longer with knowledge of the disease);
- false reassurance to those with false negative screening test results;
- increased costs (financial, psychological, clinical) for those patients having received false positive test results.

Several of these disadvantages are difficult to measure; e.g. while some authors have concluded from patient surveys that the costs of false positives is minimal [3], in fact the psychological consequences of both the fear of potential (but in fact non-existent) disease and the knowledge of real disease (longer morbidity) may affect the immune system in ways that are difficult to measure. To determine whether the potential benefits outweigh the potential drawbacks, most screening studies focus on measuring 'reduction in mortality' (the ratio of the mortality rates among those offered screening versus those who followed their usual medical care), or 'extended benefit time' (the extra survival time, over and above lead time, or the time by which the diagnosis is advanced due to having been screen-detected). Other considerations in evaluating a screening program include:

- *test sensitivity*, or the probability of obtaining a positive result given that disease is indeed present (for *diagnostic* purposes, sensitivity should be high);
- *test specificity*, or the probability of obtaining a negative result given the absence of disease (for *screening* purposes, specificity should be high, since low specificity leads to many false positives and hence higher costs);
- age of subjects being screened (e.g. screening can pose some hardships, both physical and financial, on older persons);
- invasiveness of the screening procedure (e.g. blood tests are easier to undergo than flexible sigmoidoscopy or mammography);
- disease prevalence (more prevalent disease leads to more cases likely to be screen detected, more lives affected, and higher PPV = positive predicted value = proportion of true positives among those diagnosed as positive).

The evaluation of screening programs is further complicated by the potential for bias in the estimate of the benefit. For example, self-selection bias can affect a non-randomized study such as the Breast Cancer Detection Demonstration Project [4]. When offered screening, persons at varying risks tend to present themselves, depending upon the disease targeted by the screening [5], creating an obvious bias when these results are applied to the general population. Other biases such as those due to case-group analysis, lead time, interval sampling, and overdiagnosis also occur; the first three are discussed below. To avoid or reduce the impact of some of these biases, we focus on *randomized* screening trials, in which study arm participants are offered screening at regular intervals (e.g. annual screens for 3–5 years), and the control arm participants follow their 'usual medical care'. Non-compliance in both arms is inevitable: some participants in the study arm may refuse screening, while some in the control arm may seek screening. Randomization ensures that the participant characteristics are the same in both arms, including those that lead to non-compliance of either type in either arm, arguing for an 'intention to treat' analysis [6]. Examples of randomized screening trials include:

- Health Insurance Plan of New York (HIP) breast cancer screening trial [7];

- the breast cancer screening trial in Sweden [8];
- the Canadian National Breast Screening Study (CNBSS [9–11]);
- the Minnesota colorectal cancer screening trial [12];
- the National Cancer Institute's randomized screening trial for cancers of the prostate, lung, colon/rectum, and ovaries (PLCO [13]).

Such trials allow an unbiased estimate of the reduction in mortality from the cancers under study, in spite of non-compliance (e.g. 10 800, or more than one-third, of the 30 131 study arm women in the HIP trial refused the offer of annual mammography plus clinical breast exam).

Despite these challenges, screening programs and decisions regarding their implementation must be made in the face of:

- varying factors;
- changing population demographics over time (e.g. greater or fewer numbers of persons in specific age or socio-economic groups);
- biases;
- continually evolving and developing methods of screening for the same disease.

We describe the use of computational methods to evaluate randomized screening trials, with particular attention to accurate and precise estimation of lead time and benefit time, despite potential biases and in view of unobserved factors that influence the screening outcome (e.g. the bivariate distribution of preclinical and clinical durations, or test sensitivity). We rely on the disease progression model (Section 2), with the aim of estimating average lead time, average benefit time, reduction in mortality, and the effect of length-biased sampling. Because trials are often expensive to conduct, we recommend simulating a variety of plausible scenarios (e.g. models for the joint distribution of preclinical and clinical durations, or models for test sensitivity), to ensure the robustness of the estimated quantities under a variety of circumstances, in the hopes that the results on the true underlying situation will fall somewhere within those studied by simulation.

In Section 2, we describe the disease progression model, and the issues involved with estimating lead time, benefit time, and the effect of length-biased sampling. Section 3 provides a discrete event simulation as well as some plausible models for the relevant screening parameters. Section 4 describes issues related to, and calculations associated with, the effects of length-biased sampling in screening studies, and Section 5 provides discussion and areas for further work.

2. PARAMETERS OF THE DISEASE PROGRESSION MODEL

The typical disease progression model is shown in Figure 1. The horizontal lines in this figure represent the same time line for both cases (unscreened and screen detected). For both cases, the model suggests two states of disease: preclinical (detectable by a provider-initiated screening exam), and clinical (detectable by a clinical symptom, such as an obvious lump or clinical pain). For a screen-detected case, the period corresponding to the preclinical phase is interrupted by the screen (if the test correctly identifies disease), advancing the point of detection from the usual point of clinical detection, that would have occurred in the absence

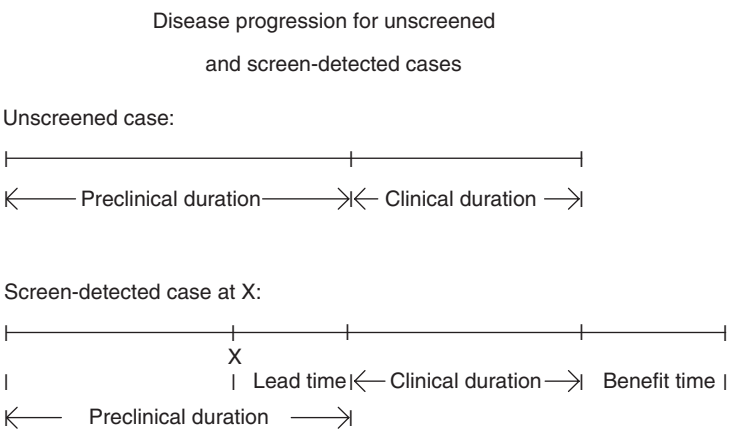

Figure 1. Disease progression model for unscreened and screen-detected cases.

of screening, by an amount known as 'lead time'. The amount of this lead time is important for measures such as 5-year survival rates that are reported commonly in cancer statistics: if the lead time is long enough to change the 'survival time since diagnosis' from, say, 4.8 years to 5.1 years, then the screen-detected case might be viewed as a 'success' in terms of the 5-year survival rate—simply because the screen found the case 0.3 years earlier. However, the screen-detected case may also experience a true benefit, in terms of added survival time denoted in Figure 1 as 'Benefit time'. The two quantities, average lead time and average benefit time among the study arm cases, are important quantities for purposes of evaluating the overall value of a screening program.

Kafadar *et al.* [14] investigated several estimators of average lead time, L, and average benefit time, B, along with estimates of their standard errors. Among them, estimates of B and L, denoted here as $\hat{B}$ and $\hat{L}$, are:

$$\begin{aligned}\hat{B} &= \text{mean benefit time}\\ &= \text{difference in average survival time since start of trial, study—control}\\ \hat{L} &= \text{mean lead time}\\ &= \text{difference in average survival time since diagnosis, control—study}\end{aligned}$$

(see also Reference [7]). To assess the accuracy and precision of these estimators, as well as the validity of the confidence intervals based on the estimated standard errors, we use computer-simulated randomized screening trials, described in the next section.

The phenomenon of lead time creates an important problem that needs to be addressed when analysing randomized screening trials. Because case diagnosis is advanced, on average, by an amount characterized by the average lead time, many more cases will appear sooner in the study arm than in the control arm. When screening ends, those study cases that might have appeared later in the absence of screening have been detected already, creating fewer cases

Table I. Cumulative incidence and mortality in HIP Study (S = Study; C = Control).

	Incidence		Mortality		Rate/100 000		Ratio
Year	S	C	S	C	S	C	S/C
1	79	58	6	2	2.00	0.66	3.04
2	138	124	11	8	1.83	1.32	1.39
3	187	165	17	19	1.90	2.09	0.91
4	249	219	24	38	2.02	3.14	0.64
5	304	295	39	63	2.62	4.18	0.63
6	367	364	58	95	3.26	5.28	0.62
7	426	439	81	124	3.92	5.92	0.66
8	497	490	108	141	4.59	5.92	0.78
9	558	565	128	172	4.85	6.44	0.75
10	617	617	147	172	5.04	6.53	0.77
11	697	680	172	193	5.38	6.70	0.80
12	767	740	198	245	5.70	6.97	0.82
13	826	799	227	271	6.06	7.15	0.85
14	889	874	253	294	6.31	7.24	0.87
15	946	927	285	314	6.67	7.25	0.92

References: [7, Table 5.1; 15, Table 1].

than will occur in the control arm during the same period following the end of screening. To evaluate the benefit of screening on 'comparable' cases, one must identify a point in time during which the survival characteristics of incident cases in both arms can be legitimately compared. In this way, any observed difference between the survival outcomes in the two arms can be attributed only to the screening intervention. Table I, adapted from Shapiro *et al.* [7] and Connor and Prorok [15], shows the effect of analysing the survival experiences of breast cancer cases in the HIP trial accumulated up to different points in time. The mortality ratio, defined as the death rate in study arm divided by death rate in control arm, ranges from a high of 3.04 at the end of year 1 to a low of 0.62 at the end of year 6; at the conclusion of the follow-up phase (year 15), the ratio was 0.92. The average benefit time, which is based on the actual cases rather than on all trial participants (most of whom remain disease-free and thus offer little information about the effect of screening), suffers from a similar difficulty (see Figure 11 in Reference [16]).

The identification of 'comparable' case groups (i.e. of two groups of cases, one from each trial arm, that theoretically have the same disease characteristics in the absence of screening) involves the comparison of the survival experiences between (a) those that could benefit from screening, with (b) those that could *not* benefit from screening. Group (a) consists of study arm participants whose disease is detected during the screening trial (say, within T years, where $T = 3$ for the HIP trial and $T = 5$ for the PLCO trial), as well as persons who refused screening (since, had they not refused, they potentially may have benefitted). Group (b) ideally consists of their 'counterparts' in the control arm—persons with the same demographical, biological, and medical characteristics, *if* they could be identified. The problem is that we do not know when the screen-detected cases would have appeared had they been in the control arm, or when the control cases would have appeared had they been in the study arm (some might be detected successfully by screening, others might test falsely negative).

Some possible solutions to this problem of 'comparable case groups' are:

1. Compare survival experiences for all cases diagnosed only up until the final year of screening (T). As explained above, such a strategy will exclude cases in the control arm that *might* have occurred earlier had they been in the study arm, and thus cases that would be deemed 'comparable' to those in the study arm.
2. Compare survival experiences for all cases diagnosed until the end of the follow-up period, say F. For HIP, where $T=3$ annual screens were offered, $F=15$; for PLCO, where $T=5$, $F=22$. In both trials, most of the cases that would be detected could not have benefitted from screening, which ended 12 and 17 years earlier, respectively. Thus, this strategy would likely underestimate the true benefit of screening. In addition, the potential for an aging population affecting participants in both arms would lead to increasing rates of disease, both from the disease targeted by the screen as well as from competing risks.
3. Compare survival experiences for all cases diagnosed up until some intermediate year C, where $T<C<F$. This is the approach taken by Kafadar and Prorok [16]. If C is too small (too close to T), important control arm cases are omitted, while if C is too large (too close to F), the effect of the screening benefit is severely diluted. Computer-simulated randomized trials (described in Section 3) suggested that $C=T+2\hat{\mu}$, where $\hat{\mu}$ is an estimate of the mean preclinical duration for the disease in question, provides nearly unbiased estimates of both the average lead time and average benefit time, with low variance and nearly valid 95 per cent confidence intervals [16].

For other approaches to the problem of comparable case groups, see References [17, 18].

3. COMPUTER SIMULATED TRIALS

To evaluate the performance of the time point defining the 'comparable case groups', C, as well as of the estimates of the mortality reduction and the average benefit time and average lead time as defined above, we can conduct a discrete event simulation of randomized screening trials. Different methods of simulating randomized screening trials have been suggested in References [19–21]. In this section, we describe one algorithm for a discrete event simulation of a randomized screening trial.

The ingredients for a simulated case of disease include: (a) a preclinical duration; (b) a clinical duration; (c) a screening program (frequency of screening examinations); (d) number of screens (baseline plus T further screens); (e) a value that describes the sensitivity of the screening test (which may depend on characteristics of the case, such as the point during the preclinical phase during which the screen occurred, age of patient, or number of previous negative screens); (f) a potential benefit, if the case is screen detected (which also may depend on characteristics of the case, including those noted for test sensitivity). In the simulation used in Reference [16], quantities (a) and (b), preclinical and clinical durations, were generated from the generalized bivariate gamma distribution, with specified means, variances, and the correlation between them. For case k, this simulated vector will be denoted (d_k, c_k). A simple choice for item (c), the screening frequency, is annual, with T annual screens beyond the baseline (item (d)); typical choices in actual randomized trials are $T=3$ (e.g. HIP) to $T=5$ (e.g. PLCO and the Canadian National Breast Screening Study). While screening test

sensitivity, β, probably depends on various factors, one may set it to a constant initially (e.g. 0.80–0.95), to see if the results are greatly affected by this choice (item (e)). For cases that are screen-detected, one also needs to generate a potential benefit time (item (f)), which also probably depends on several factors (such as on d_k, the preclinical duration).

A series of randomized trials can then be simulated with a model for disease incidence in both the study and control arms. The actual number of simulation runs will depend on the ultimate purposes. For example, for purposes of comparing the accuracy and precision of different estimators of average lead time or average benefit time, 500 runs may be sufficient, whereas, for comparing different confidence interval procedures having nominal 90–95 per cent coverage probabilities, several thousands of runs may be needed. The time at which the preclinical phase begins (a_k), its duration (d_k), and the clinical duration (c_k) define the characteristics of a case in the control arm (the end of the preclinical phase signals the start of the clinical phase). For the 'subjects' assigned to the study arm, a screening pattern is superimposed on the disease history from which the lead time can be calculated. A study case will present as either a screen-detected case, with a stated probability of detection at each screen (which may depend upon the sojourn time, age of the subject, number of previous false screens, etc.), or as an interval case, if screening fails to detect the cancer or if the preclinical phase falls entirely between two screens. A positive lead time may imply a potential benefit; different models can be used to generate a random benefit time. For example, one may model the benefit time to be a function of the preclinical duration: very small for very short durations, reasonably large for intermediate durations, and essentially constant for very long preclinical durations.

Figure 2 shows an algorithm for simulating a case in a randomized screening trial. We start with case k, with simulated values of a_k (beginning of preclinical phase), d_k (preclinical duration, or sojourn time), and c_k (clinical duration). Step 1 initializes the screen number ($j=0$) and the number of missed screens ($nmiss=0$). If the time when the preclinical duration began, a_k, is not less than the time of the jth screen, then advance j until the case could be screen detected (test: $a_k<t_j$; Step 2a). Otherwise (Step 2b), consider the time at which the preclinical duration would have ended, in the absence of screening, which is a_k+d_k (Step 3). If this time point occurs before the jth screen (Step 3a), then the case fell completely between two screens ('interval case'), resulting in zero lead time, $L_k=0$, and survival time since diagnosis, S_k, equal to the clinical duration, c_k. If this time point a_k+d_k exceeds t_j (Step 3b), then the case is still in its preclinical phase and thus potentially detectable by screening, with test sensitivity β. To simulate the success of the detection, we generate a pseudo-random uniform variate u (Step 4); if $u\leqslant\beta$ (Step 4a), then the screen successfully detected disease: lead time is calculated as the difference between the time when the case would have surfaced in the absence of screening (a_k+d_k) and the time at which the screen took place ($t_{j+nmiss}$), and survival time since diagnosis is the sum of the clinical duration, the lead time, and the benefit time, $S_k=c_k+L_k+\text{benefit}_k$. Conversely, if $u>\beta$ (Step 4b), then the screen failed to detect disease (number of missed screens, *nmiss*, is increased by 1, and the algorithm proceeds to Step 5). If the time at which the preclinical duration would have ended in the absence of screening is less than the time of the next screen ($a_k+d_k<t_{j+nmiss}$), then the case becomes an interval case, with lead time $L_k=0$ and survival time since diagnosis $S_k=c_k$ (Step 5a). Otherwise, the case has another opportunity to be screen detected (Step 5b): another uniform variate u is generated and compared with β (return to Step 4). This completes all the possible scenarios for case k, and the simulation moves to the next case.

Notation:

k = case number

a_k = time of start of preclinical duration for case k

d_k = length of preclinical duration for case k

c_k = length of clinical duration for case k

L_k = lead time for a screen-detectable case

S_k = survival time for a case (screened or unscreened)

$benefit_k$ = benefit time for a screen-detectable case

β = screening test sensitivity (e.g., 0.80 or 0.90; may depend on age, preclinical duration, etc.)

j = screen number

t_j = time of j^{th} screen

Step 1: Initialize

- case k
- screen number $j = 0$
- number of screens that failed to detect preclinical disease $nmiss = 0$

Step 2: Preclinical duration starts before screen j?

- (2a) No (case is not yet in its preclinical phase):
  advance to next screen: $j \leftarrow j + 1$; return to Step 2
- (2b) Yes (case is eligible for screen detection): Go to Step 3

Step 3: Time of clinical duration occurs before screen j?

- (3a) Yes (interval case): Lead time $L_k = 0$; Survival time $S_k = c_k$ = clinical duration
- (3b) No (case is eligible for screen detection): Go to Step 4

Step 4: Screen-detection with probability β: Generate $u \sim$ Uniform[0,1]; $u \leq \beta$?

- (4a) Yes (screen-detected case): Lead time $L_k = a_k + d_k - t_{j+nmiss}$;
  Survival time $S_k = c_k + L_k + benefit_k$
- (4b) No (screening test failed to detect disease): $nmiss \leftarrow nmiss + 1$; Go to Step 5

Step 5: Time of clinical duration occurs before $t_{j+nmiss}$?

- (5a) Yes (interval case): Lead time $L_k = 0$; Survival time $S_k = c_k$ = clincial duration
- (5b) No (case is eligible for screen detection): Go to Step 4

Figure 2. Algorithm for a discrete-event simulation of a randomized screening trial.

Table II. One-half fraction of a simulation experiment with three factors: sojourn time (mean+variance); clinical duration (mean+variance); correlation between them.

	Sojourn time		Clinical duration		
Scenario	Mean	Variance	Mean	Variance	Correlation
(A)	2	1	2	1	0.9
(B)	2	1	4	4	0.3
(C)	4	4	2	1	0.3
(D)	4	4	4	4	0.9

A simulation that follows this algorithm has the advantage of flexibility. Various distributions for the preclinical and clinical durations may be used; e.g. the generalized bivariate gamma distribution can model both fast- and slow-growing disease (e.g. pancreatic and prostate cancer) because its parameters can be chosen to have specified characteristics [22]. The simulation can also vary the number of screens, from which one can assess an optimal screening interval for various assumptions about the joint distribution of preclinical and clinical duration and test senstivity, as well as different values for $F =$ number of years of follow-up of participants in the study. Finally, since the simulation involves many factors which can assume different levels, the investigator is wise to implement experimental design strategies such as fractional factorial designs [23]. For example, Table II gives a one-half fraction of a 2^3 factorial design for three factors (mean and variance of the sojourn time; mean and variance of the clinical duration; correlation between them).

This particular simulation program was written in Fortran using subroutines available from *statlib* (via URL address `http://lib.stat.cmu.edu`) to generate Wichman–Hill uniform random deviates [24] and gamma deviates (`ranlibf.uuen`, from B.W. Brown); it is available from the first author upon request.

4. LENGTH-BIASED SAMPLING

An additional problem that arises with screening trials is the effect of *length-biased sampling*. Length-biased sampling arises in this situation because cases with longer sojourn times are more likely to be detected by screening than those which have shorter durations. Zelen [25] noted: 'People who are diagnosed by an early detection program do not constitute a random sample of preclinical cases. Cases found by screening tend to be less advanced... . Women who are found earlier in a detection program tend to be asymptomatic longer, i.e. have slower-growing disease'. Slower-growing disease among screen-detected cases in the HIP study indeed tended to be less advanced: the proportion of cases having negative nodes was 63 per cent among the 132 study cases detected by screening, 47 per cent among the 91 interval cases and 73 cases among those who refused screening in the study arm, and 46 per cent among the 284 cases in the control arm. Bias caused by length-biased sampling cannot be eliminated by the randomization in the trial.

For a program involving just one single screen, Cox and Lewis [26, p. 67] show heuristically that the density of a random variable that is subjected to length-biased sampling, say $f_{Y^*}(\cdot)$

is related to that of the unsampled random variable Y, say $f_Y(\cdot)$, via

$$f_{Y*}(\cdot) = \lim_{m\to\infty} y \cdot m_y \,\mathrm{d}y / \sum_{i=1}^{m} Y_i$$

$$= \lim_{m\to\infty} y \cdot (m_y/m)\,\mathrm{d}y / \sum_{i=1}^{m} Y_i/m = y \cdot f_Y(y)/\mu_Y \tag{1}$$

where Y_i, $i = 1,\ldots,m$ denote the length-biased sampled intervals, $m_y\,\mathrm{d}y$ is the number of the m intervals having lengths between y and $y + \mathrm{d}y$, and μ_y is the mean of the unsampled density of Y. If $Z_1,\ldots,Z_m$ are the clinical durations associated with $Y_1,\ldots,Y_m$ with probability density function (pdf) $f_Z(z)$ and joint pdf $f_{Y,Z}(y,z)$, then, by extension, the joint density $f_{Y^*,Z^*}(y,z)$, when Y only is subject to length-biased sampling, is related to the joint density $f_{Y,Z}(y,z)$, via

$$f_{Y^*,Z^*}(y,z) = f_{Z^*|Y^*}(z \mid y) f_{Y^*}(y) = f_{Z|Y}(z \mid y) \cdot y f_Y(y)/\mu_Y = y \cdot f_{Y,Z}(y,z)/\mu_Y \tag{2}$$

Schotz and Zelen [27, equation (13)] provide an equivalent formula when a four-phase process is subjected to length-biased sampling during the second phase. From (2), the proportional increase in the mean of Y^* over the mean of Y is

$$E(Y^*)/E(Y) = \left[\int_0^\infty \int_0^\infty y(y \cdot f_{Y,Z}(y,z)/\mu_Y)\,\mathrm{d}y\,\mathrm{d}z\right]/\mu_Y = (1 + CV_Y^2) \tag{3}$$

where CV_Y denotes the coefficient of variation (relative standard deviation) of Y, as can be seen also from the earlier derivation (1) using $f_{Y^*}(y)$. Our primary interest, however, is in the effect of the length-biased sampling on the distribution of clinical durations, rather than the preclinical durations. Considering first the mean of Z^*:

$$E(Z^*) = \int_0^\infty \int_0^\infty z f_{Y^*,Z^*}(y,z)\,\mathrm{d}y\,\mathrm{d}z = \int_0^\infty \int_0^\infty z f_{Z^*|Y^*}(z|y) f_{Y^*}(y)\,\mathrm{d}y\,\mathrm{d}z \tag{4}$$

$$= \int_0^\infty \int_0^\infty z f_{Z|Y}(z|y) f_{Y^*}(y)\,\mathrm{d}y\,\mathrm{d}z = \int_0^\infty \int_0^\infty z f_{Y,Z}(y,z)[f_{Y^*}(y)/f_Y(y)]\,\mathrm{d}y\,\mathrm{d}z$$

$$= E[g(Y)Z] \tag{5}$$

where $g(Y) = Y/\mu_Y$. With only one screen, $f_{Y*}(y) = y f_Y(y)/\mu_Y$, and

$$E(Z^*)/E(Z) = E(YZ)/(\mu_Y\mu_Z) = (\rho\sigma_Y\sigma_Z + \mu_Y\mu_Z)/(\mu_Y\mu_Z) = (1 + \rho CV_Y CV_Z) \tag{6}$$

where μ_Y, σ_Y^2, CV_Y (respectively, μ_Z, σ_Z^2, CV_Z) denote the mean, variance, and relative standard deviation of Y (respectively, Z), and ρ is the correlation between Y and Z. Thus, the proportional increase in the average clinical duration among those individuals who are screen-detected, even when the screening has no benefit, depends upon ρ and the relative standard deviations. When ρ is positive (as is commonly assumed for a disease such as cancer [28–30]), this ratio exceeds one. While CV_Z can be estimated from the cases that arise during the control arm of a randomized screening trial, estimates of neither ρ nor CV_Y are

available. For this reason, a variety of plausible models for the joint density between Y and Z should be investigated. For a very simple model, one in which both variables are exponentially distributed (so $CV_Y = CV_Z = 1$), the length-biased sampling of the sojourn times results in an increase of 100ρ per cent in the apparent clinical duration among the screen-detected cases, even when the screening procedure offers no benefit in terms of increased survival since diagnosis (above and beyond what has been taken into account by lead time, or advanced diagnosis of the disease). The periodic screening case is a bit more complex; results will be reported in a forthcoming article.

5. DISCUSSION

The decision to implement screening programs depends upon several well-conceived methods of evaluation. The principal and most objective method is the randomized screening trial, which unfortunately also involves high costs. Consequently, the decision to implement a screening program should also involve some preliminary investigations using computer generated randomized screening trials. Such simulations can involve a variety of conditions for the unobservable quantities such as the joint distribution of the preclinical and clinical durations, test sensivity and specificity, and definitions for comparable case groups, and can be evaluated in terms of the consequent mortality reduction, average lead time, average benefit time, and the added survival time due to the length-biased sampling of the sojourn times. The effect of length-biased sampling can be studied analytically in simple cases, but more general models for the joint distribution of preclinical and clinical durations will require either computational methods for evaluating integrals such as those in Section 4, or discrete event simulation, or both. The application of these methods can be expected to apply in future screening trials such as PLCO, data from which are being collected and will be ready for analysis within a few years. The decision to screen or not to screen will depend upon quantitative methods that increasingly rely on computational methods.

ACKNOWLEDGEMENTS

Part of this research was conducted while Dr. Kafadar was a Guest visitor in the Biometry Research Branch, Division of Cancer Prevention, National Cancer Institute, May 2003.

REFERENCES

1. Prorok PC, Connor RJ, Baker SG. Statistical considerations in cancer screening programs. *Urologic Clinics of North America* 1990; **17**:699–708.
2. Goldberg JD, Wittes JT. The evaluation of medical screening procedures. *The American Statistician* 1981; **35**: 4–11.
3. Cockburn J, Staples M, Hurley SF, De Luise T. Psychological consequences of screening mammography. *Journal of Medical Screening* 1994; **1**:7–12.
4. Beahrs OH, Shapiro S, Smart C. Report of the working group to review the National Cancer Institute—American Cancer Society Breast Cancer Detection Demonstration Projects. *Journal of the National Cancer Institute* 1979; **62**:640–709.
5. Prorok PC, Connor RJ. Screening for the early detection of cancer. *Cancer Investigations* 1986; **4**:225–238.
6. Byar DP, Simon RM, Friedewald WT *et al.* Randomized clinical trials: perspectives on some recent ideas. *New England Journal of Medicine* 1976; **295**:74–80.
7. Shapiro S, Venet W, Strax P, Venet L. *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986*. Johns Hopkins University Press: Baltimore, 1988.

8. Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, Andersson I, Bjurstam N, Fagerberg G, Frisell J, Tabar L, Larsson L. Breast cancer screening with mammography: overview of Swedish randomized trials. *The Lancet* 1993; **341**(8851):973–978.
9. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Cancer Screening Study: 1. Breast cancer detection and death rates among women aged 40–49 years. *Canadian Medical Association Journal* 1992; **147**:1459–1476.
10. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Cancer Screening Study: 2. Breast cancer detection and death rates among women aged 50–59 years. *Canadian Medical Association Journal* 1992; **147**:1477–1488.
11. Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Cancer Screening Study-2: 13-year results of a randomized trial in women aged 50–59 years. *Canadian Medical Association Journal* 1992; **147**:1477–1488.
12. Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, Ederer F. Reducing mortality from colorectal cancer by screening for fecal occult blood. *New England Journal of Medicine* 1963; **328**:1365–1371.
13. Gohagan JK, Levin DL, Sullivan D, Prorok PC. The prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials* 2000; **21**(Suppl. 6):249S–406S.
14. Kafadar K, Prorok PC, Smith PJ. An estimate of the variance of estimators for lead time and screening benefit in randomized cancer screening trials. *Biometrical Journal* 1998; **40**:801–821.
15. Connor RJ, Prorok PC. Issues in the mortality analyses of randomized controlled trials of cancer screening. *Controlled Clinical Trials* 1994; **15**:81–99.
16. Kafadar K, Prorok PC. Alternative definitions of comparable case groups and estimates of lead time and benefit time in randomized cancer screening trials. *Statistics in Medicine* 2003; **21**:83–111.
17. Aron JL, Prorok PC. An analysis of the mortality effect in a breast cancer screening study. *International Journal of Epidemiology* 1986; **15**:36–43.
18. Baker SG, Kramer BS, Prorok PC. Statistical issues in randomized trials of cancer screening. *BMC Medical Research Methodology* 2002; **2**:11; `http://www.biomedcentral.com/1471-2288/2/11`
19. Habbema JDF, Van Oortmarssen GJ, Van Putten DJ. An analysis of survival differences between clinically and screen-detected cancer patients. *Statistics in Medicine* 1983; **2**:183–279.
20. Etzioni R, Self SD. On the catch-up time method for analyzing cancer screening trials. *Biometrics* 1995; **51**: 31–43.
21. Kafadar K, Prorok PC. Computer simulation experiments of randomized screening trials. *Computational Statistics and Data Analysis* 1996; **23**:263–291.
22. Chen J, Prorok PC, Graff KM. An age dependent stochastic model of periodic screening: length bias at a prevalence screen. *Mathematical Biosciences* 1983; **65**:93–123.
23. Box GEP, Hunter WS, Hunter JS. *Statistics for Experimenters*. Wiley: New York, 1974.
24. Wichmann BA, Hill ID. Algorithm 183: an efficient and portable pseudo-random number generator. *Applied Statistics* 1982; **31**:188–190.
25. Zelen M. Theory of early detection of breast cancer in the general population. In *Breast Cancer: Trends in Research and Treatment*, Heuson JC, Mattheiem WH, Rozenweig M (eds). Raven Press: New York, 1976; 287–301.
26. Cox DR, Lewis PAW. *The Statistical Analysis of Discrete Time Events*. Oxford Press: London, 1972.
27. Schotz WE, Zelen M. Effect of length sampling bias on labeled mitotic index waves. *Journal of Theoretical Biology* 1971; **32**:383–404.
28. Fontana RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhm JR, Bernatz PE, Payne WS, Pairolero PC, Bergstalh EJ. Screening for lung cancer: a critique of the Mayo lung project. *Cancer* 1991; **67**:1155–1164.
29. Spratt JS, Meyer JS, Spratt JA. Rates of growth of human solid neoplasms, Part I. *Journal of Surgical Oncology* 1995; **60**:137–146.
30. Spratt JS, Meyer JS, Spratt JA. Rates of growth of human solid neoplasms, Part II. *Journal of Surgical Oncology* 1996; **61**:68–83.

