

*Key words:* cancer screening, sojourn time, length bias, bivariate gamma distribution, logrank test, survival function, catch up time, confidence interval coverage probability

## **The effect of defining comparable case groups on estimates of lead time and benefit time in randomized screening trials**

*Karen Kafadar*

Department of Mathematics \*  
University of Colorado-Denver  
Denver, Colorado 80217-3364

*Philip C. Prorok*

Screening Section, Biometry Branch  
Division of Cancer Prevention & Control  
National Cancer Institute  
Bethesda, Maryland 20892-7354

### *ABSTRACT*

Randomized screening trials provide an effective means of assessing the benefit of screening for disease. Unlike clinical trials, however, where strict eligibility criteria assure the comparability of cases in the arms of a treatment trial, the cases detected by screening may arise from length biased sampling which can bias the estimates of the screening benefit and lead time. We examine, via simulation, several methods for defining comparable case groups and their effects on estimates of average lead time and average benefit time as well as on confidence intervals for these parameters, the mortality reduction, and the logrank test statistic. We propose a new method for defining comparable case groups which uses an estimate of the mean sojourn time. The focus on bias in estimating lead time and benefit time leads to an interpretation of length biased sampling in terms of case survival time.

### **1. Introduction.**

Randomized trials to determine the existence of a benefit of screening for a disease such as cancer involve special considerations that distinguish them from randomized treatment trials. Both types of trials require that the evaluation of the "treatment" be made on cases that are comparable in terms of their progression, former treatments, and history. In the latter type of trial, strict protocols define the entry criteria to ensure that the case histories of all trial participants, both those receiving the standard treatment or placebo as well as those on the "new" treatment, are comparable. In this way, any observed difference

---

\* This work was performed while Dr. Kafadar was a Guest Researcher in the Screening Section, Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute.

between the outcomes in the two arms can be attributed only to the treatments and not to a predisposition towards more favorable disease histories in one arm over the other. With randomized screening trials, however, cases evolve as the trial proceeds. Although the two arms include equally asymptomatic patients at the outset of the trial, the benefit of screening is ascertained on the basis of the cases which arise during the trial. This estimate of the benefit may be biased due to overdiagnosis (cases which may never have surfaced clinically otherwise are diagnosed and treated), lead time (cases are detected earlier, and therefore have longer survival times since diagnosis, even in the absence of any real survival benefit), and length biased sampling (cases with long sojourn times, and thus, perhaps, different types of disease, are more likely to be screen-detected than those with short sojourn times).

Randomized screening trials often are conducted as "stop-screen" designs; i.e., screening is offered to the study group for only a limited time, say from year zero (initial screen) to year  $T$ , with follow-up to year  $F$ . This is the case with the breast cancer screening trial of the Health Insurance Plan of New York (HIP; Shapiro et al. 1988) and the National Cancer Institute's randomized screening trial for cancers of the prostate, lung, colorectal, and ovarian cancers (Gohagen et al. 1994). Additional trials using this design are or have been conducted in Sweden (Nystrom et al. 1993) and Canada (Miller et al. 1992) for breast cancer, and in Minnesota for colorectal cancer (Mandel et al. 1993).

Stop-screen designs require a comparison of case groups which arise in the two arms of the trial (Connor and Prorok 1994, p.87). Although it may seem reasonable at first glance to compare case histories for only those cases detected up until the final year of screening, such a comparison would be biased for two reasons. First, because of imperfect detection (sensitivity  $< 100\%$ ), some cases in the screened arm, preclinical during the screening period, may surface only after screening ends. Second, cases detected by screening are subject to length biased sampling. If such cases tend to have a better prognosis than the average case type, the result is an overestimate of the benefit of screening. Conversely, if the comparison includes many cases diagnosed long after screening ends, beyond the point at which screening could have had any influence, the estimate of a possibly positive benefit will be diluted. An unbiased assessment of the screening benefit thus requires comparable case groups.

Aron and Prorok (1986) address the problem of identifying comparable case groups for analysis of a

screening benefit (p.38). They propose that only cases diagnosed up to the "catch-up point" be used in the analysis, where the "catch-up point" is defined as "the earliest time at which the behaviour of the cumulated groups of cases in the study and control populations can be compared in an unbiased fashion," and "subsequent to this point, cases diagnosed in the study and control populations are unaffected by the screening programme and hence are comparable as well." The use of the term "catch-up" arises from their idea to choose that point in time at which the cumulative incidence curves from the two populations first become equal; presumably, before that point, the screened cumulative incidence curve exceeds that of the control population due to the effects of lead time (advance in diagnosis), while after screening ends, control incidence eventually "catches up" and remains equal to that of the study population. The "catch-up point" has thus come to refer to this point in time where equalization occurs; case groups should be comparable before and after this point (Connor and Prorok 1994, p.89).

Cumulative incidence curves from the HIP trial do indeed exhibit such behavior (Table 2 of Aron and Prorok 1986, p.38); screening ended at year 4 since start of the trial; by year 7, cumulative incidence in the control arm had exceeded that in the study arm. One could choose to analyze the survival experiences of either the 421 cases in the screened arm and the 437 cases in the control arm, or else the first 371 cases in each arm obtained by linearly interpolating between the numbers of cases at year 7 and the 365 study cases and 363 control cases diagnosed by year 6 (see Figure 1). In fact, Aron and Prorok choose to analyze cases only through year 5 following start of study. To confirm their choice, they examine the age at entry into the trial for the two arms and find no evidence of a difference in the age distribution of the participants at either year 5 or year 6.

The advantage of this definition is that it can be used in the absence of unobtainable knowledge about the underlying disease history. Ideally, we would select a group of cases in the control arm which mirror precisely the characteristics among those cases arising from the screened arm, such as sojourn time and aggressiveness of disease. The "catch-up" method, using only those cases diagnosed by the time point at which cumulative incidence curves first become equal, or nearly so as Aron and Prorok chose to do in the HIP trial, is designed to minimize the effects of length biased sampling of study cases while maintaining enough power (case accumulation) to detect a mortality reduction if it exists. The difficulty is that

"catch-up" may never occur. In fact, we show in the Appendix that the expected cumulative incidence curves will never cross if the distribution of the sojourn time has infinite support (e.g., exponential). In practice, however, it is likely that this distribution can be well approximated by one with finite support, and the fact that "catch-up" can be observed even in simulations using exponential sojourn times is explained by statistical variation. Aron and Prorok recommend using the maximum follow-up time in the event that the cumulative incidence curves do not cross.

Etzioni and Self (1995) undertook a simulation to study the effect of using this "catch-up point" in a logrank test for mortality reduction due to screening. Using three different sojourn time distributions, for each simulation, they calculate a mortality reduction at year  $C$  following start of study, where  $C$  is either a fixed case ascertainment time of 3 to 10 years or  $C$  is the "catch-up point". For either scenario, they calculate the actual size of the logrank test and find that it often exceeds the nominal size (0.05); they speculate that this bias may decrease with increased follow-up times. Not surprisingly, the distribution of "catch-up point" for the various trials in their simulation is highly dispersed; in fact, the catch-up point reaches the maximum year of follow-up (20 years) in 19%, 28%, and 54% of their simulated trials with exponential, mixture, and Weibull sojourn time distributions, respectively (p.37).

Because of the simplicity of the "catch-up point" for purposes of analysis, it is desirable to investigate the effect of this bias on screening parameters with a goal towards improving the method to reduce bias while retaining the power to detect a benefit if it exists. A wisely-chosen set of comparable cases will enable us to assess bias that may be present in an analysis of cases detected before such time as determined by a "point of comparability." In this paper, we study the effect of the "catch-up point" particularly on the estimates of the benefit time and lead time in the study and incidentally on the logrank test for a difference in mortality. This emphasis allows an interpretation of length biased sampling in terms of the survival distribution of cases in the study arm. In Section 2, we describe these estimates of average benefit time and average lead time. Various methods of assessing comparability among cases are offered in Section 3, including a new method that uses an estimate of the mean sojourn time. In section 4, we investigate, via simulation, the effect of these methods on both point estimates and interval estimates of average benefit time and average lead time, as well as on the estimate of the reduction in mortality and on the logrank test

statistic. The connection between comparable case groups and the survival component of the bias caused by length biased sampling is discussed in Section 5, along with a summary and directions for further work involving length bias.

## 2. Estimating average lead time and average benefit time.

In this paper, we focus on accurate and precise estimation of both average benefit time,  $B$ , and average lead time,  $L$ , using a very simple simulation model. Kafadar and Prorok (1994, 1995) show that the following estimators of these quantities perform quite well across a variety of shapes in the underlying preclinical and clinical durations as well as the correlation between them:

$$\text{average benefit time : } \hat{B} = \text{ave} \{ x_i - \hat{F}_C^{-1}(\hat{F}_S(x_i)) \} \quad (1)$$

$$\text{average lead time : } \hat{L} = \text{ave} \{ y_j - \hat{R}_S^{-1}(\hat{R}_C(y_j)) \} - \hat{B} \quad (2)$$

where

$x_i = i^{\text{th}}$  largest survival time (since entry) for a study case

$y_j = j^{\text{th}}$  largest survival time (since diagnosis) for a control case

$\hat{F}_S =$  Kaplan-Meier survival curve based on survival times, since entry, for study cases

$\hat{F}_C =$  Kaplan-Meier survival curve based on survival times, since entry, for control cases

$\hat{R}_S =$  Kaplan-Meier survival curve based on survival times, since diagnosis, for study cases

$\hat{R}_C =$  Kaplan-Meier survival curve based on survival times, since diagnosis, for control cases

The success of the method (in terms of both bias and variance of the estimators) depends on case group comparability. Kafadar and Prorok (1995) show that their average lead time estimator outperforms other methods; e.g. Zelen and Feinleib (1969), Morrison (1982), Hutchison and Shapiro (1968), Shapiro, Goldberg, Hutchison (1974). Although not used explicitly in (1) and (2), their approach also leads to an estimate of the average sojourn time,  $\hat{\mu}$ :

$$\hat{\mu} = N_0 / (\hat{\lambda}\hat{\beta}) \quad (3)$$

where  $\hat{\lambda}$  estimates the incidence of cases as the average number of control cases per year and  $\hat{\beta}$  estimates the screening sensitivity from the number of screen-detected cases at the initial screen ( $N_0$ ), the following screen ( $N_1$ ), and between these two screens ( $N_{0,1}$ ):

$$\hat{\beta} = (N_0 - N_1) / (N_0 + N_{0,1} - \hat{\lambda}),$$

truncated at 0 and 1.

Kafadar, Prorok, and Smith (1995) show that  $\hat{B}$  and  $\hat{L}$  are asymptotically normally distributed, and they derive, via influence functions, their asymptotic variances:

$$\widehat{Var}(\hat{B}) = s_{\bar{x}}^2 + s_{\bar{y}}^2; \quad \widehat{Var}(\hat{L}) = s_{\bar{u}}^2 + s_{\bar{v}}^2 \quad (4)$$

where  $\bar{x}$  and  $\bar{y}$  (respectively,  $\bar{u}$  and  $\bar{v}$ ) are the sample mean times to death from start of trial (respectively, mean times to diagnosis from start of trial) among cases in the study and control arms. Approximate 95% confidence limits on  $B$  and  $L$  can be determined via  $\hat{B} \pm 1.96 \cdot [\widehat{Var}(\hat{B})]^{1/2}$  and  $\hat{L} \pm 1.96 \cdot [\widehat{Var}(\hat{L})]^{1/2}$ . Both estimators use the Aron-Prorok "catch-up point" to identify comparable case groups. In the next section, we specify this point, modifications of it, and proposals for other methods of defining comparable case groups.

### 3. Methods of defining comparable case groups.

#### A. The Aron-Prorok rule, $C$ .

Aron and Prorok (1986) provide guidelines for calculating a "catch-up point," here denoted by  $C$ :

- (1) If the cumulative incidence curves cross, define  $C = \min \{ k: C_k \leq S_k \}$ , where  $C_k$  and  $S_k$  are the cumulative incidence at year  $k$  in the control and study arms, respectively;
- (2) If the cumulative incidence curves do not cross, define  $C = \text{maximum year of follow-up}, F$ .

Additional considerations are taken into account, such as similarity in the two case groups with respect to covariates such as age at entry into trial. Connor and Prorok (1994) illustrates these considerations on data from the HIP trial.

#### B. Modified Aron-Prorok rule, $C^*$ .

In the simulated randomized trials to be described in the following section, oftentimes the cumulative incidence curves from the two case groups do not cross; hence,  $C = F$ . In actual randomized screening trials such as the HIP breast cancer screening trial,  $F = 15$ , even though the last screen took place over a decade earlier. Thus, many cases in the study group could not benefit from screening, so the estimate of the average benefit time is highly diluted. To reduce the likelihood of these situations, the Aron-Prorok "catch-up rule" may be modified with an additional step:

- (1) If the cumulative incidence curves cross, define  $C^* = \min \{ k: C_k \leq S_k \}$ , where  $C_k$  and  $S_k$  are the cumulative incidence at year  $k$  in the control and study arms, respectively;
- (2) If the cumulative incidence curves do not cross, define  $C^* = \min \{ k: \text{logrank test comparing } \{C_k\} \text{ and } \{S_k\} \text{ first fails to reject the null hypothesis of equal incidence at } \alpha = 0.10 \}$ ;
- (3) If neither (1) nor (2) occurs, define  $C^* = \text{maximum year of follow-up}$ .

*C. The "mu hat rule,"  $C_\mu$ .*

As indicated in the introduction, intuition suggests that comparable case groups will be attained somewhat after screening ends, but not so long afterwards that the benefit of screening has been diluted. Because of lead time, screen-detected cases are identified earlier than they would have been in the absence of screening. The cases in the control arm which are comparable to these screen-detected cases might not appear for some time after screening ends. When the case sojourn time has an exponential distribution, the average of the distribution of these lead times is the same as the average sojourn time (Zelen and Feinleib 1969). If we denote by  $\mu$  the average sojourn time (in years), this intuition suggests that most screen-detected cases will have their counterparts in the control arm diagnosed, on average, within  $\mu$  years following the end of screening. If the sojourn time really is exponential with mean  $\mu$ , then we can expect to have identified  $\exp[-(\mu + \hat{\mu}^{1/2})/\mu] \times 100\%$  of the control cases comparable to these screen-detected cases within  $(\mu + \hat{\mu}^{1/2})$  years after screening ends (e.g., 83.7% if  $\mu = 1.5$  years). Using an estimate of  $\mu$  defined in (3), the "mu hat rule" thus defines the point of comparability as  $C_\mu = T + \hat{\mu} + \hat{\mu}^{1/2}$ , where  $T$  denotes the final year of screening.

*D. Average difference in sojourn times,  $C_D$ .*

The main source of bias in length biased sampling of cases in the study group arises from the increased sojourn time, on average, for these cases. An important component of assessing comparability of case groups from the two arms of the trial is the similarity in their distributions of sojourn times. As Connor and Prorok (1994) indicate, these distributions should be the same for cases diagnosed *before* the point of comparability as well as for cases diagnosed *after* that time point until the end of follow-up.

If we concentrate on only the *averages* of these distributions, this criterion suggests that both:

$$\begin{aligned} d_{pre}(C_D) &= [ b_S(C_D) - b_C(C_D) ] / SE[ b_S(C_D) - b_C(C_D) ] \approx 0 \\ d_{post}(C_D) &= [ a_S(C_D) - a_C(C_D) ] / SE[ a_S(C_D) - a_C(C_D) ] \approx 0 \end{aligned} \quad (5)$$

where  $b_S(C_D)$  and  $b_C(C_D)$  [respectively,  $a_S(C_D)$  and  $a_C(C_D)$ ] are the average sojourn times for the cases in the study and control arms diagnosed before (respectively, after) time  $C_D$ . One way of assuring both criteria equally is to choose that year where the Euclidean distance from the point  $(d_{pre}, d_{post})$  to the origin is minimized. We thus define  $C_D = \{i = 1, \dots, F: a_{pre}^2(i) + a_{post}^2(i) \text{ is minimum}\}$ . Notice that  $C_D$  is a theoretical point of comparability since case sojourn times are unobservable. Although simulation details are described in the next section, Figure 2 illustrates the definition of  $C_D$  for one trial in a simulation study where the sojourn time and clinical duration have a bivariate gamma distribution with means 2 and 4 and variances 1 and 4, respectively, and a correlation of 0.3 between them. Here, the point  $(d_{pre}, d_{post})$  is closest to the origin for year 16 with a Euclidean distance of 0.716 (the next closest year, 20, has Euclidean distance 0.723), so  $C_D = 16$ .

#### E. Maximum lead time, $C_L$ .

One component of the survival time since diagnosis of screen-detected cases is lead time; by definition, the lead time for such cases is positive. If we knew the maximum time at which screen-detected cases would have surfaced in the absence of screening, say  $C_L$ , we would be sure that all cases diagnosed in the two arms *after* that point in time are comparable. Since we assume that case groups are comparable by the end of follow-up, it follows that the groups of cases in the two arms diagnosed by time  $C_L$  should likewise be comparable. (See also Connor and Prorok 1994, p.89.) As with  $C_D$ ,  $C_L$  is a theoretical point of comparability, since actual lead times are unobservable. Both  $C_D$  and  $C_L$  are useful for purposes of comparison with the first three methods ( $C$ ,  $C^*$ ,  $C_\mu$ ).

### 4. Effect of "catch-up" year on estimates of average benefit/lead time.

#### A. Methods: Simulation.

Different methods of simulating randomized screening trials have been suggested in the literature; see Habbema et al. (1983), Etzioni and Self (1995), Kafadar and Prorok (1995). To investigate the distribution of the "catch-up point"  $C$ , the modification  $C^*$ , and the proposed rule  $C_\mu$ , as well as their effects on estimates of average benefit time, average lead time and mortality reduction, we simulate a series of randomized trials where case sojourn times and clinical durations are random with a bivariate gamma distribution having prescribed means, variances, and correlation (Chen, Prorok, and Graff 1983). For the "subjects" assigned to the study arm, we then impose a screening pattern on their disease history and calculate the lead time for a screen-detected case. Subjects in the control arm present as cases at the end of their preclinical phases; those in the study arm present either as screen-detected cases (with constant probability 0.80 of detection at each screen) or as interval cases (if screening fails to detect the cancer or if the preclinical phase falls entirely between two screens). The simulation consists of a "stop-screen" design with an initial screen followed by five annual screens; at the end of five years, both arms are presumed to continue with their usual medical care. Case ascertainment and follow-up extends to 20 years since start of study: cases diagnosed after year 20 are not counted, even if their preclinical phases began earlier. Once a case in a trial is screen-detected, lead time is calculated as the difference between the time of the screen and the time when the preclinical phase would have ended in the absence of screening. For positive lead times, a random benefit time can also be generated (e.g., see Kafadar and Prorok 1995, Section 4). In this study, we concentrate on the effect of the "catch-up point" and other methods described above on estimates of average benefit time, average lead time, mortality reduction, and the logrank test, so all benefit times are zero. Several enhancements to this simulation would include age-dependent incidence and sojourn distributions, variable sensitivity dependent on age, number of previous false screens, and sojourn time, and random compliance.

The simulation program is written in Fortran using the IMSL subroutine *kapmr* to calculate the Kaplan-Meier survival curve (IMSL 1992). Analyses of the simulation output data are performed in *S-Plus*, Version 3.2 for the Sun SPARC, SunOS 5.x (Statistical Sciences Inc. 1993).

*B. Results.*

Four different specifications for parameterizations are investigated, corresponding to parameters of the bivariate distribution for the sojourn time (mean, variance), clinical duration (mean, variance), and the correlation between them:

Scenario	Sojourn time		Clinical Duration		Correlation
	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

These specifications allow investigation of the effects of short versus long sojourn time, short versus long clinical duration, and low versus high correlation, using a  $2^{3-1}$  fractional factorial design. In each case, test sensitivity is set to 0.80 (20% false negative probability).

*1. Distribution of rules for comparability.*

Table 1 tabulates the frequency of occurrence for different values of  $C$ ,  $C^*$ ,  $C_\mu$ ,  $C_D$ , and  $C_L$ , based on 500 simulated randomized trials. Most widely dispersed are the distributions for  $C$ ,  $C^*$ , and  $C_D$ . Consistent with Etzioni and Self (1995), a large number of trials require observation to maximum year of follow-up (25-35% of the trials).  $C^*$  is also quite dispersed at the opposite extreme: when the sojourn time mean and variance are small,  $C^* \leq 5$  for 25-27% of the trials, before screening ends. Judging by the dispersion in  $C_D$ , the rule for determining  $C_D$  may be somewhat unreliable, since many values of  $i$ ,  $i = 1, \dots, 20$ , have similar values of  $d_{pre}^2(i) + d_{post}^2(i)$  [cf. (5)].

By design, the distribution of  $C_L$  is much less dispersed. The maximum lead time among all screen-detected cases is likely to be at least the mean sojourn time from the time screening ends, so  $C_L$  will often exceed  $T + \mu$ . For longer sojourn times [scenarios (A) and (D), whose sojourn mean and variance are 4],  $C_L$  can be as large as 20, but typically is less than 15 (87% of the trials). The "mu hat rule",  $C_\mu$ , attempts to mimic  $C_L$  and thus is also quite consistent, interestingly,  $C_\mu$  is more consistent than  $C_L$  when the sojourn time mean and variance are 4. The reduced variability for the distribution of  $C_\mu$  translates into better performance for the estimators of average benefit time, average lead time, mortality reduction, and the logrank test statistic.

2. *Distribution of error in average benefit time and average lead time point and interval estimates.*

The variability in the comparability rules noted above translate into variability in point estimates of average benefit time and average lead time as well as in confidence interval coverage. Figures 3 and 4 show the distributions of the estimate of average benefit time (nominally zero for all cases) and the error in the average lead time (which varies for each trial). For both figures:

- (a) The distributions are most dispersed, and often quite biased, using  $C_D$  (cf. Figure 3B and 4C);
- (b) The distributions are dispersed using  $C$ , but often with less or no bias (exceptions: Figure 4D);
- (c) The use of  $C_D$  would not be helpful in achieving good estimates of the average benefit time or average lead time, either in terms of bias or precision, even if it could be observed;
- (d) The use of  $C_L$  would lead to unbiased, consistent estimates of average benefit time and average lead time (exception: Figure 4C), but it is unobservable;
- (e) The "mu hat rule",  $C_\mu$ , is a good proxy for  $C_L$ , as the distributions of  $\hat{B}$  and  $\hat{L}$  are fairly consistent with little or no bias; larger biases are associated with longer sojourn times.

Table 2 demonstrates the success of confidence intervals (using eqn (4)) in terms of their non-coverage rates in each tail (nominally 2.5%). Non-coverage rates can be as high as 24.4% for the average lead time estimator  $\hat{L}$  using the catch-up rule  $C$  and even higher, 39.2%, using the modified  $C^*$  rule. Using the "mu hat rule", coverage rates are closer to the nominal 95%, although there is a tendency for the intervals to be too high rather than too low.

The use of the "mu hat rule" does not completely eliminate the effect of the variability in  $C_\mu$ . A major reason for the noncoverage rates in these intervals is due to the fact that the variance estimator in (4) actually treats  $C_\mu$ , or any point of comparability, as fixed. However, because

$$\begin{aligned} \text{Var}(\hat{B}) &= E [ \text{Var} ( \hat{B} | C_\mu ) ] + \text{Var} [ E ( \hat{B} | C_\mu ) ] \\ \text{Var}(\hat{L}) &= E [ \text{Var} ( \hat{L} | C_\mu ) ] + \text{Var} [ E ( \hat{L} | C_\mu ) ] \end{aligned}$$

the variance estimator considers only the first of these two components of variation, so confidence limits based on asymptotic normality of  $\hat{B}$  and  $\hat{L}$  are much too narrow. Figure 5 shows that, although  $\hat{B}$  has small bias when averaged over all trials using the "mu hat rule," the error in  $\hat{B}$  conditionally on  $C_\mu$  decreases as

$C_\mu$  increases from year 2 to year 20, final year of follow-up. Thus, while  $E(\hat{B} | C_\mu) \neq B$ , it appears that  $E[E(\hat{B} | C_\mu)] = E(\hat{B}) = B$ . Conversely, Figure 6 shows that the error in  $\hat{L}$  increases as  $C_\mu$  increases. (Note in these figures that the boxes for  $C_\mu \geq 15$  are often based on only very few trials. The figures do not display boxes with widths proportional to their sample sizes to avoid crowding the display at the right end.)

#### 4. Logrank test statistic and estimate of mortality reduction.

Figure 7 shows little distinction in the distributions of the estimates of the mortality reduction for the five rules. The estimates of what is nominally a null benefit can be as extreme as  $\pm 20\%$ , with greater variability associated with the longer preclinical durations. It is somewhat disturbing to see estimates of over 40% reduction when using the  $C$  or  $C^*$  rules; most commonly, these extreme estimates occur when  $C$  or  $C^*$  is less than 4, before screening ended at year 5.

The logrank test statistic for these four simulation scenarios tends to be slightly biased, but the direction of the bias is not predictable. Figure 8 shows that the catch-up rule  $C$  is generally the most consistent, followed by the modified rule  $C^*$ . For longer sojourn times, the extreme values of the statistic using  $C^*$  is always associated with those trials for which  $C^* = 2$  (17 trials in Figure 7C, 18 trials in Figure 7D). Aside from these trials, the use of either  $C$  or  $C^*$  would be preferable over any other rule, even those based on unobtainable knowledge ( $C_D, C_L$ ). Table 2 provides the sizes of the logrank test for four scenarios as the percent of trials for which the test statistic was significant for rejecting the null hypothesis of no mortality reduction. Contrary to the simulating scenario of Etzioni and Self (1995), we see here a conservative statistic, on average yielding only 2-4% Type I errors using  $C$  and 1-4% using  $C^*$ . The use of  $C_\mu$  is slightly less conservative, with Type I errors of 3.4-4.2% across the four scenarios evaluated here. The  $C_D$  rule yields invalid tests (7-9%), and the  $C_L$  rule yields a test with approximately nominal coverage except in scenario (A).

#### 5. Discussion.

A primary motivation for this work was the search for a rule to define comparable case groups that would be less variable and hence translate into less variability in estimates of average benefit time and average lead time. The proposed "mu hat rule" appears to achieve this goal while maintaining the nominal

size of the logrank test for mortality. In addition, interval estimates are more accurate, in that their coverage is closer to the nominal coverage rates, than the "catch-up rule" originally proposed by Aron and Prorok (1986).

The "catch-up rule" defined by Aron and Prorok does appear to succeed in minimizing length bias, if length bias can be measured as the difference in the average sojourn time between study cases and control cases diagnosed by time  $C$  as well as those diagnosed after time  $C$  until the end of follow-up. The distributions of  $d_{pre}(C)$  and  $d_{post}(C)$  [cf. (5)] using the "catch-up rule", as well as the other rules of comparability, are shown in Figures 9 and 10. Not surprisingly, the distributions of  $d_{pre}(C_D)$  and  $d_{post}(C_D)$  are least biased (by definition of  $C_D$ ), followed by those of  $d_{pre}(C_L)$  and  $d_{post}(C_L)$ , for all scenarios. Among the rules based on observable quantities,  $C$  often leads to the least biased distributions, followed by  $C_\mu$  and  $C^*$ . These figures help to explain the success in using  $C$  for purposes of only testing for a mortality reduction via the logrank test.

One approach to finding comparable case groups would be that time point at which the bias in  $\hat{B}$  or  $\hat{L}$  is smallest. However, such a rule, like  $C_D$  and  $C_L$ , would be based on unobservable quantities. Furthermore, in simulations not shown here, fixed-year rules yield highly biased estimates of  $B$  and  $L$ , presumably because different trials yield different patterns of cumulative incidence. This observation is consistent with the bias in the logrank test statistic using "fixed catch-up rules" noted by Etzioni and Self (1995). An adaptive rule that is based on an estimate of the mean sojourn time is thus more appealing.

Finally, we note that our attention to bias in estimating lead time and benefit time leads to a survival interpretation of the bias caused by length biased sampling. The survival times since diagnosis of comparable case groups involve both lead time and benefit time from screening. Before this point of comparability,  $\hat{B}$  will overestimate  $B$  and  $\hat{L}$  will underestimate  $L$ , and the biases may not completely cancel. Thus, one interpretation for the bias due to length biased sampling is

$$\text{Length bias}(i) = \text{Difference in observed survival since diagnosis} - B - L$$

where the difference in observed survival times since diagnosis arising from cases in the two arms diagnosed by year  $i$ . The estimation of this quantity is the subject of work in progress.

*References*

- Aron, J.L. and Prorok, P.C. (1986) An analysis of the mortality effect in a breast cancer screening study, *International Journal of Epidemiology* **15**, 36-43.
- Chen, J., Prorok, P.C., and Graff, K.M. (1983), An age dependent stochastic model of periodic screening: Length bias at a prevalence screen, *Mathematical Biosciences* **65**, 93-123.
- Connor, R.J. and Prorok, P.C. (1994), Issues in the mortality analyses of randomized controlled trials of cancer screening, *Controlled Clinical Trials* **15**, 81-99.
- Etzioni, R. and Self, S.D. (1995), On the catch-up time method for analyzing cancer screening trials, *Biometrics* **51**, 31-43.
- Gohagen, J.K. Prorok, P.C., Kramer, B.S., Cornett, J.E. (1994), Prostate cancer screening in the prostate, lung, colorectal, and ovarian cancer screening trial of the National Cancer Institute, *Journal of Urology* **152**, 1905-1909.
- Habbema, J.D.F., Van Oortmarssen, G.J., Van Putten, D.J. (1983), An analysis of survival differences between clinically and screen-detected cancer patients, *Statistics in Medicine* **2**, 279-183.
- Hutchinson, G.B. and Shapiro, S. (1968), Lead time gained by diagnostic screening for breast cancer, *Journal of the National Cancer Institute* **41**, 665-681.
- IMSL, Inc. (1992), *User's Manual: IMSL Stat/Library*, IMSL: Sugar Land, Texas.
- Kafadar, K. and Prorok, P.C. (1994), A Data-Analytic Approach for Estimating Lead Time and Screening Benefit Based on Survival Curves in Randomized Trials. *Statistics in Medicine* **13**, 569-586.
- Kafadar, K. and Prorok, P.C. (1995), Computer Simulation of Randomized Cancer Screening Trials to Compare Methods of Estimating Lead Time and Benefit Time, submitted.
- Kafadar, K., Prorok, P.C., Smith, P.J. (1995), An estimate of the variance of estimators for lead time and screening benefit in randomized cancer screening trials. Submitted.
- Mandel (1993)
- Miller, A.B. (1992), Candian trial *Can Med Assoc J* **147**, 1459-1476; 1477-1488.
- Morrison, A., The effects of early treatments, lead time, and length bias on the mortality experienced by cases detected by screening, *International Journal of Epidemiology* **11**, 261-267.
- Nystrom, L., Rutqvist, L.E., Wall, S., Lindgren, A., Lindqvist, M., Ryden, S., Andersson, I., Bjurstam, N., Fagerberg, G., Frisell, J., Tabar, L., Larsson, L. (1993), Breast cancer screening with mammography: Overview of Swedish randomized trials, *The Lancet* **341(8851)**, 973-978.
- Shapiro, S., Goldberg, J.D., Hutchinson, G.B. (1974), Lead time in breast cancer detection and implications for periodicity of screening, *American Journal of Epidemiology* **100**, 357-366.
- Shapiro, S., Venet, W., Strax, P., Venet, L. (1988), *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963-1986*, Johns Hopkins University Press: Baltimore (1988).
- Statistical Sciences Inc. (1993), *S-Plus User's Manual, Version 3.2*, Seattle, Washington.

Zelen, M. and Feinleib, M. (1969), On the theory of screening for chronic diseases, *Biometrika* **56**, 601-613.

Appendix

Expected cumulative incidence in study and control arms of a randomized screening trial

The notation is the same as in Section 3 of Kafadar and Prorok (1994). Let:

$N_{(j-1,j)}$  = Number of interval cases in study arm detected between screens  $j-1$  and  $j$ ,  $j = 1, \dots, T$

$U_j$  = Number of cases in control arm during interval  $j$ ,  $[j-1, j)$ ,  $j = 1, \dots, F$

$N_j$  = Number of cases in study arm detected at screen  $j$ ,  $j = 0, 1, \dots, T$

$N_{T+m}$  = Number of cases in study arm detected at in year  $m$  after screening ends, i.e.  $[T+m-1, T+m)$

$U_{+j}$  = Cumulative incidence in control arm through year  $j$

$N_{+j}$  = Cumulative incidence in study arm through year  $j$

$\lambda$  = case incidence rate (constant)

$\beta$  = sensitivity of screening test (constant)

$G_{ij} = \int_i^j G(x) dx$  where  $G(x)$  = survival function of sojourn time distribution

$T$  = final year of screening

$F$  = final year of follow-up

Using the formulas in Kafadar and Prorok (1984, p.585), after  $y \leq T$  years of screening:

$$E(U_{+y}) = \lambda \cdot y$$

$$E(N_{+y}) = E(N_0 + N_{(0,1)} + N_1 + N_{(1,2)} + \dots + N_{y-1,y} + N_y)$$

$$= \lambda \beta \mu + \sum_{j=1}^y [\lambda + \lambda \beta (1 - \beta)^j G_{j,\infty}]$$

$$= \lambda \cdot y + \lambda \beta \mu + \lambda \beta \sum_{j=1}^y (1 - \beta)^j G_{j,\infty} > \lambda \cdot y$$

Since  $\lambda$ ,  $\beta$ ,  $1-\beta$ ,  $\mu$ , and  $G_{j,\infty}$  are greater than 0, the expected cumulative incidence in the control arm cannot exceed that in the study arm before screening ends at  $T$ . When  $y = T+M+1$ ,  $M=0, \dots, F-1$ :

$$E(U_{+y}) = \lambda \cdot y = \lambda (T + M + 1)$$

$$E(N_{+y}) = E(N_0 + N_{(0,1)} + N_1 + N_{(1,2)} + \dots + N_{T-1,T} + N_T) + E(N_{T+1} + \dots + N_{T+M+1})$$

$$= \lambda T + \lambda \beta \mu + \lambda \beta \sum_{j=1}^T (1 - \beta)^j G_{j,\infty} + (M + 1)\lambda - \lambda \beta \sum_{m=0}^M \sum_{k=0}^T (1 - \beta)^k G_{k+m, k+m+1}$$

$$= \lambda(T + M + 1) + \lambda \beta \left[ \sum_{j=0}^T (1 - \beta)^j G_{j,\infty} - \sum_{m=0}^M \sum_{j=0}^T (1 - \beta)^j G_{j+m, j+m+1} \right]$$

$$= \lambda(T + M + 1) + \lambda \beta \left[ \sum_{j=0}^T (1 - \beta)^j (G_{j,\infty} - \sum_{m=0}^M G_{j+m, j+m+1}) \right]$$

$$= \lambda(T + M + 1) + \lambda \beta \left[ \sum_{j=0}^T (1 - \beta)^j (G_{j+M+1,\infty}) \right]$$

If the sojourn distribution has infinite support, then  $G_{j+M+1,\infty} \neq 0$ . If its support is finite, so that for some

$M$ ,  $G_{M+1,\infty} = 0$ , then  $E(U_{+y}) = E(N_{+y})$ , where  $y = T + M + 1$ .

*List of Tables*

*Table 1.* Distribution of years determined by five rules of comparability.

*Table 2.* Confidence interval coverage for  $B$  and  $L$  and size of logrank test using five rules of comparability.

*List of Figures*

*Figure 1.* Cumulative incidence in HIP trial at years 6 and 7. Solid line = control arm; dashed line = study arm; dotted line indicates time point (6.11 years) at which curves cross via linear interpolation.

*Figure 2.* Calculating  $C_D$  for one simulated randomized screening trial (Sec. 3D). Sojourn time mean 2, variance 1; clinical duration mean 4, variance 4, correlation 0.3. Character denotes the year,  $I$ , and the value  $(d_{pre}(I), d_{post}(I))$  is plotted [cf. (5)]. In this trial,  $C_D = 16$  (Euclidean distance 0.716), followed closely by  $I = 20$  (Euclidean distance 0.723). For details of the simulation, see Section 4A.

*Figure 3.* Distribution of benefit time estimator using five rules of comparability.

*Figure 4.* Distribution of lead time estimator using five rules of comparability.

*Figure 5.* Distribution of benefit time estimator by value of  $C_\mu$ .

*Figure 6.* Distribution of error in lead time estimator by value of  $C_\mu$ .

*Figure 7.* Distribution of estimate of mortality reduction using five rules of comparability.

*Figure 8.* Distribution of logrank test statistic using five rules of comparability.

*Figure 9.* Distribution of difference in average sojourn time between study cases and control cases diagnosed up to the point of comparability [ $d_{pre}(C_\mu)$ , etc.], using five rules of comparability.

*Figure 10.* Distribution of difference in average sojourn time between study cases and control cases after the point of comparability [ $d_{post}(C_\mu)$ , etc.], using five rules of comparability.

Table 2  
Confidence interval coverage for average benefit time and average lead time  
and size of logrank test, using five rules of comparability

*Sojourn time mean = 2, variance = 1; Clinical duration mean = 2, variance = 1; Correlation = 0.9*

Rule	Benefit Interval		Lead Time Interval		Size of Logrank
	Too high	Too low	Too high	Too low	
$C_{\mu}$	7.2	2.2	1.6	8.0	4.0
$C_D$	3.2	1.4	13.6	6.0	8.8
$C_L$	2.6	2.0	10.8	1.4	5.4
$C^*$	18.0	4.2	11.2	37.2	1.8
$C$	8.8	4.2	17.2	18.0	2.0

*Sojourn time mean = 2, variance = 1; Clinical duration mean = 4, variance = 4; Correlation = 0.9*

Rule	Benefit Interval		Lead Time Interval		Size of Logrank
	Too high	Too low	Too high	Too low	
$C_{\mu}$	3.0	3.0	2.2	3.8	4.2
$C_D$	6.8	2.4	4.8	18.4	8.0
$C_L$	2.4	2.0	1.4	9.8	4.2
$C^*$	7.8	4.6	8.2	24.4	1.0
$C$	5.2	5.0	9.8	22.6	2.8

*Sojourn time mean = 2, variance = 1; Clinical duration mean = 4, variance = 4; Correlation = 0.3*

Rule	Benefit Interval		Lead Time Interval		Size of Logrank
	Too high	Too low	Too high	Too low	
$C_{\mu}$	4.6	1.8	1.0	3.6	3.4
$C_D$	7.0	1.6	5.2	17.2	6.6
$C_L$	4.0	1.0	1.2	17.2	3.6
$C^*$	21.8	5.0	9.2	39.2	3.4
$C$	14.4	4.2	10.0	24.4	2.8

*Sojourn time mean = 4, variance = 4; Clinical duration mean = 2, variance = 1; Correlation = 0.3*

Rule	Benefit Interval		Lead Time Interval		Size of Logrank
	Too high	Too low	Too high	Too low	
$C_{\mu}$	8.8	2.2	1.8	7.2	3.6
$C_D$	4.8	2.8	17.6	8.4	7.2
$C_L$	3.8	3.6	7.6	5.0	3.6
$C^*$	23.8	3.8	12.2	37.8	3.6
$C$	12.6	4.6	16.8	16.0	3.8

Note: Standard errors are approximately 1.0%.

"Interval too high:" Lower limit is above true average; "Interval too low:" Upper limit is below true average.