Improving Evaluations of Cancer Screening through Better Methods of Estimating Preclinical Duration Distributions

Justin Benjamin Weinstock
Chester, Virginia

B.A., University of Virginia, United States, 2015
M.S., University of Virginia, United States, 2017

A Dissertation Presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
May 2020

Abstract

In a randomized controlled cancer screening trial, the screen-detected cancers present a length-biased sample of all preclinical durations, as cases that progress more slowly prior to diagnosis are more likely to be caught by regular screening than their faster-paced counterparts. This leads to an overestimate of the life-extending benefits of the screening test being evaluated. Previous research has shown that the severity of the length-biased sampling effect depends on the joint distribution of preclinical and clinical durations, but these simulation studies did not make data-driven choices for the distributions of these cancer growth periods.

We discovered that a mixture of two exponential distributions fit the clinical durations from three historic screening trials well after developing a simple exploratory procedure to estimate the parameters of such a mixture model. Furthermore, we found that, when simulating preclinical durations from the same type of mixture distribution, the parameters could be inferred from the pattern of diagnoses in the trial, a key finding given that preclinical durations are unobservable in a real trial. We used these simulated results to train a predictive model to estimate the distribution of preclinical durations from observable trial outcomes. Finally, we calculated the mean length-biased sampling effect under a variety of preclinical duration distributions, screening test sensitivity models, and screening programs. Using our approach to predict preclinical duration distribution parameters from trial outcomes could allow for the length-biased sampling effect to be more accurately specified for a real cancer screening trial.

Acknowledgments

While this dissertation is the end product of five years of graduate study in Statistics, it also culminates eight incredible years at the University of Virginia and twenty-one years of formal education. I would never have made it this far without the support from many teachers, friends, role models, and family members along the way. To avoid writing my acknowledgments the length of another dissertation chapter, I will try my best to be brief in expressing my gratitude.

I would like to thank all of my former teachers for inspiring my love of learning and the desire to further my education. This extends not only to my professors at the University of Virginia, but also to my teachers at Marguerite Christian Elementary School, Matoaca Middle School, and Maggie L. Walker Governor's School. I hope to be able to challenge and motivate my future students as they did for me.

I would like to recognize my favorite university professor and my advisor, Professor Karen Kafadar. It is no exaggeration that Professor Kafadar was the most coveted advisor among my Ph.D. cohort, and I feel very fortunate to have been able to work with her these past few years. Professor Kafadar encouraged me to think outside of the box and to pursue new ideas without fear of mistakes or failure. Most importantly, from all of the stories I have heard in our weekly meetings, Professor Kafadar has taught me to value all the people I meet and friendships I make throughout my career.

I would like to express my sincere gratitude to my dissertation committee members: Dr. Philip Prorok, Professor Xiwei Tang, and Professor Jianhui Zhou. Without their direction and willingness to help, this dissertation would not have been possible. I am

especially grateful to Dr. Prorok for providing access to real screening trial data for this research.

I would like to acknowledge all of my classmates from the Statistics Department. Without their friendship and collaboration, I am not sure that I could have made it through the first few semesters of graduate coursework, let alone this dissertation.

I would like to express my appreciation for the friends who have been a part of this journey. They are the reason that Charlottesville feels like home. From basketball games to tailgates to the nights that will never be forgotten, I will forever cherish the memories we have made together.

I am especially grateful to my best friend, Cynthia, for brightening my days with her spontaneity, encouragement, and contagious smile. I have always admired her compassion and work ethic, and she inspires me to want to be a better person.

Last, but certainly not least, I would like to thank my family for their unconditional love and support. My brother was both my biggest defender and biggest critic when we were growing up. While our sibling rivalry pushed us to become better students (and better athletes), I never doubted he was my greatest supporter the whole time. My mom and dad taught me the value of hard work and challenged me to never settle for meeting expectations. They helped me to set high goals for myself, and I would not be the man I am today without their faith and guidance.

Contents

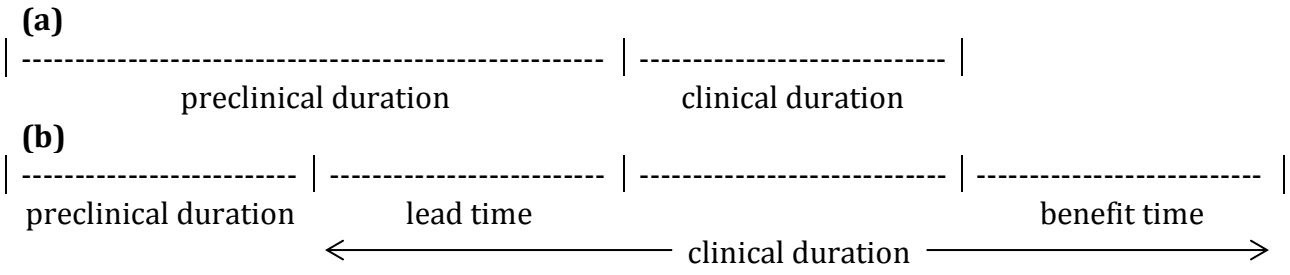
1. Introduction

Cancer is one of the leading causes of mortality worldwide. In the United States alone, the National Cancer Institute estimated that 609,640 people died from cancer in the year 2018, or approximately 163.5 deaths per 100,000 people [1]. With roughly one in five deaths in the United States caused by cancer, billions of dollars are spent each year on cancer research [2].

A key focus of cancer research is early detection through cancer screening. The goal of cancer screening is to check for and identify cancer earlier than it would have been diagnosed from symptoms alone. Because screening can detect cancers at an earlier stage, the prognosis is presumed to be better and the treatments more likely to be successful [3]. Many cancer screening tests are widely used as a part of modern medicinal practices. For example, the United States Preventative Services Task Force recommends women aged 50 to 74 have a biennial mammogram to screen for breast cancer [4]. Starting at age 50, all adults are recommended to begin screening for colorectal cancer, either with an annual stool-based test or a visual exam, like a colonoscopy, at least every ten years [5].

To understand the mechanism of cancer screening, we rely on a model for cancer growth as a two-stage process [6]. The first stage is the preclinical duration, the time during which cancer can only be detected by a screening test but otherwise is not clinically apparent. The length of the preclinical duration is unknown, as it requires knowledge of the time that cancer growth begins. The second stage is the clinical duration, the period during which cancer grows in the body post-diagnosis until its endpoint, cure or death. The clinical duration begins at the time of diagnosis and ends at death from or cure of the disease.

Figure 1.1: the timeline of a patient's life from the start of cancer growth without screening (a) and with screening (b).



This two-stage model is modified for the person whose disease is screen-detected during the preclinical phase. Figure 1.1 compares the preclinical and clinical durations for a patient without screening (a) to the same patient with screening (b) [7]. For the unscreened patient, a cancer diagnosis cannot be made until symptoms become clinically apparent and confirmed by a doctor. When the patient is screened for cancer, diagnosis can be made earlier than in the unscreened case, as the cancer can be detected before symptoms are present. When this occurs, the patient's survival time, or time from diagnosis to endpoint, increases by the length of time corresponding to the advanced diagnosis. This period is referred to as the lead time, measuring how much earlier cancer is detected in the presence of screening. The screened patient may also have an improved prognosis because the cancer has been detected and treated sooner, so it is possible that the screen-detected cancer patient will live longer than the unscreened patient would. This added lifetime is called the benefit time. Thus, relative to the unscreened patient, the screen-detected cancer patient's time from diagnosis to death is extended both by the early diagnosis (lead time) and the life-extending benefits of this early diagnosis and treatment (benefit time) if such a benefit exists.

Unfortunately, in reality, the same subject cannot be observed both in the presence and absence of screening as in Figure 1.1. To assess the benefits of any cancer screening test, comparisons must be made at the group level through a randomized controlled screening trial [8]. A cancer screening trial enrolls a large cohort of healthy subjects from the at-risk population targeted by a certain screening test. For example, a breast cancer screening trial may enroll healthy women aged 40 years or older. The subjects who agree to participate in the trial are generally divided into two arms: treatment and control. In the treatment arm, the subjects are enrolled in the screening program, meaning that they are offered screening every *r* years for an *s* year period. The subjects in the control arm receive usual care and are monitored for the development of cancer in the absence of screening, a baseline to which we can compare the screened subjects. After several years of follow-up, population death rates between the two arms are compared, as well as the survival times for the cancer patients in both arms. Because this study design randomly assigns subjects to the screened and control arms, we can expect that all relevant background characteristics to cancer prognosis, such as age or medical history, are evenly distributed between the two arms due to randomization. Thus, the only difference between the two arms is the presence of screening, and any significant difference in average group prognoses, measured by time from diagnosis to death, must be attributed effects stemming from the screening program.

When comparing the subjects who developed cancer in the screened arm and the subjects who developed cancer in the control arm, there are two estimates potentially calculated to assess the life-extending benefit of a screening test. One is the mortality ratio. The mortality rate at time *t*, calculated for either the treatment or control arm, records the

proportion of study participants who died from cancer within the first *t* years from the start of the trial. The mortality ratio compares the mortality rate in the treatment arm to the rate in the control arm [9]. The mortality ratio can be reported at various times *t* throughout the follow-up period of the screening trial, such as five-year and ten-year mortality ratios. The second estimate to assess the screening test is mean benefit time. As a reminder, the benefit time measures how much longer a cancer patient lives when diagnosed earlier through screening. The mean benefit time is an average estimate of this quantity across all cancer patients in the treatment arm of the trial [10]. These two estimates – mortality reduction and mean benefit time – are the possible metrics to assess the benefits of screening and will be discussed further in Section 2.

It is important to note that, of the cancer patients in the treatment arm, not all had their cancer diagnosed by a screening test. In fact, the cancer diagnoses among the treatment arm patients can be grouped into four general categories [7]. One of these four categories is diagnosis through screen-detection. A second case is a patient who was offered screening by the trial, but refused, and was then diagnosed through symptoms alone; this type of patient is commonly referred to as a refusal. A third category of cancer patient is an interval case, which is a patient who was diagnosed by symptoms in between two scheduled screens. For the interval case to occur, either the preclinical duration is shorter than the time until the next screen or the previous screening test(s) produced a false negative prior to the diagnosis by symptoms. The fourth and final category of cancer diagnosis in the treatment arm is post-screening, which refers to a patient diagnosed with cancer during the follow-up period after the conclusion of the screening program offered.

Among other causes, such as the length of the preclinical duration and the time at which it begins, one of the reasons why not all treatment group cancers are diagnosed by screening is the sensitivity of the test. The sensitivity of a screening test measures the proportion of cases of cancer correctly diagnosed by the test. Tests with low sensitivity will produce a higher rate of false negative cancer diagnoses, which explains why some subjects who are regularly screened in the treatment group instead have their cancers diagnosed by symptoms [11]. The sensitivity of a test is not believed to be the same for all subjects. Sensitivity is believed to depend on variables related to the cancer's growth, such as the length of the preclinical duration and how long the cancer has been growing at the time of the screen, as well as covariates related to the patient's health, such as age [12].

Besides false negatives, another problematic screening test result is a false positive, which occurs when the test suggests the presence of a cancer but the subsequent diagnostic workup is negative. While this is a detrimental outcome for the patient, false positives usually do not impact the estimates of mortality reduction or mean benefit time in a cancer screening trial because false positive test results are usually overturned after follow-up testing and the subjects are not counted among the cancer patients in the treatment arm. Although false positives do not affect the comparison of survival among confirmed cases in the trial, false positives, as well as overdiagnosed cases (to be discussed on page 14), do contribute unnecessarily, and mightily, to the healthcare burden [13].

Many of the common screening tests used today were shown to be beneficial by previous cancer screening trials, while several other screening tests have been deemed unnecessary by such studies. Two noteworthy examples are the Health Insurance Plan of Greater New York (HIP) trial and the National Cancer Institute's Prostate, Lung, Colorectal,

and Ovarian Cancer Screening Trial (PLCO). The HIP trial was one of the first randomized cancer screening trials to demonstrate the value of a screening program, consisting of an annual mammogram plus clinical breast exam, for reducing mortality from breast cancer. The trial enrolled over 60,000 asymptomatic female plan members aged 40-64 between 1963 and 1966, offering usual care to the subjects randomly assigned to the control arm and four annual mammograms with clinical breast exams to the treatment arm subjects [14]. After ten years of follow-up, there was a 30% reduction in breast cancer mortality between the treatment and control arm cancer patients [15]. The PLCO began in 1992 and enrolled 76,685 men and 78,216 women, aged 55 to 74, to assess the benefits of screening tests for the four title cancer types relative to usual care [16]. The PLCO found a significant reduction in colorectal cancer incidence and mortality among subjects who were offered screening in the form of a sigmoidoscopy, with one screen at baseline and a second screen during the next three-to-five years [17]. However, among the men screened for prostate cancer (an annual digital rectal examination for four years with an annual blood test for prostate-specific antigen for six years), the men and women screened for lung cancer (an annual chest X-ray for four years), and the women screened for ovarian cancer (an annual transvaginal ultrasound for four years with an annual blood test for the CA-125 tumor marker for six years), there was no significant reduction in mortality relative to their counterparts receiving usual care [18, 19, 20] and, in the case of ovarian cancer, a high rate (44%) of false positive results later confirmed as negative [20].

In addition to these historical randomized screening trials, several trials are currently underway, including four trials assessing various screening tests for colorectal cancer. The Nordic-European Initiative on Colorectal Cancer (NordICC) Study is enrolling

just below 70,000 men and women aged 55-64, two thirds of whom will receive usual care and one third of whom will receive a one-time colonoscopy with the removal of all detected lesions [21]. The NordICC study hopes to quantify the benefits of the colonoscopy, which is widely used to screen for colorectal cancer despite no randomized screening trials providing evidence of its utility. The other three screening trials are designed to evaluate the fecal immunochemical test (FIT) for colorectal cancer, which is related to the new and increasingly used in-home colorectal cancer screening product, Cologuard© [22]. The Screening of Swedish Colons (SCREESCO) Trial enrolled 200,000 people aged 59-62; 20,000 were offered a one-time colonoscopy, 60,000 were offered a FIT at baseline and after two years with a follow-up colonoscopy for any positive test result, and 120,000 were controls [23]. The Spanish COLONPREV Study enrolled just over 50,000 healthy adults aged 50-69; 26,703 were offered a one-time colonoscopy, and 26,599 were offered a FIT every two years with a follow-up colonoscopy for any positive test result [24, 25]. The Colonoscopy vs. Fecal Immunochemical Test in Reducing Mortality from Colorectal Cancer (CONFIRM) Study is enrolling 50,000 individuals aged 50-75 from Veterans Affairs medical centers in the United States; half will be offered a one-time colonoscopy, while the other half will be offered an annual FIT with a follow-up colonoscopy for any positive test result [26, 27].

Although the randomized controlled cancer screening trial is the optimal study design to assess the benefits of a cancer screening test, the analysis of survival from diagnosis involves challenges that must be taken into account. These screening trials are susceptible to both length-bias and overdiagnosis. A length-biased sampling method is a data collection procedure in which “bigger” outcomes are more likely to be selected than

"smaller" ones [28]. As a result, the sample has a disproportionate number of "bigger" outcomes relative to the population, so estimates from the sample data are based on an inaccurate representation of the population. In a cancer screening trial, subjects with longer preclinical durations are more likely to have their cancer detected by screening than subjects with shorter preclinical durations, especially at the start of the trial [29]. There is simply a larger window in time for screening to intervene and detect the cancer. Incidences of cancer with longer preclinical durations tend to have longer clinical durations, so the screen-detected cases will have longer times from diagnosis to endpoint even in the absence of screening benefit. Because of this phenomenon, cancer screening trials, with many slow-growing cases detected by screening, may overestimate the mean benefit time of screening.

The second important source of bias in cancer screening trials is overdiagnosis. In cancer screening, an overdiagnosis occurs when a patient's cancer is diagnosed by screening, but said cancer never would have progressed to an advanced enough stage to be detected in the absence of screening [30]. If there is overdiagnosis in a cancer screening trial, there will no longer be balance between the screened and control arms. When a cancer screening trial is designed using randomization, it is assumed that each subject in the screened arm of the study has a match in the control arm. For every fast- or slow-growing case in the screened arm, randomization ensures a counterpart (fast- or slow-growing) in the control arm. However, an overdiagnosed case in the screened arm has no such counterpart in the control arm because this "matching case" would never have been diagnosed without screening. Thus, when we compare the times to death from the screened arm cancer cases to those from the control arm, the screened arm will have some

very long clinical durations that are not matched in the control arm, which will result in overestimated screening benefit.

Section 1 has introduced randomized controlled screening trials, the study design used to rigorously assess the benefits of cancer screening tests. Two estimates of the life-extending benefits of screening tests, mortality reduction and mean benefit time, were defined, along with the biases in screening trials that can affect the mean benefit time. Finally, a number of noteworthy screening trials, both past and present, were described, highlighting the importance of research on the analysis of cancer screening trial data. The following section will review the statistical literature discussing the simulation and analysis of randomized cancer screening trial data, placing a special emphasis on those publications addressing the effects of length-bias sampling and overdiagnosis. Summarizing the limitations of these current methodologies will transition to the presentation of our research agenda, which has the goal of estimating the bias in mean lead and benefit time calculations from screening trial data.

2. Literature Review

Historically, the mortality ratio has been the more common measure of the benefit of screening applied to randomized controlled screening trial data. This is evidenced by the results published from the HIP trial and PLCO that were presented in Section 1 [15, 17, 18, 19, 20]. Kafadar and Prorok note that mean benefit time may be a more useful measure of the effectiveness of screening both because of the continuous nature of the estimate produced and its interpretability [10]. First, mortality is a binary response – death or survival – so large trials are needed to detect whether there is a significant improvement in mortality resulting from screening, especially when most trial participants in both the treatment and control arms survive or are never even diagnosed with cancer during the trial. Conversely, using benefit time, a continuous variable, to assess a screening test should offer greater power to detect a significant benefit from screening if one exists. The second advantage of mean benefit time relative to the mortality ratio is the interpretability of the measure. The possibility for extended lifetime from a cancer diagnosis by screening is just one factor that doctors consider when weighing the pros and cons of a screening test; this improved prognosis along with other advantages, such as reduced treatment costs and better patient quality of life as a result of less aggressive treatments due to the early cancer detection, must be weighed against disadvantages of the screening test, including its cost, the potential for overdiagnosis, and the prevalence of false positive results [7]. Since the assessment and recommendation of screening tests is done by doctors and organizations like the American Cancer Society and National Cancer Institute, it is important that the measure quantifying a screening test's effectiveness be easily interpretable for doctors and patients. The mean benefit time, interpreted as the average

added survival time that a patient lives when the cancer is detected by screening relative to usual medical care (i.e. clinically apparent symptoms), is simply more intuitive than the mortality ratio or percentage reduction in mortality. Moreover, as shown by Connor and Prorok using the HIP trial data, the estimated mortality ratio is also highly dependent on the time of comparison [9]. Once comparable case groups between the treatment and control arms arise, this should not be an issue for mean benefit time.

While mean benefit time is a useful measure of a screening test's effectiveness, an individual patient's lead and benefit time are unobservable in randomized controlled screening trials, so estimators of the mean lead and benefit time must be derived. Kafadar and Prorok first proposed estimators of these two quantities based on survival curves assuming that both lead time and benefit time are additive effects [31]. They recognized that the difference in survival curves from the time of diagnosis involves both lead and benefit time, but the difference in survival curves from the start of the trial involves only benefit (assuming no bias from overdiagnosis). As a result, the estimate for mean benefit time is

$$\hat{B} = average_i \left\{ x_i - \hat{F}_C^{-1}\left(\hat{F}_S(x_i)\right)\right\}, \tag{2.1}$$

where

$x_i = i^{\text{th}}$ survival time from entry in the treatment arm,

$\hat{F}_C = \text{Kaplan} - \text{Meier estimate of the survival curve from entry for the control arm,}$

$\hat{F}_S = \text{Kaplan} - \text{Meier estimate of the survival curve from entry for the treatment arm.}$

The corresponding estimate for the mean lead time is

$$\hat{L} = average_j \left\{ w_j - \hat{R}_S^{-1} \left(\hat{R}_C(w_j) \right) \right\} - \hat{B}, \tag{2.2}$$

where

$$w_j = j^{\text{th}} \text{ survival time from diagnosis in the control arm,}$$

$$\hat{R}_C = \text{Kaplan} - \text{Meier estimate of the survival curve from diagnosis for the control arm,}$$

$$\hat{R}_S = \text{Kaplan} - \text{Meier estimate of the survival curve from diagnosis for the treatment arm.}$$

Another pair of estimators for mean lead and benefit time was also proposed based on the difference in average group outcomes [32]. As mentioned previously, a difference in survival from entry between the treatment and control cancer cases in a randomized controlled screening trial must be attributed to the benefit time (assuming no overdiagnosis). In addition, a difference in time to diagnosis between the treatment and control cancer cases must be attributed to lead time. Thus, the estimate for mean benefit time is

$$\hat{B}' = \bar{v}_S - \bar{v}_C, \tag{2.3}$$

where $\bar{v}_j =$ average survival time from entry for the treatment (S) or control (C) arm.

The corresponding estimate for mean lead time is

$$\hat{L}' = \bar{u}_C - \bar{u}_S, \tag{2.4}$$

where $\bar{u}_j =$ average time to diagnosis for the treatment (S) or control (C) arm.

These estimators were proven to be asymptotically equivalent to the estimators of mean lead and benefit time based on Kaplan-Meier survival curves [32], and simulations showed that both pairs of estimators resulted in roughly the same estimates for mean lead and benefit time under various simulated distributions [7, 32]. Furthermore, the asymptotic variance of the means-based $\hat{B}'$ and $\hat{L}'$ estimators is easily calculated [32].

With reliable estimators for the mean lead and benefit time in a randomized controlled screening trial, it is also necessary to identify comparable cases between the treatment and control groups for comparison. Comparable case groups are defined as "two groups of cases, one from each trial arm, that theoretically have the same disease characteristics in the absence of screening" [7]. This begs the following question: at what time point in the trial are both case groups comparable? Recall that, in a screening trial, treatment arm subjects are offered screening only for the length of the screening program, say T years, but they continue to be tracked for follow-up until the end of the study, say F years ($T < F$). However, since treatment arm subjects receive only usual care during the follow-up period, the same as control arm subjects, they have no screening benefit (not screen-detected). Thus, any estimate of mean benefit time made using all diagnoses in the trial through year F will be "diluted" by those post-screening cases in the treatment arm [9]. Conversely, comparing only the diagnoses made through the end of screening at year T would also be a mistake. The screen-detected cases are diagnosed earlier than their unscreened counterparts by an interval known as the lead time, so many of the cases in the control arm that would match with a screen-detected diagnosis in the treatment arm (whoever they are) to create comparable case groups are not yet diagnosed by year T [7, 10]. Therefore, to identify comparable case groups for comparison between the treatment

and control arms, a catch-up point C must be chosen, where $T < C < F$, allowing additional time for the later control arm diagnoses to occur that will match with earlier screen-detected counterparts (assuming no overdiagnosis).

To choose the catch-up point C, Kafadar and Prorok discuss several methods, ultimately settling on the two best options [10]. The first option is to choose C as the first year for which the control arm cumulative cancer incidence equals the treatment arm cumulative cancer incidence [33]. However, in the presence of overdiagnosis or a preclinical duration with infinite support, control arm cumulative cancer incidence may be less than treatment arm cumulative cancer incidence for all years between T and F. In that case, C is chosen as the first year for which the logrank test fails to reject the null hypothesis of equal incidence between arms at $\alpha = 0.1$ [34]. If no such insignificant result exists between years T and F, meaning that there is excess incidence throughout the follow-up period (possibly due to overdiagnosis), the catch-up point cannot be identified [35]. A second method of choosing a catch-up point C is based on the mean preclinical duration [10]. A method-of-moments estimate for the mean preclinical duration is

$$\hat{\mu} = N_0/(\hat{\lambda}_1\hat{\beta}), \tag{2.5}$$

where

$$N_0 = \text{number of treatment arm cases detected at the first screen,}$$

$$N_1 = \text{number of treatment arm cases detected at the second screen,}$$

$$N_{0,1} = \text{number of treatment arm cases detected at between the first and second screen,}$$

$$\hat{\lambda}_1 = \text{expected first} - \text{year number of cases in treatment arm in absence of screening,}$$

$$\hat{\beta} = \text{estimated sensitivity} = (N_0 - N_1)/(N_0 + N_{0,1} - \hat{\lambda}_1).$$

We can base an estimate of $\hat{\lambda}_1$ on the annual average number of cases in the control arm after the first several years of the trial. From the method-of-moments estimate of $\hat{\mu}$, the catch-up point is chosen to be

$$C = T + 2\hat{\mu}. \tag{2.6}$$

In simulations with various preclinical and clinical duration distributions, as well as various methods of assigning a benefit time to screen-detected cases, Kafadar and Prorok found that the second method of determining the catch-up point, based on the method-of-moments estimate of the mean preclinical duration, resulted in more accurate confidence interval estimates for the mean lead and benefit times (coverage probability closer to the nominal confidence level) [10].

Although we have reliable estimates in some circumstances for mean lead and benefit time in a randomized controlled screening trial and a method of choosing comparable case groups for comparison between the treatment and control arms, these estimates can still be biased due to length-biased sampling and overdiagnosis. Moreover, both of these biases tend to favor screening [10]. However, because a cancer patient's preclinical duration is unobservable, it is impossible to identify in real data whether an individual cancer is an oversampled slow-growing case due to length-bias or an overdiagnosis. The effects of these two biases must be addressed through simulation under a variety of screening programs and plausible models for the cancer growth durations [7].

When simulating a screening trial, an important attribute of the screening test to consider is its sensitivity. Several studies have focused on estimating the sensitivity of a screening test using real trial data. Day and Walter noted a strong negative correlation between joint confidence interval estimates of the mean preclinical duration and test sensitivity, highlighting the practical difficulty of distinguishing between the effects of the distribution of preclinical durations and test sensitivity on the pattern of diagnoses in a trial [36]. Day proposed an estimator for sensitivity based on the distribution for the preclinical durations, though correctly selecting this distribution was a major assumption of the method and estimates were biased when the mean was misspecified [37]. Duffy et al used a three-state Markov chain (no disease, preclinical state, clinical state) to model "birth" into the preclinical state and "death" into the clinical state, assuming an exponential distribution for time-to-birth and time-to-death. The mean preclinical duration was estimated assuming 100% sensitivity, and sensitivity was subsequently updated using the parameters of the Markov model [38]. Straatman et al modelled the number of diagnoses at certain screens and in certain between-screen intervals using a multinomial distribution, where the multinomial probabilities stemmed from an exponential distribution assumed for the preclinical durations, and derived maximum likelihood estimates for the mean lead time and test sensitivity [39]. Hakama et al estimated test sensitivity as

$$\beta = 1 - \frac{\alpha P_1}{P_0 - (1-\alpha)P_{10}} \tag{2.7}$$

where

$\alpha =$ the proportion of treatment arm subjects who attended screening,

$P_1 =$ the rate of interval cases among those screened,

$P_0 =$ the rate of interval cases in the control arm,

$P_{10} =$ the rate of interval cases among the refusals [40].

In each of these cases, an overall test sensitivity is estimated, though Straatman et al recognized that constant test sensitivity is not a realistic assumption [39]. For simulation studies, test sensitivity, whether assumed to be constant or increasing as the preclinical duration progresses, must be tested in addition to the distributions of preclinical and clinical durations for its impact on mean lead and benefit time and the length-biased sampling effect [12, 29].

Useful measures to quantify the effects of length-biased sampling on mean lead and benefit time are the relative increases in expected preclinical and clinical durations [29]. $E[Y^*]/E[Y]$, where Y is random variable denoting the preclinical durations in the general population and Y^* denotes the preclinical durations for the length-biased sampled screen-detected cases, measures the proportional mean increase in preclinical duration for screen-detected cases relative to the general population. Similarly, $E[Z^*]/E[Z]$ measures the proportional mean increase in clinical duration (Z) for screen-detected cases relative to the general population. Because estimates for the mean benefit time rely on subjects' survival times from diagnosis in a screening trial, knowing how much longer the screen-detected subjects would have lived in the absence of screening can help to debias the mean benefit time estimate.

Kafadar and Prorok undertook a large simulation study to quantify the effects of a length-biased sample of preclinical durations on the distribution of observed clinical durations and estimates of mean lead and benefit time [29]. They used a bivariate gamma distribution to jointly model the preclinical (Y) and clinical durations (Z) as follows:

$$f(y,z) = \phi\left(\frac{{\lambda_1}^{r_1} y^{r_1-1} e^{-\lambda_1 y}}{\Gamma(r_1)} * \frac{{\lambda_2}^{r_2} z^{r_2-1} e^{-\lambda_2 z}}{\Gamma(r_2)}\right) + (1-\phi)\left(\frac{{\lambda_3}^{r_3} y^{r_3-1} e^{-\lambda_3 y}}{\Gamma(r_3)} * \frac{{\lambda_4}^{r_4} z^{r_4-1} e^{-\lambda_4 z}}{\Gamma(r_4)}\right). \quad (2.8)$$

r_i is the shape parameter in the univariate gamma density, and λ_i is the rate parameter. $0 \leq \phi \leq 1$ is a mixing proportion that represents the proportion of fast-growing cases. In Kafadar and Prorok's simulation, twenty-one different scenarios were created for the model of the preclinical and clinical durations, using different values for the parameters r_i, λ_i, and ϕ, resulting in various combinations of short, moderate, and long preclinical and clinical durations. Different screening test sensitivities and screening programs were also simulated. Across these various trials, the joint distribution of preclinical and clinical durations played the largest role in determining the effect of length-biased sampling. The next most important factor was the time in between screens, followed by the test sensitivity. However, the bivariate gamma distribution was not a particularly good model for cancer growth periods. Although the distribution was easy to simulate and enabled exact computations, its correlation structure was not realistic; conditional on being in the fast- or slow-growing group, a case's preclinical and clinical durations were uncorrelated.

Heltshe et al also undertook a simulation study similar to that of Kafadar and Prorok, but with two notable differences in the setup of the simulated trial [12]. First, Heltshe et al modelled the preclinical and clinical durations with a bivariate lognormal distribution, which they argue is more flexible than the bivariate gamma, allowing for greater possibilities in terms of the preclinical and clinical duration distribution shapes and scales. It is also simpler to specify the correlation between preclinical and clinical duration in the bivariate gamma with just one correlation parameter ρ, where the correlation in the bivariate gamma distribution is a function of all nine model parameters. Second, Heltshe et al set the test sensitivity to be a function of the length of the preclinical duration and the number of previous false negative screens; Kafadar and Prorok used constant test sensitivity. In Heltshe et al's approach, the test sensitivity $\tilde{\beta}_j(k, y)$ was defined as

$$\tilde{\beta}_j(k, y) = \beta_0 + \beta_1 y + \beta_2(k - 1), \tag{2.9}$$

where

$$y = \text{length of the preclinical duration,}$$

$$k - 1 = \text{number of previous false negative screens.}$$

For a given preclinical duration y, Heltshe et al allowed the test sensitivity to increase monotonically throughout y using either a uniform, beta, or Normal cumulative density function. Sensitivity began at 50% at the start of the preclinical duration and converged to 100% as roughly halfway through the preclinical duration. Similar to Kafadar and Prorok, Heltshe et al simulated a number of different trial scenarios, matching different preclinical and clinical distributions with several correlations and screening test sensitivities. Heltshe

et al found that mixtures of extremely different disease progressions, such as a bimodal model with both slow and fast disease growth, resulted in the largest effect of length-biased sampling on the observed clinical durations, with up to a 40% inflation in the mean clinical duration observed. The effect of length-biased sampling was less severe for models with homogeneous disease growth. The effects of length-biased sampling also generally became more severe as the correlation between preclinical and clinical duration increased. With respect to the model for test sensitivity, Heltshe et al found that increasing sensitivity to detect the longer preclinical durations exacerbated the effects of length-biased sampling (compared to constant test sensitivity). While the effects of length-biased sampling were less if test sensitivity were held constant, Heltshe et al's model is likely to be more realistic of the true sensitivity of screening tests. They saw no large differences across the different distribution functions (uniform, beta, Normal) used to model sensitivity throughout the preclinical period.

While the two aforementioned publications (Kafadar and Prorok [29] and Heltshe et al [12]) used simulations to address the effects of length-biased sampling on the clinical durations observed, no such simulation studies have assessed the effects of overdiagnosis. In general, the number of overdiagnoses in a trial is estimated by comparing the cumulative cancer incidence between the treatment and control arms. A consistent difference in the number of cases between these two arms that persists into follow-up may be attributable to overdiagnosis [35]. For example, with ovarian cancer screening in the PLCO, 212 cases were diagnosed in the treatment arm by the end of follow-up, while only 176 cases were diagnosed in the control arm in this time. This suggests that, if left unscreened, roughly thirty-six cancers detected by screening might never have caused any symptoms or

affected the women who were diagnosed by screening; the transvaginal ultrasound and annual blood test for the CA-125 tumor marker are prone to overdiagnosis [41]. However, in the HIP trial, cumulative cancer incidence in the control arm first exceeds that of the treatment arm after seven years, suggesting no overdiagnosis from an annual mammogram and clinical breast exam in that trial [33].

Kafadar and Prorok have proposed a method of selecting cases of overdiagnosis in randomized controlled screening trial data [42]. They suggest that, because the overdiagnosed cases are extremely unlikely to die from cancer, these patients may still be alive at the end of the follow-up period or perhaps dead from another cause. Using the screen-detected subjects who are still alive at the end of follow-up or dead from another cause as the pool of potentially overdiagnosed cases, they randomly select n_{OD} of these cases to be deemed overdiagnoses, where n_{OD} is equal to the difference in cancer incidence between the treatment and control arms, which is the estimated number of overdiagnoses as described in the previous paragraph.

Kafadar and Prorok also proposed equations to estimate what the mean clinical duration would have been in the absence of screening for the length-bias sampled treatment arm cases, both in the presence and absence of overdiagnosis [42]. This allows for comparison between these cases and the control arm to assess the effect of length-bias on the clinical durations that are screen-detected. Kafadar and Prorok first propose an estimator of the mean clinical duration in the absence of screening for the length-bias sampled cases assuming that there is no overdiagnosis. If there is no overdiagnosis, with comparable case groups, the total clinical duration across all cases – a product of the

number of cases (n) and the mean clinical duration in the absence of screening ($\bar{z}$) – should be equal for both the treatment (S) and control (C) groups:

$$n_C \bar{z}_C = n_S \bar{z}_S. \tag{2.10}$$

Because, as discussed in Section 1, the treatment group cases can arise in four ways, the right-hand side of Equation (2.10) can be split as follows:

$$\begin{aligned} n_C \bar{z}_C = {} & n_{refused} \bar{z}_{refused} + n_{interval} \bar{z}_{interval} + n_{post-screen} \bar{z}_{post-screen} \\ & + n_{screen\ detected} \bar{z}^* \end{aligned} \tag{2.11}$$

There is no lead or benefit time associated with refusal, interval, or post-screening cases, so the observed time from diagnosis to death for these patients is the clinical duration in the absence of benefit from screening. The only unknown value in the equation is $\bar{z}^*$, the mean clinical duration in the absence of screening for the length-bias sampled screen-detected cases, so we can solve for $\bar{z}^*$ algebraically. If a screening trial does have overdiagnosis, Equation (2.11) cannot be used because the right-hand side must include overdiagnosed cases that have no counterparts in the control group, as they would never have been diagnosed without the screening exam. Kafadar and Prorok select overdiagnosed cases in the data using the method described in the previous paragraph: randomly selecting n_{OD} subjects from the screen-detected cases still alive at the end of follow-up or dead from another cause. Once these subjects are removed from the data, $n_C \bar{z}_C = n_S \bar{z}_S$, and $\bar{z}^*$ can be

estimated using the remaining screen-detected cases in the treatment group that were not identified as overdiagnoses.

To summarize, Section 2 has presented a method for identifying comparable case groups for comparison between the treatment and control arms of a randomized controlled screening trial. Estimators have been presented for the mean lead time and mean benefit time in a screening trial that can be used to assess the effectiveness of a screening test. However, screening trials are known to provide a biased sample of survival times due to length-biased sampling and overdiagnosis, which lead to overestimated mean lead and benefit times. While the effects of length-biased sampling on the observed survival times are known to be driven by the joint distribution of the preclinical and clinical durations, the models used in simulation, the bivariate gamma and the bivariate lognormal, are not data-driven selections, and the sensitivity to misspecifying the joint distribution of the preclinical and clinical durations has not been studied. Moreover, while the amount of overdiagnosis in a sample can be estimated, the effect of overdiagnosis on estimates of mean lead and benefit time has not been studied.

3. Modelling Control Arm Clinical Durations

From the literature, specifically the simulation studies from Kafadar and Prorok [29] and Heltshe et al [12] discussed in Section 2, it is clear that correctly specifying the joint distribution of the preclinical and clinical durations is necessary to accurately assess the effects of length-biased sampling. While modelling the preclinical durations proves to be quite a challenge, as these times are inherently unknown because they measure the period in which cancer is growing prior to diagnosis, the lengths of subjects' clinical durations are readily available from randomized controlled screening trial data. We decided that it was best to try to specify the distribution of clinical durations first, and, given the strong correlation between a subject's preclinical and clinical durations, we could then assume the same type of distribution could plausibly be fit to the preclinical durations, albeit with different parameters.

To model the distribution of clinical durations, we must work with the times from diagnosis to death from the control arm cancer patients. We focused on just the controls because these subjects were not exposed to screening, so their times from diagnosis to death do not include any lead or benefit time. The times from diagnosis to death in the treatment arm would include both the lead time and benefit time for screening, so it would not be useful to consider the times from diagnosis to death for these subjects.

For this analysis, we began with the HIP trial data [15] because it was readily available. Kafadar and Prorok [10], and Connor and Prorok before them [9], had already identified the catch-up point in the HIP trial to be roughly seven years from a patient's enrollment in the study. By year 7, 437 cases of breast cancer were diagnosed in the control arm. The greatest limitation of the HIP trial data is the high rate of censoring in the

times from diagnosis to death; 205 (47%) of the 437 control group cases were censored because the subject was either still alive at the end of follow-up or dead from a cause unrelated to the cancer. Death from an unrelated cause was viewed as a censor because the true clinical duration (time from diagnosis to death caused by cancer) would have lasted longer had the patient not suffered an unrelated demise. Of these 205 censored cases, 167 (38% of all 437 control group cases) were still alive when they exited the study, so the censored clinical durations in the data largely correspond to the longest clinical durations for breast cancer patients.

A Kaplan-Meier survival curve was estimated from the 437 control group clinical durations diagnosed prior to catch-up in the HIP trial. We wanted to see how various common survival distributions, including the exponential, Weibull, gamma, log-Normal, and log-logistic distributions, would fit to this observed survival distribution. We attempted to specify the parameters of these common survival distributions such that the parametric survival function would match with the Kaplan-Meier survival curve for the HIP trial data. Unfortunately, none of the five common survival distributions provided a good fit (results not pictured).

As suggested by Kafadar and Prorok when modelling the joint distribution of preclinical and clinical durations with a bivariate gamma distribution [29], the clinical durations could be split into two groups: fast-developing and slow-developing cases. We believed that, because no single common survival distribution fit the HIP trial clinical durations well, it was necessary to model the clinical durations using a mixture of two distributions. The exponential distribution is the simplest of the five common survival distributions mentioned above, as it has only a single rate parameter to specify (all four of

the other distributions are defined by two parameters), so we decided to begin this new approach with a mixture of two exponential distributions.

The survival function, or probability of surviving beyond time x, for exponentially distributed data is

$$S(x) = e^{-\lambda x}. \tag{3.1}$$

The rate parameter λ can be estimated using $\bar{x}^{-1}$ in the absence of censoring, or, when censoring renders the sample mean survival time difficult to estimate, λ can be estimated from the slope of log-transformed survival function:

$$\log S(x) = -\lambda x. \tag{3.2}$$

When a mixture of two exponential distributions is used to model survival time data, the survival function now includes components for both the fast- and slow-developing cases:

$$S(x) = pe^{-\lambda_F x} + (1-p)e^{-\lambda_S x}, \tag{3.3}$$

where

$p =$ proportion of fast-developing cases,

$\lambda_F =$ rate parameter for the fast-developing cases,

$\lambda_S =$ rate parameter for the slow-developing cases.

With a mixture of two exponential distributions, there are now three parameters to specify and no simple mathematic solution for estimation.

The most common approach for parameter estimation in a mixture distribution is the expectation-maximization (E-M) algorithm [43]. In the case of a mixture of two exponential distributions, the E-M algorithm seeks to maximize the complete data log-likelihood for the unknown parameters in the mixture model:

$$\log L = \sum_{i=1}^{2} \sum_{j=1}^{n} z_{ij}\{\log p_i - \lambda_i x_j\}, \tag{3.4}$$

where

z_{ij} = indicator whether case j arose from component i of the mixture,

p_i = fraction of population from component i of the mixture,

λ_i = rate parameter for component i of the mixture,

x_j = time to death for case j.

The algorithm begins by assigning starting values to the z_{ij}'s, as the z_{ij}'s are all unknown. In the subsequent maximization step using the assigned z_{ij} values, the unknown parameters λ_1 and λ_2 are estimated using maximum likelihood estimation within each component i. The p_i's may also be estimated using the assigned values to the z_{ij}'s:

$$\hat{p}_i = \sum_{j=1}^{n} z_{ij}/n. \tag{3.5}$$

Following the maximization step, an expectation step adjusts the assigned z_{ij} values given the parameters estimated in the maximization step:

$$E\left[z_{ij}\middle|x_j\right] = \frac{p_i e^{-\lambda_i x_j}}{\sum_{h=1}^{2} p_h e^{-\lambda_h x_j}}. \tag{3.6}$$

The E-M algorithm continues iteratively, cycling between maximization and expectation steps until convergence. The final values of the p_i's and λ_i's are the estimated parameter values for the mixture distribution.

While the E-M algorithm is the most common approach to parameter estimation in a mixture distribution, it does have some drawbacks. First and foremost, the final parameter estimates have been shown to be highly dependent on the starting values for the algorithm, with a mixture of two exponential distributions used as an example to highlight this limitation [44]. Second, the complete data log-likelihood becomes more complex to specify and the algorithm more difficult to use in the presence of censored or truncated data [45], such as that from randomized controlled screening trials. Finally, the algorithm can also be computationally slow to yield estimates. All of these limitations suggest that there may be a better approach, or at least a simpler one, to estimate the three parameters in a mixture of two exponential distributions.

3.1 Procedure for Parameter Estimation in Mixture of Exponential Distributions

When fitting an exponential distribution to survival time data, the rate parameter can be estimated from the slope of the plot of the log-transformed survival function against time

as shown in Equation (3.2). We believed that a similarly simple approach could be developed for parameter estimation in a mixture of two exponential distributions that would be easier to apply than the E-M algorithm. Although the survival function for the mixture of two exponential distributions shown in Equation (3.3) provides no simple function when log-transformed, there is one scenario in which such a transformation can aid in parameter estimation. Consider a very large time t' such that almost all fast-developing cases would have died prior to t'. If almost all of the fast-developing cases have died before time t', then the probability of a fast-developing case surviving beyond time t' is roughly zero, meaning that the term in Equation (3.3) pertaining to the fast-developing cases would be approximately equal to zero, which reduces the equation to:

$$S(t') \approx (1-p)e^{-\lambda_S t'}. \tag{3.7}$$

When Equation (3.7) is log-transformed, it becomes a simple linear function that can be used to estimate p and λ_S:

$$\log S(t') \approx \log(1-p) - \lambda_S t'. \tag{3.8}$$

Thus, two of the three parameters in the mixture of two exponential distributions can be estimated by plotting the log-transformed survival function against time and fitting a straight line to the plot in the domain of large time points. The third parameter, λ_F, can be estimated using the survival probability for some small time t''; $\hat{S}(t'')$, $\hat{p}$, and $\hat{\lambda}_S$ can be substituted into Equation (3.3) to solve for $\hat{\lambda}_F$.

There are two conflicting challenges of estimating p and λ_S using Equation (3.8). First, what qualifies as "a very large time" depends on the distribution of times to death for the fast-developing cases and, therefore, λ_F, which has yet to be estimated. Because of this concern, it would seem that p and λ_S should be estimated by fitting a straight line to the log-transformed survival function for the largest times to death available in the sample data. This conservative approach should ensure that the survival times being used to estimate λ_S are not fast-developing cases. However, the second challenge of estimating p and λ_S using Equation (3.8) is the high rate of censoring as survival times increase in randomized controlled screening trial data. Censored data, of course, arise in any clinical trial (loss to follow-up, low mortality from disease under study, etc.), but cancer screening trials have very high rates of censoring, especially for cancers that tend to have higher five-year survival rates anyway (e.g., breast versus pancreatic cancer). Nonetheless, this challenge may well occur in other kinds of clinical trials also, so an approach to taking it into account will benefit the analysis in other applications. In a screening trial, a high rate of censoring for long survival times means that there will be few observed deaths and many censored times in the domain of large time points, so it could be difficult to accurately estimate p and λ_S using the conservative approach with few very large observed times to death and a potentially fairly flat empirical survival curve in this domain.

To balance both of the aforementioned challenges in the application of our method, we estimate the parameters of a mixture of two exponential distributions using an iterative process. Our idea for the procedure is outlined as follows:

1. Calculate the Kaplan-Meier estimated survival curve, and plot $\log \hat{S}(t)$ against time. Fit a line to the plot in the domain of the largest observed times to death. Estimate $(1-p)$ and λ_S using Equation (3.8).
2. For some small time point t^*, estimate λ_F by substituting $\hat{S}(t^*)$, $\hat{p}$, and $\hat{\lambda}_S$ from step 1 into Equation (3.3). For screening trial data, a value of t^*in the 100 to 500 day range seems appropriate.
3. Given the estimate of λ_F from step 2, determine the time point t^{**} by which most (perhaps 95% or 99%) of the fast-developing cases will have died using the exponential cdf.
4. Step 1 can be repeated. However, instead of fitting the line in the domain of the largest observed times to death, the line can now be fit in the domain of times just larger than t^{**}.
5. Step 2 can now be repeated given the updated estimates of p and λ_S from step 4.
6. Steps 3-5 repeat iteratively until there is no change in the parameter estimates.

In the first iteration, which is completed in steps 1 and 2 of the procedure, we implement the conservative approach designed to ensure that only slow-developing cases are used to estimate λ_S. However, because we want to consider only the largest observed times to death in this iteration to estimate $(1-p)$ and λ_S, there are likely only sparse observations to use because of censoring and the long right tail of the exponential distribution producing few extremely long survival times. As a result, the initial estimates of $(1-p)$ and λ_S may be too small. The second iteration, which is completed in steps 3 through 5, addresses these drawbacks by reevaluating $(1-p)$ and λ_S in a domain of time closer to zero than the

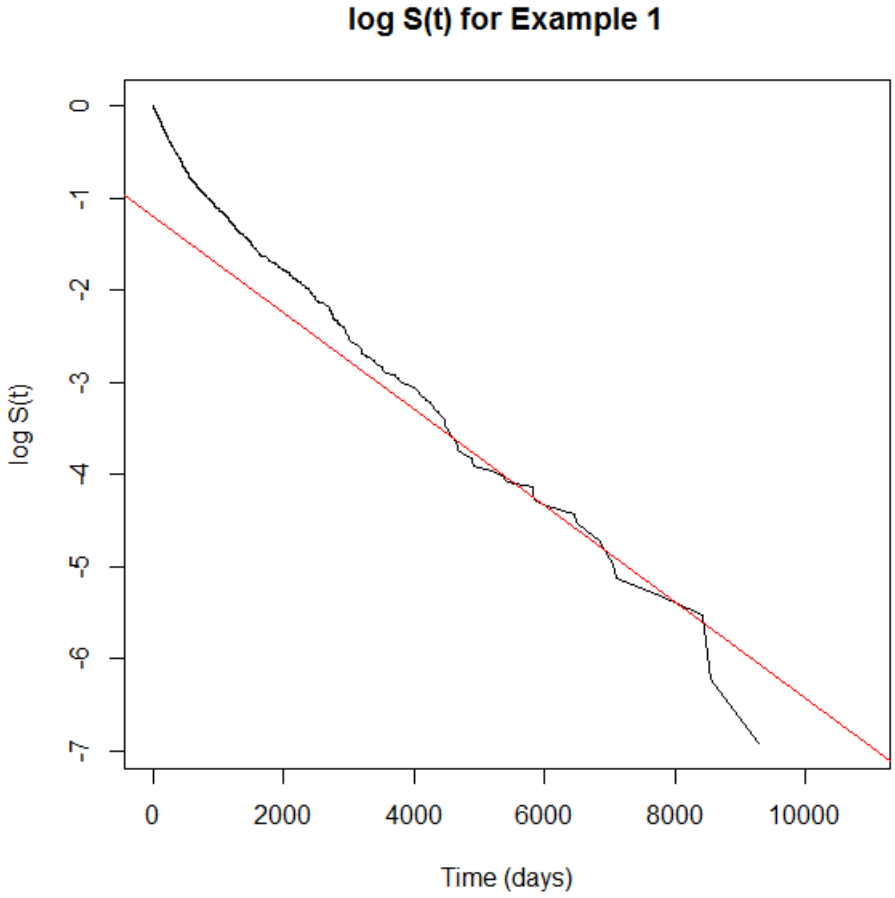
extreme end, where there are more observed death times but it is still expected that only slow-developing cases are used to estimate λ_S. All subsequent iterations seek to adjust the estimates of $(1-p)$ and λ_S given updated information about λ_F and the domain of time where almost all observed deaths are slow-developing cases until a stable model is achieved.

Perhaps our procedure could be viewed as a "poor man's version" of the E-M algorithm. Like the E-M algorithm, our procedure iterates between separating observations based on the component of the mixture distribution they likely originated and updating parameter estimates based on these divisions. A formal E-M procedure for estimating the parameters in screening trial survival times may be difficult to obtain because of the high rate of censoring. The potential advantages of our approach versus a formal E-M algorithm are discussed further in Section 3.6.

3.2 Simulations

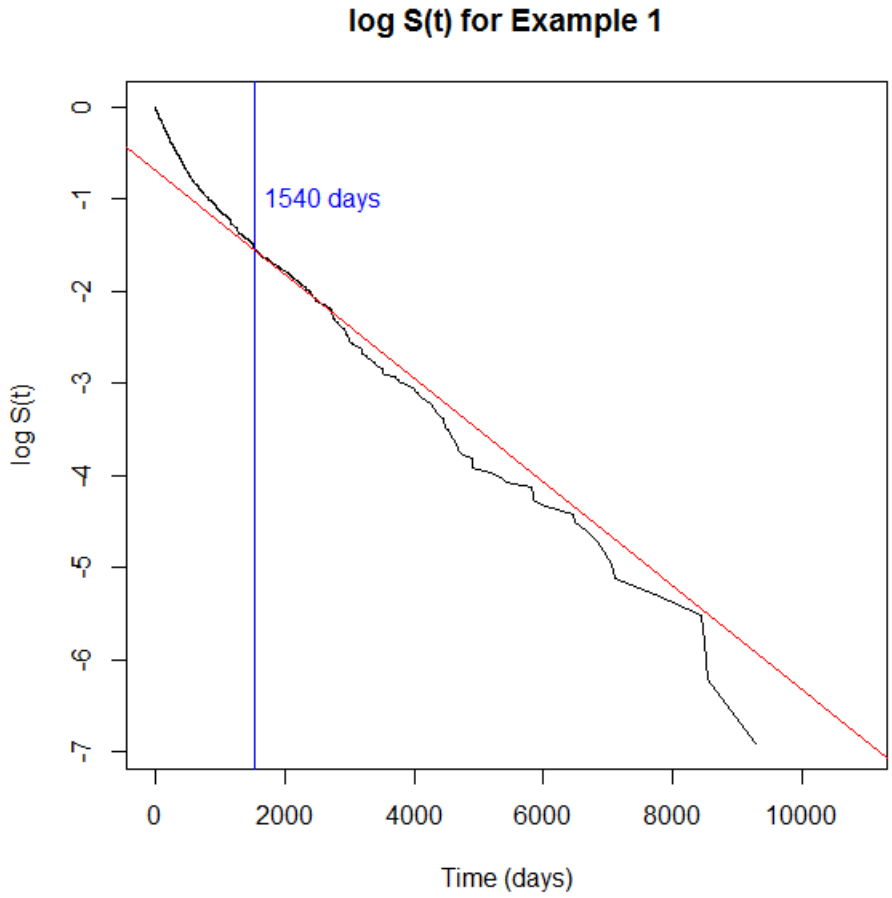
While our simple procedure should be reliable in theory, we wanted to assess it in a few simulated examples before returning to the HIP trial data. The goal of these simulations is to determine the accuracy of the procedure in estimating the three parameters of the mixture distribution, as the true parameter values will be set by the simulation design. Example 1 will be presented in detail, as it introduces the general simulation design and illustrates the application of our procedure. In the subsequent simulations, only the mixture distribution parameters or rate of censoring change in the simulation design, so these examples may be discussed more briefly with a focus primarily on the accuracy of our procedure.

Figure 3.1: plotting $\log \hat{S}(t)$ against time for example 1 and fitting a line (red) in the domain of the largest observed times to death.



In example 1, 1,000 survival times are generated from a mixture of exponential distributions that is 50% fast-developing cases with a mean duration of 1 year and 50% slow-developing cases with a mean duration of 5 years: $p = 0.5$, $\lambda_F = 1/365$, and $\lambda_F = 1/(5 * 365)$. For this example, none of the survival times are censored. To begin our procedure, we calculated the Kaplan-Meier estimated survival curve for these 1,000 cases and plotted $\log \hat{S}(t)$ against time, as shown in Figure 3.1. Considering those times to death in the 5,000 to 8,000 day range to be the largest available, we fit a line to the plot with an intercept of $\log 0.3$ and a slope of $1/(5.25 * 365)$. This provided our initial estimates of

Figure 3.2: plotting $\log \hat{S}(t)$ against time for example 1 and fitting a line (red) in the domain of times just larger than $t^{**} = 1{,}540$ days.



$\hat{p} = 0.7$ and $\hat{\lambda}_S = 1/(5.25 * 365)$. Moving on to step 2 of the procedure, we selected $t^* = 100$ days, and calculated $\hat{S}(100) = 0.86$ from the simulated data. We were able to solve for $\hat{\lambda}_F = 1/514$ using the initial estimates for $\hat{p}$ and $\hat{\lambda}_S$. In step 3 of the procedure, we calculated that, if $\lambda_F = 1/514$, 95% of the fast-developing cases should end within the first 1,540 days from diagnosis. Revisiting the plot of $\log \hat{S}(t)$ against time in step 4 of the procedure, we wanted to fit a line in the domain of times just longer than 1,540 days. This is shown in Figure 3.2, and we obtain new estimates of $\hat{p} = 0.49$ and $\hat{\lambda}_S = 1/(4.85 * 365)$. In step 5, we return to $\hat{S}(100) = 0.86$ and adjust our estimate of $\hat{\lambda}_F$ given the new values of

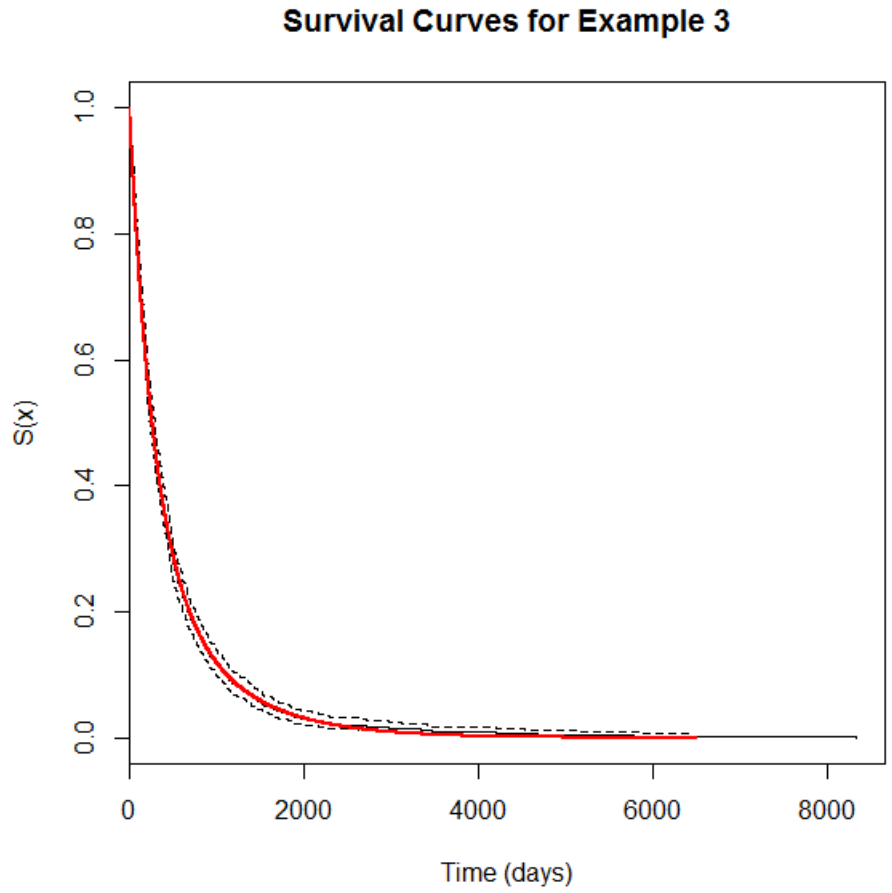
Figure 3.3: results of examples 1-3.

Example	True Parameters			Estimated Parameters			Iterations
	p	λ_F	λ_s	p	λ_F	λ_s	
1	0.5	1/365	1/1825	0.49 (2%)	1/390 (6%)	1/1770.25 (3%)	2
2	0.25	1/730	1/2555	0.26 (4%)	1/1202 (40%)	1/2737.5 (7%)	3
3	0.75	1/273.75	1/1460	0.65 (13%)	1/267 (3%)	1/821.25 (78%)	2

$\hat{p}$ and $\hat{\lambda}_S$ to be $1/390$. Now, if $\lambda_F = 1/390$, 95% of the fast-developing cases should end within the first 1,168 days from diagnosis. Once again revisiting the plot of $\log \hat{S}(t)$ against time, we see that the slope is no different in the domain of times just longer than 1,168 days than it is in the domain just longer than 1,540 days, which means that our estimates of $\hat{p}$ and $\hat{\lambda}_S$ would not change. Thus, our algorithm stops after this second iteration, and our parameter estimates are $\hat{p} = 0.49$, $\hat{\lambda}_F = 1/390$, and $\hat{\lambda}_S = 1/(4.85 * 365)$.

In examples 2 and 3, the three mixture distribution parameters are varied, but there is still no censoring. Example 2 represents a slower distribution: $p = 0.25$, $\lambda_F = 1/(2 * 365)$, and $\lambda_F = 1/(7 * 365)$. Example 3 represents a faster distribution: $p = 0.75$, $\lambda_F = 1/(0.75 * 365)$, and $\lambda_F = 1/(4 * 365)$. The final parameter estimates from simulated examples 1 through 3 are listed in Figure 3.3, along with relative errors for the estimates. Overall, our procedure is fairly accurate in estimating the mixture distribution parameters. Some of the smaller relative errors in parameter estimation may even be explained by sampling variability, though this effect was made to be small with a simulated sample size of 1,000 for each example. However, in examples 2 and 3, the method struggled to estimate

Figure 3.4: comparing the survival curve fit by our procedure (red) to the Kaplan-Meier survival curve for the simulated control arm clinical durations (black) in example 3.



the rate parameter for the smaller component of the mixture distribution. For instance, only 25% of cases were slow-developing in example 3, and, while the procedure estimated the rate of the fast-developing cases fairly accurately, the rate parameter for the slow-developing cases was significantly overerestimated. Because the parameter was properly specified for the larger component of the mixture, the survival curve fit by our procedure still matches fairly closely to the empirical survival curve from the simulated data, as shown in Figure 3.4, and it can be calculated that the survival probabilities from a

Figure 3.5: results of examples 1, 4, and 5 (where, in all three examples, the true parameter values are $p = 0.5$, $\lambda_F = 1/365$, and $\lambda_F = 1/1825$).

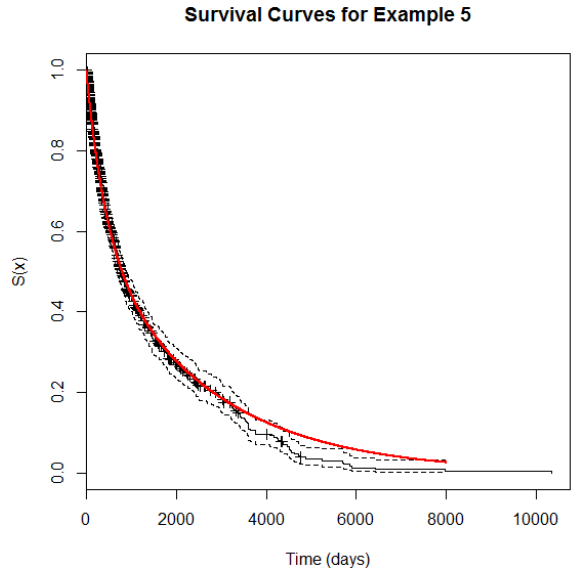
Example	Censoring Probability	Estimated Parameters			Iterations
		p	λ_F	λ_s	
1	0	0.49 (2%)	1/390 (6%)	1/1770.25 (3%)	2
4	0.25	0.43 (14%)	1/342 (7%)	1/1934.5 (6%)	2
5	0.4	0.4 (20%)	1/420 (13%)	1/2555 (29%)	2

distribution with the true parameters and the survival probabilities from a distribution with our estimated parameters are very similar.

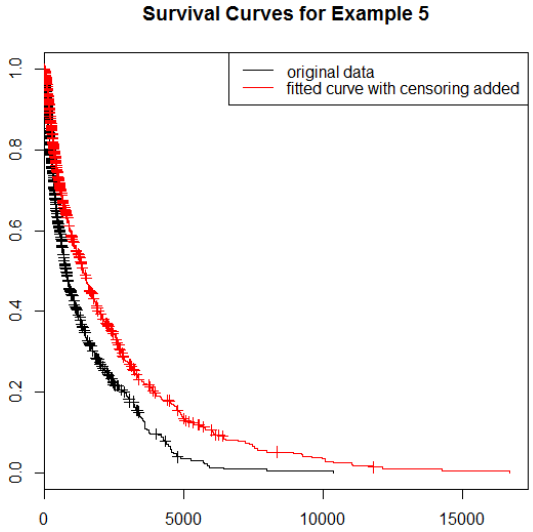
Examples 4 and 5 introduced censoring to the simulated survival times. For both examples, 1,000 survival times are generated from a mixture of two exponential distributions with $p = 0.5$, $\lambda_F = 1/365$, and $\lambda_F = 1/(5 * 365)$. Each observation has the same 25% chance of being censored in example 4, and the censoring probability increases to 40% in example 5. In either example, a case that has been selected to be censored will have its true survival time rescaled by a random observation from the Uniform(0, 1) distribution so that the time of censoring will occur prior to the true time of death. These two simulated examples are designed to illustrate random censoring. The final parameter estimates and relative errors, along with the results from example 1, which used the same simulated distribution but without censoring, are listed in Figure 3.5. It is clear that the relative errors of the estimated parameters increase with the rate of censoring. Furthermore, while the size of the relative errors may be acceptable in example 4 with its

Figure 3.6: comparing the survival curve fit by our procedure (red) to the Kaplan-Meier survival curve for the simulated control arm clinical durations (black) in example 5.

(a) the fit from our procedure plotted without censoring added.



(b) the fit from our procedure plotted with random 40% censoring added.



moderate censoring probability, they are awfully large in example 5 with its high censoring probability. When the censoring is random and occurring at such a high rate, there is a lot of early dropout and fewer observed deaths, so the Kaplan-Meier survival curve estimated from the data will decrease more slowly than the true underlying survival function of the distribution. This causes our algorithm to underestimate the rate parameter for both the fast- and slow-developing components. As shown in Figure 3.6a, the curve estimated by our procedure fits the majority of the simulated data for example 5 well, overlapping the Kaplan-Meier curve for roughly the first 3,500 days. Unfortunately, the model fit by our procedure is more heavy-tailed than the Kaplan-Meier curve because it is overfitting the slow decrease of the Kaplan-Meier estimated survival function caused by heavy censoring. Figure 3.6b shows the distribution fit by our procedure with random 40% censoring added

compared to the simulated sample data from example 5. Here, it is clear that λ_F and λ_S have been underestimated, which was not as evident in Figure 3.6a when our estimated distribution was plotted without censoring. This suggests that adjustments should be made to the parameter estimates from our procedure in the event of heavy censoring as seen in example 5, such as systematically increasing the estimates for λ_F and λ_S to improve the overall fit. Moreover, we need a method of estimating the censoring probability so that we may generate plots like Figure 3.6b to evaluate the fit of the distribution estimated by our procedure while giving consideration to the censoring. For instance, had example 5 been real data instead of a simulation, we would not have known that every observation had a 40% chance of censoring to generate Figure 3.6b from the distribution estimated by our procedure. In Section 3.3, we will present an idea for estimating the censoring probability as a function of the case's time-to-death.

3.3 Procedure to Estimate the Censoring Probability as a Function of Time-to-Death

When using our procedure to estimate the parameters of a mixture of two exponential distributions, it is important for the rate of censoring to be considered when assessing the estimated fit. However, to add simulated censoring to the survival times from our estimated distribution, we must approximate the probability of being censored for each case. In examples 4 and 5 in the previous section, censoring was purely random; every case had the same probability of being censored. In reality, there is usually some aspect of the data collection method causing the censoring. For randomized controlled screening trials, a subject's survival time is censored for one of two reasons: (i) the subject dies from another cause not related to cancer or (ii) the subject is still alive at the end of the trial's

follow-up period. As we saw in the HIP trial data presented at the beginning of Section 3, most of the censored survival times in screening trial data are cases where the subject is still alive at the end of the study; this reason accounted for over 80% of the censoring in the HIP trial. If a survival time is censored because the subject is still alive at the end of the follow-up period, it is much more likely that the subject has a long survival time than a short one. Thus, we believe it is useful to estimate the censoring probability as a function of the subject's true survival time, as the length of the survival time is one of, if not the, leading cause of censoring.

To estimate the probability of censoring for a case with survival time in a certain interval, say $[t_i, t_j)$, we need to compare the number of observed uncensored survival times in that interval to the expected number of survival times in the same interval. The difference between the number of observed uncensored cases and the expected number of cases can be attributed to censoring, and the probability of being censored can be estimated by dividing this difference by the expected number of cases in the interval. While the number of observed uncensored cases in a certain time interval can be counted in the sample data, the expected number of cases in this interval depends on the underlying distribution of survival times. Using our procedure described in Section 3.1, we can estimate this underlying survival time distribution and calculate the expected number of cases from a sample of size n that occur in $[t_i, t_j)$ by

$$n * \Big(S(t_j) - S(t_i)\Big). \tag{3.9}$$

Now, to develop a function for assigning a censoring probability to cases based on their survival time, we need to calculate the estimated censoring probability for several intervals $[t_i, t_j)$. We can then plot the estimated censoring probabilities against the midpoints of the intervals to get an idea of the functional form of the relationship between censoring probability and survival time. A simple way to do this is by subsetting the data into bins of width w; this creates the intervals $[0, w)$, $[w, 2w)$, etc. However, choosing w in this case is a bit of a balancing act. If w is too wide, then the domain of survival times in the sample data will be divided into a very small number of bins, meaning that there will be very few points to plot to try to assess the functional form of the relationship between censoring probability and survival time. If w is too narrow, then the observed uncensored counts and expected counts in each interval will be fairly small, and the censoring probabilities will not be estimated very precisely. Ultimately, the choice of w is unique to the sample data in question, specifically its sample size and the domain of survival times; if the sample size is large or the domain of survival times is small, then a smaller value for w may be selected. It is also worth noting that, in some instances, especially intervals pertaining to short survival times where the probability of censoring is low, the number of observed uncensored cases in the interval may be greater than the expected number of cases just due to chance. If this happens, the estimated censoring probability calculated as described in the previous paragraph would be negative, which is clearly impossible. In the event that this does happen, we recommend assigning a very small censoring probability, perhaps between 0 and 5%, when plotting to determine the functional form of the relationship between censoring probability and survival time.

With a function to assign censoring probabilities to cases based on their survival time, we are now almost able to simulate the effects of censoring on our estimated distribution of survival times, as exemplified in Figure 3.6b. When a case has been probabilistically selected to be censored, we just need to determine at what point in its survival time the censoring occurs prior to death. Our idea is to rescale the survival time by multiplying it by a random value from a beta random variable, say rB. The beta distribution is ideal because its support is $[0, 1]$, the same as an allowable proportion, so a random beta observation can be treated as the proportion of the survival time that is observed prior to censoring. Moreover, the beta distribution is flexible in its shape, meaning that its parameters can be specified in a way that certain values of rB are likelier than others. For example, if a censored case is equally likely to be censored at any point during its survival time, then the beta$(1,1)$ distribution can be used. If a censored case is most likely to be censored somewhere in the middle of its survival time, then the beta$(3,3)$ distribution may be appropriate. If a censored case is more likely to be censored as it gets deeper into its survival time, the beta$(2,1)$ distribution is a reasonable selection. Ultimately, we believe the choice of beta distribution parameters is best made using subject matter expertise on the data collection method and why censoring is occurring. As has already been discussed, censoring in randomized controlled screening trials usually occurs because the subject is still alive at the end of the study, which suggests that censoring is not likely to occur shortly after the subject's diagnosis early in the survival time, but rather later on in the survival period when the subject may reach the end of the follow-up period; for this reason, we like the beta$(2,1)$ distribution to rescale censored survival times in simulated screening trial data. It is also worth noting that, in some instances, the survival

times should be capped prior to being rescaled by rB. In screening trials, the maximum possible observed survival time, censored or uncensored, is the length of the study; if an individual survival time is longer than the length of the study, it makes sense to apply rB to the study length to determine how long the individual is measured prior to censoring rather than multiplying the true survival time by rB. Thus, for simulated screening trial data, we will assign the censor time c to a case probabilistically selected to be censored as

$$c = rB * \min(z, F), \tag{3.10}$$

where

$$z = \text{true length of the uncensored survival time,}$$

$$F = \text{length of follow} - \text{up for the screening trial.}$$

3.4 Additional Simulations

Returning to the simulated examples carried out in Section 3.2, we will now introduce censoring probability as a function of the case's survival time. For both example 6 and 7, 1,000 survival times are generated from a mixture of two exponential distributions with $p = 0.5$, $\lambda_F = 1/365$, and $\lambda_F = 1/(5 * 365)$; this is the same scenario as examples 1, 4, and 5. In example 6, the censoring probability increases linearly with survival time, starting at 0% for a survival time of zero days and reaching a 100% censoring probability at 4,000 days. In example 7, the probability that a survival time is observed, which is one minus the probability of censoring, decreases in a quadratic function as the survival times extend, starting at 100% for a survival time of zero days and reaching a 0% chance of being fully observed (uncensored) at 4,000 days. For both examples, a survival time that has been

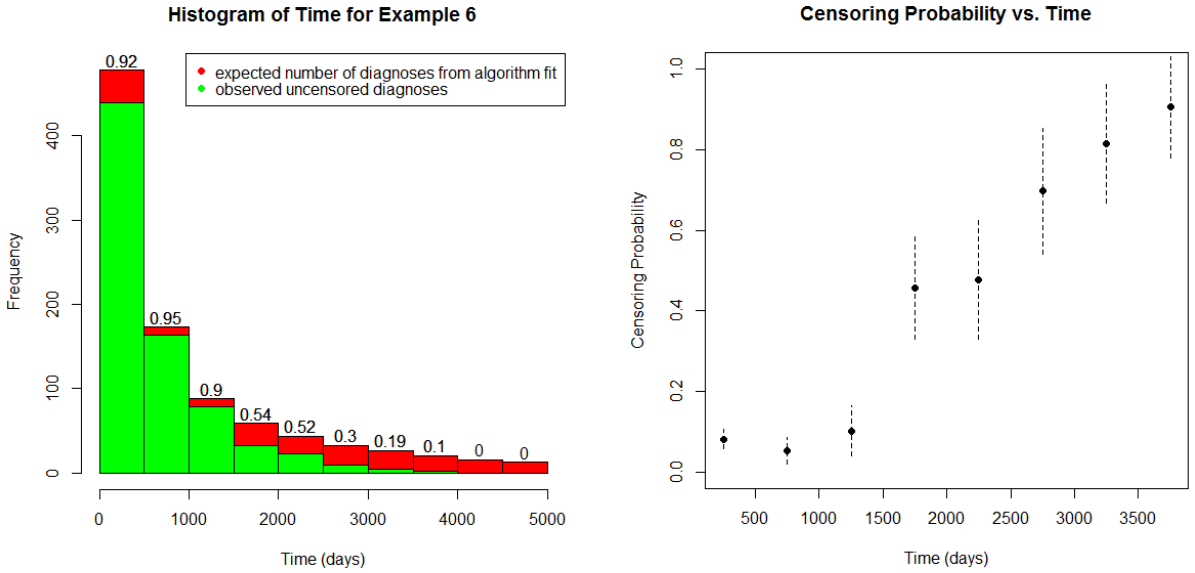
probabilistically selected for censoring will be rescaled by a random observation from a beta(2,1) distribution as discussed in Section 3.3.

Starting with example 6, our procedure estimated the mixture distribution parameters to be $\hat{p} = 0.48$, $\hat{\lambda}_F = 1/344$, and $\hat{\lambda}_S = 1/(5.75 * 365)$. These estimates had relative errors of 4%, 6%, and 13%, respectively. Given the roughly 20% overall censoring probability in example 6, these relative errors are similar to those in example 4, which had a 25% random censoring probability for each observation. However, different from example 4, rate parameter for the slow-developing cases now has the largest relative error, which may be explained by the fact that the longer cases are more likely to be censored in example 6, making their rate parameter more difficult to estimate. To plot the fit estimated by our procedure for comparison to the simulated data, we needed to also estimate the censoring probability as described in Section 3.3. We selected $w = 500$ days, which divided the domain of simulated survival times into eight intervals, as there were no observed survival times greater than 4,000 days in length. In each interval, the number of observed uncensored survival times was counted, and the expected number of cases was calculated using Equation (3.9), where $S(x)$ was the survival function for a mixture of exponential distributions with $p = 0.48$, $\lambda_F = 1/344$, and $\lambda_S = 1/(5.75 * 365)$. Figure 3.7a displays these results in a histogram, where the green bar shows the number of observed uncensored cases in the interval, the red bar shows the expected number of cases, and the proportion of expected cases that were observed uncensored labels the bin. Figure 3.7b plots the censoring probability, which is the complement of the proportion of observed uncensored cases, against time. 95% margins of error are also added to each interval's estimated censoring probability, primarily to reflect the sample size (expected count) in

Figure 3.7: estimating censoring probability as a function of survival time for example 6.

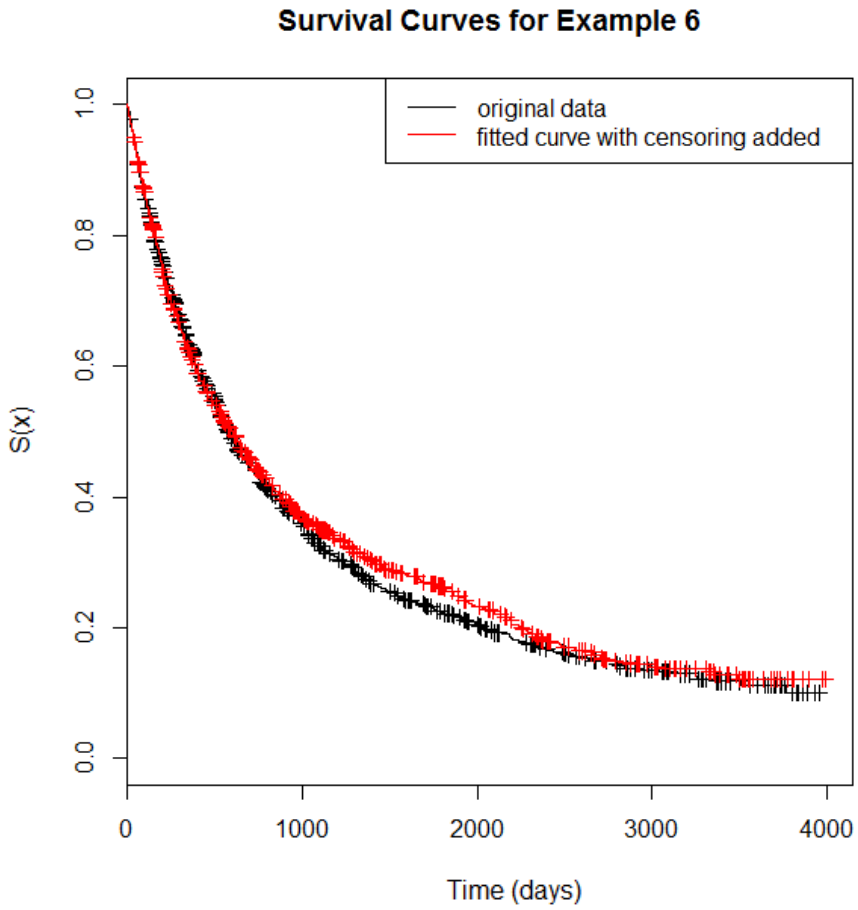
(a) histogram displaying the proportion of observed uncensored cases in each survival time interval.

(b) plot of estimated censoring probability against survival time with margins of error



each interval used to estimate this probability. Recall that the true function assigning censoring probabilities is linear; the probability of censoring is equal to the case's time-to-death in days divided by 4,000, and all times greater than 4,000 days have a censoring probability of 100%. While the function does not appear perfectly linear in Figure 3.7b, as the estimated censoring probability remaining roughly the same for the first three intervals causes the function to seem perhaps polynomial in its increase, it can be described as roughly linear overall. This result suggests that we should prefer the simpler functional explanation to describe the relationship between censoring probability and survival time, rather than trying to specify a complex functional relationship from such a small number of time interval data points. Finally, now that we have an idea of the function to assign censoring probability based on survival time, we may generate data with censoring from a

Figure 3.8: comparing the survival function estimated by our procedure with censoring added (red) to the Kaplan-Meier survival curve for the simulated control arm clinical durations (black) in example 6.



mixture of exponential distributions with $p = 0.48$, $\lambda_F = 1/344$, and $\lambda_S = 1/(5.75 * 365)$ to compare to the simulated data from example 6. This comparison is plotted in Figure 3.8. Because the parameters estimated by our procedure have small relative errors compared to the true simulated parameters, and because we were able to reliably estimate the censoring probabilities, we see that our estimated fit matches well with the original data.

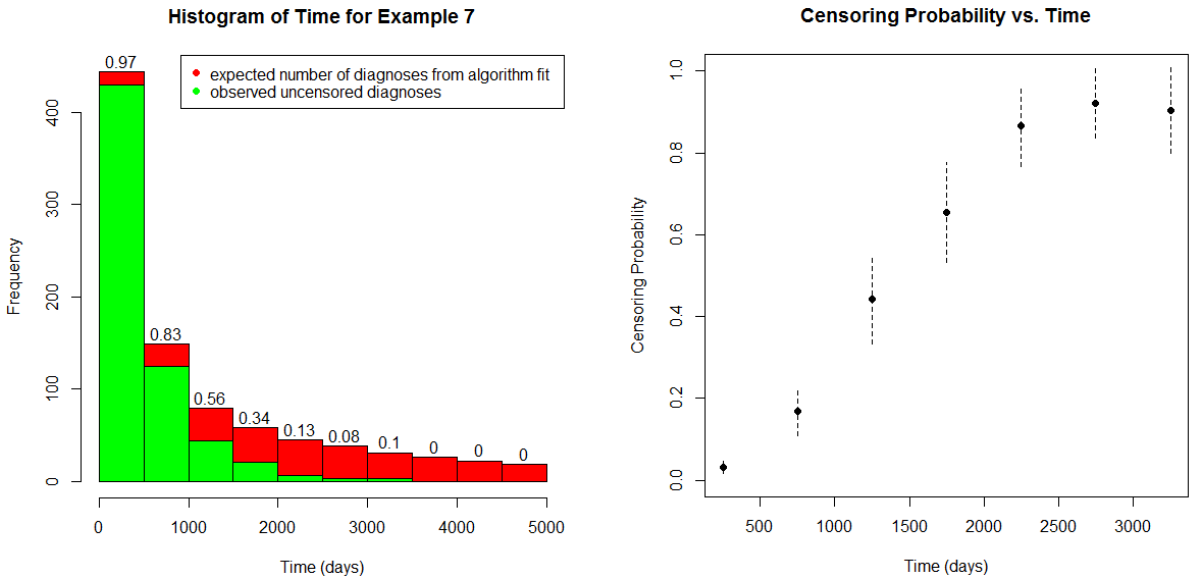
Proceeding to example 7, our procedure estimated the mixture distribution parameters to be $\hat{p} = 0.44$, $\hat{\lambda}_F = 1/315$, and $\hat{\lambda}_S = 1/(7.5 * 365)$, which had relative errors

of 12%, 16%, and 33%, respectively. The overall censoring probability was approximately 35% in example 7, almost twice that of example 6, and the relative errors increased with the censoring rate, just as we saw when comparing examples 1, 4, and 5 in Section 3.2. Additionally, similar to example 6, with the longer cases more likely to be censored, the rate parameter for the slow-developing cases had the largest relative error. With example 5, we discussed adjusting underestimated rate parameter estimates in the presence of high random censoring, but, when the censoring probability increases with survival time, perhaps only the rate parameter for the slow-developing cases may need to be manually increased. To estimate the functional relationship between censoring probability and survival time in order to plot the fit estimated by our procedure for comparison to the simulated data, we again divided the domain of simulated survival times into eight intervals of $w = 500$ days. The number of observed uncensored survival times was counted in each interval, and the expected number of cases was calculated using Equation (3.9), with $S(x)$ specified by our estimated parameter values. Figure 3.9a compares the number of observed uncensored cases to the expected number of cases in each interval, and Figure 3.9b plots the censoring probability against time with 95% margins of error added to each interval's estimated censoring probability. In Figure 3.9b, the probability of censoring appears to increase linearly with survival time over the first 2,500 days before levelling off around 90%; this matches fairly closely to the true function assigning censoring probability to cases, which decreases the probability a case is observed in a quadratic function as the survival time increases. Just as in example 6, our method for estimating the functional relationship between censoring probability and survival time is rather accurate. Using our estimated function to assign censoring probability based on

Figure 3.9: estimating censoring probability as a function of survival time for example 7.

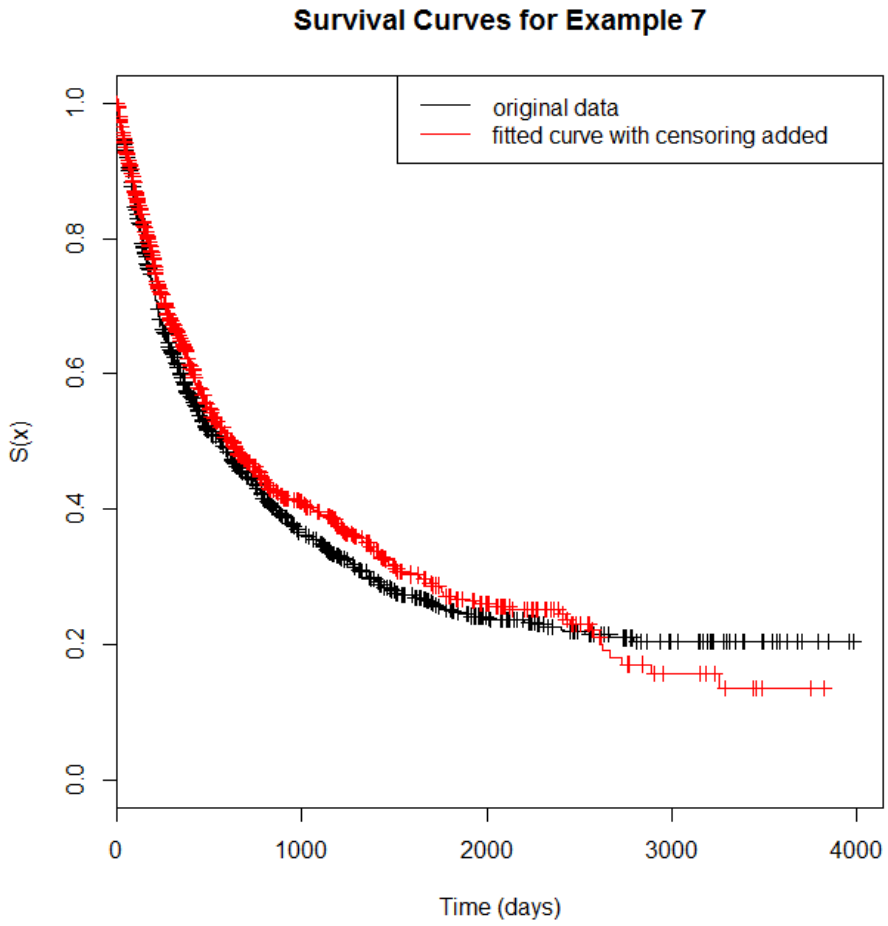
(a) histogram displaying the proportion of observed uncensored cases in each survival time interval.

(b) plot of estimated censoring probability against survival time with margins of error



survival time, we generated data with censoring from a mixture of exponential distributions with $p = 0.44$, $\lambda_F = 1/315$, and $\lambda_S = 1/(7.5 * 365)$ to compare to the simulated data from example 7. This comparison is plotted in Figure 3.10. Our estimated curve (red) matches up fairly well with the simulated sample data (black) for the first 2,500 days, though there are larger differences between the estimated and observed survival probabilities for more extreme survival times. Such a discrepancy is likely driven by the 33% relative error we made in estimating λ_S. Again, when the probability of censoring is so high for longer survival times resulting in very few observed uncensored survival times, it can be difficult to estimate λ_S using our graphical procedure, so the estimate of $\lambda_S$may need to be artificially adjusted when plotting the estimated survival function against the

Figure 3.10: comparing the survival function estimated by our procedure with censoring added (red) to the Kaplan-Meier survival curve for the simulated control arm clinical durations (black) in example 7.



Kaplan-Meier curve for the simulated sample data in order to improve the fit from the estimated mixture of exponential distributions.

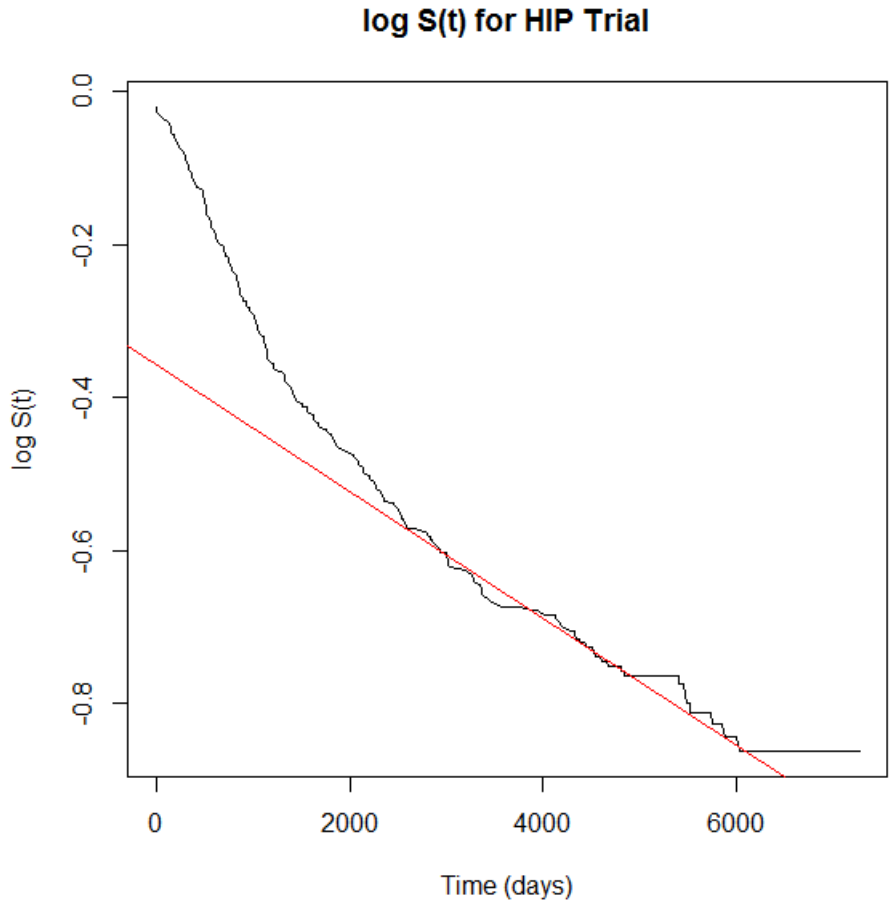
3.5 Using Data from the HIP and PLCO Trials

In the simulated examples of Sections 3.2 and 3.4, our procedures to estimate the parameters in a mixture of two exponential distributions and estimate censoring probability as a function of survival time both proved to be reliable in most scenarios. If the rate of censoring is high, we may underestimate the exponential rate parameters,

specifically λ_S when the rate of censoring increases with survival time. However, this error can be adjusted by manually increasing λ_S when plotting the estimated survival function against the Kaplan-Meier curve for the simulated sample data.

Now that we have tested the application of our procedures, we want to return to the HIP trial data discussed at the very beginning of Section 3 and see how well a mixture of two exponential distributions fits to the survival time data recorded in the HIP trial. Recall that the HIP trial data includes 437 cases in the control group at the catch-up point, and there is a heavy rate of censoring (47%). To begin our procedure, we calculated the Kaplan-Meier estimated survival curve and plotted $\log \hat{S}(t)$ against time, as shown in Figure 3.11. We considered times to death longer than 4,000 days to be the largest observed, and we fit a line to the plot with an intercept of $\log 0.7$ and a slope of $1/(33 * 365)$, giving initial estimates of $\hat{p} = 0.3$ and $\hat{\lambda}_S = 1/(33 * 365)$. In step 2 of our procedure for parameter estimation, we selected $t^* = 500$ days, and calculated $\hat{S}(500) = 0.87$ from the sample data. We solved for $\hat{\lambda}_F = 1/1165$ using the initial estimates for $\hat{p}$ and $\hat{\lambda}_S$. In step 3, we determined that, if $\lambda_F = 1/1165$, 95% of the fast-developing cases should end within the first 3,490 days from diagnosis. This new domain of survival times to estimate p and λ_S (times longer than 3,490 days) is not very different from the domain we used in step 1 (times longer than 4,000 days) to get the initial estimates $\hat{p}$ and $\hat{\lambda}_S$. Revisiting Figure 3.11, the line we fit to the relationship between log-transformed survival probability and time does not change when we try to fit this line in $\{t > 3490\}$ instead of $\{t > 4000\}$. Thus, our procedure stops after just one iteration, and our parameter estimates are $\hat{p} = 0.3$, $\hat{\lambda}_F = 1/1165$, and $\hat{\lambda}_S = 1/(33 * 365)$. However, we know that our procedure may underestimate λ_S when there is heavy censoring for large survival

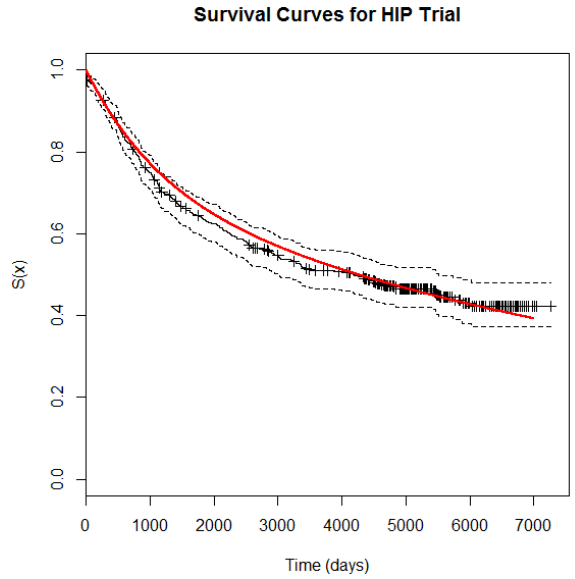
Figure 3.11: plotting $\log \hat{S}(t)$ against time for the HIP trial data and fitting a line (red) in the domain of the largest observed times to death.



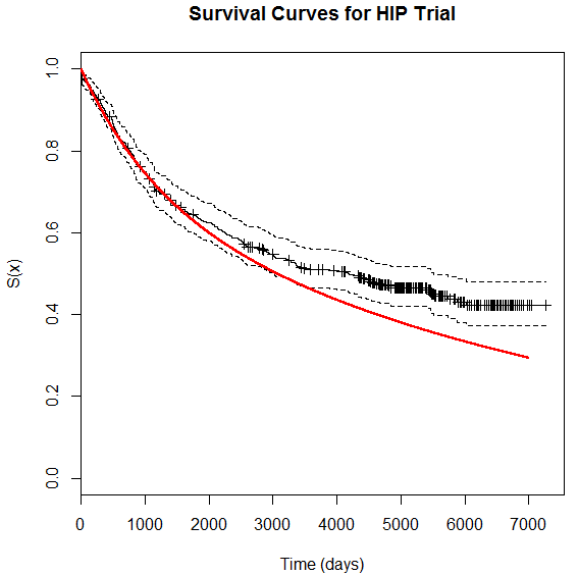
times, so we want to plot our estimated distribution against the Kaplan-Meier curve for the HIP trial data. We will consider this plot prior to estimating the function for assigning censoring probabilities, as we do not want to use an erroneous value of λ_S to estimate said probabilities. Figure 3.12a shows the Kaplan-Meier curve for the HIP trial data overlaid with our estimated mixture distribution. We see that our estimated distribution does not capture the majority of the Kaplan-Meier curve well, especially not between 1,000 and 4,000 days, because it is overfitting the trend in the heavily censored data beyond 4,500 days. We want to increase our estimate of λ_S to better capture the trend prior to the

Figure 3.12: comparing the survival curve fit by our procedure (red) to the Kaplan-Meier survival curve for the control arm clinical durations (black) in the HIP trial data.

(a) using the initial value of $\lambda_S = 1/(33 * 365)$.



(b) using the adjusted value of $\lambda_S = 1/(22 * 365)$.

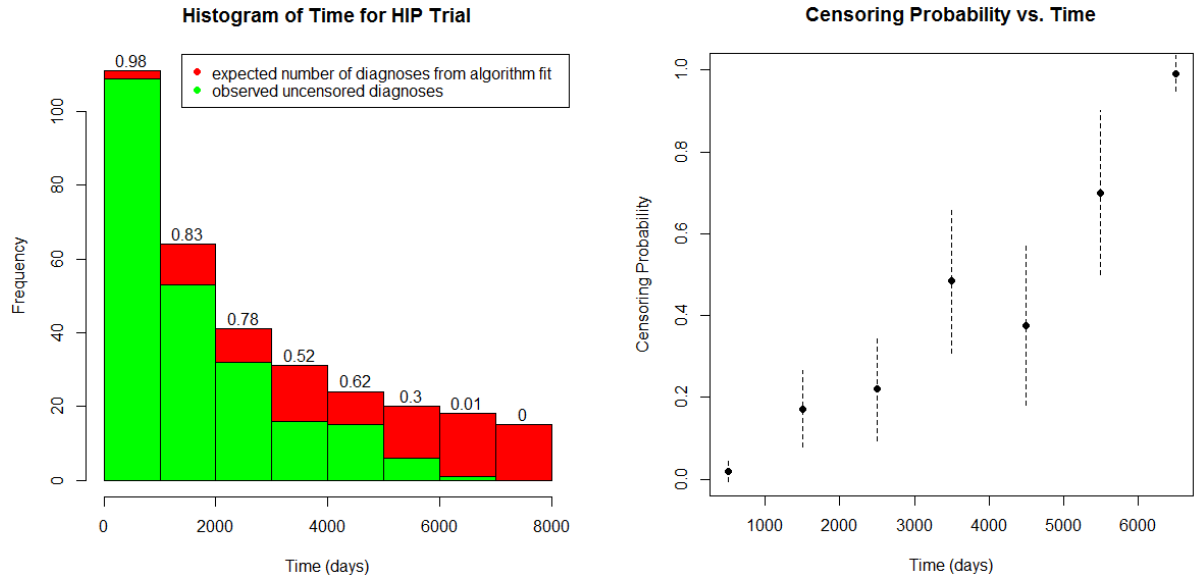


heavily censored survival times. By changing $\hat{\lambda}_S = 1/(22 * 365)$, we obtain the plot shown in Figure 3.12b. The updated distribution estimate matches with the Kaplan-Meier curve for the HIP trial data very well for the first 2,000 days when the rate of censoring is low, and the differences between the estimated distribution and the Kaplan-Meier curve beyond 2,000 days may likely be explained by the Kaplan-Meier curve flattening out with so many censored survival times and so few observed uncensored cases. To verify this explanation, we need to add censoring to our estimated mixture distribution of survival times, which requires us to now estimate censoring probability as a function of survival time. Applying the procedure discussed in Section 3.3, we divide the domain of the HIP trial survival times into intervals of width $w = 1{,}000$ days, counting the number of observed uncensored survival times in each interval and calculating the expected number of cases using Equation

Figure 3.13: estimating censoring probability as a function of survival time for the HIP trial data.

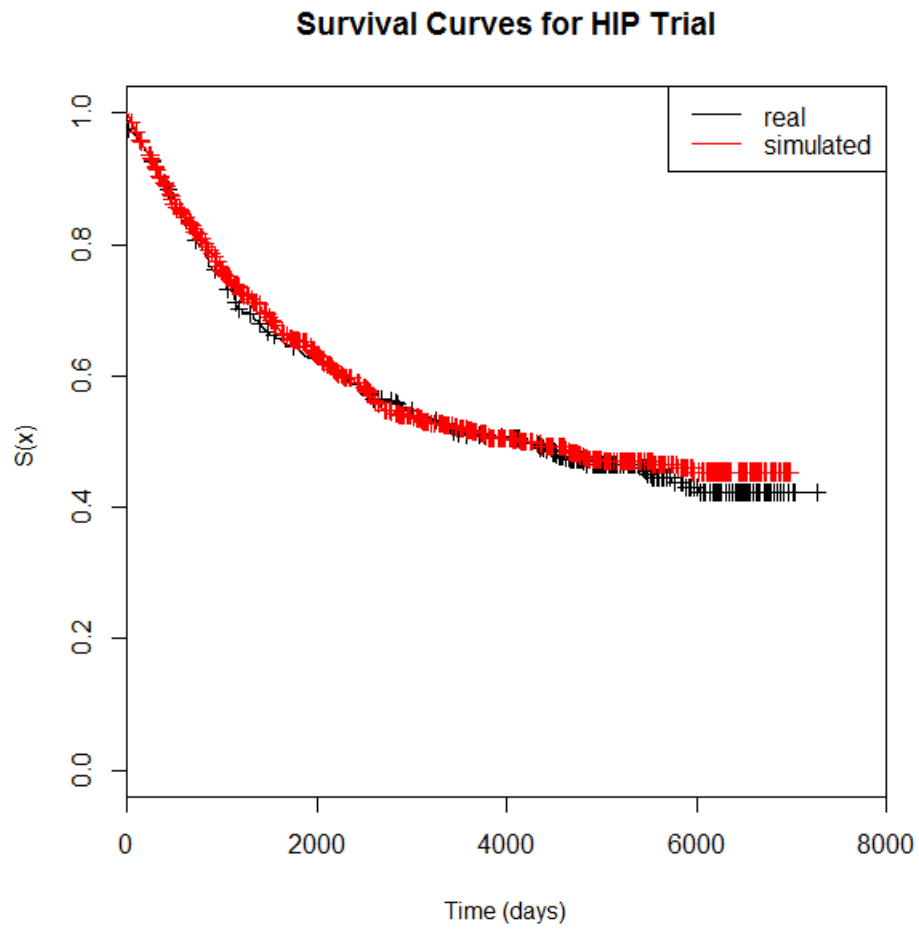
(a) histogram displaying the proportion of observed uncensored cases in each survival time interval.

(b) plot of estimated censoring probability against survival time with margins of error



(3.9), where $S(x)$ is the survival function for a mixture of exponential distributions with $p = 0.3$, $\lambda_F = 1/1165$, and $\lambda_S = 1/(22 * 365)$. Figure 3.13a compares the number of observed uncensored cases to the expected number of cases in each interval, and Figure 3.13b plots the censoring probability against time with 95% margins of error added to each interval's estimated censoring probability. In Figure 3.13b, the probability of censoring appears to increase linearly with survival time over the first 6,000 days, reaching practically 100% in the $[6000, 7000)$ interval of survival times. We chose to model the censoring probability as a linear function that increases from 0% for a survival time of zero days to 100% for a survival time of 6,000 days. Additionally, for simulated survival times from our estimated mixture of exponential distributions that were probabilistically

Figure 3.14: comparing the survival function estimated by our procedure with censoring added (red) to the Kaplan-Meier survival curve for the control arm clinical durations (black) in the HIP trial.



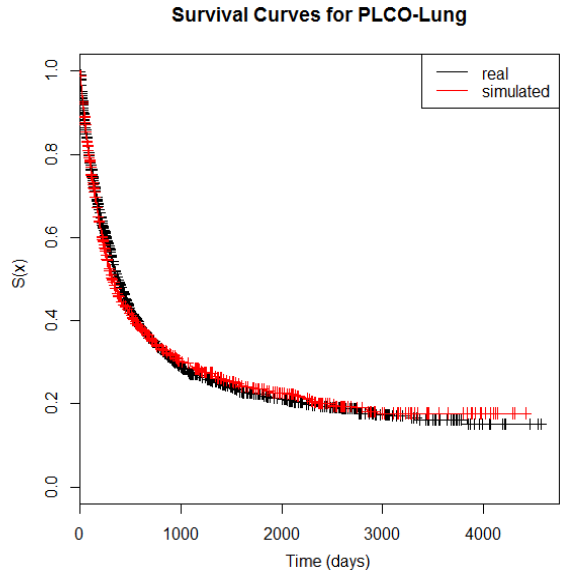
selected to be censored, we decided to cap them at 7,000 days (roughly the longest observed censored survival time in the HIP trial) and rescale using a random observation from beta(2,1) variable to simulate the time at which censoring occurs. Figure 3.14 compares simulated data with censoring from a mixture of two exponential distributions with $p = 0.3$, $\lambda_F = 1/1165$, and $\lambda_S = 1/(22 * 365)$ to the Kaplan-Meier survival curve for the HIP trial data. Because the two survival curves overlap so well, it is clear that a mixture of two exponential distributions is an appropriate model for the survival times in the HIP

trial, and our procedures have done a remarkable job of estimating the parameters of this mixture distribution in the presence of heavy censoring.

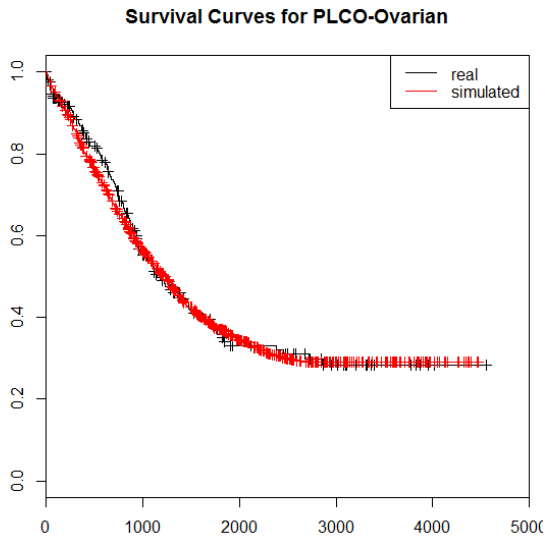
Although the mixture of two exponential distributions is a good model for the survival times from diagnosis in the HIP trial, we wanted to confirm that this type of distribution would also be a good fit for survival times from other cancer types, as well. We decided to look at lung and ovarian cancer survival times from the PLCO [16]. There were 1,719 lung cancer diagnoses in the control arm of the PLCO, and approximately 31% of these survival times were censored. There were also 209 ovarian cancer diagnoses, of which 47% had censored survival times. Without going into repetitive detail about the application of our estimation procedures, I will simply present the results to assess how well a mixture of two exponential distributions fits to the survival times from these screening trials. For the lung cancer survival times in the PLCO, we fit a mixture of exponential distributions with $p = 0.6$, $\lambda_F = 1/238$, and $\lambda_S = 1/(6 * 365)$. We found that the probability of a case being observed uncensored decreased as a quadratic function with survival time, similar to the function assigning censoring probabilities to cases in example 7. Figure 3.15a compares simulated data from our fitted mixture of two exponential distributions with censoring to the Kaplan-Meier survival curve for the PLCO lung cancer data. For the ovarian cancer survival times in the PLCO, we fit a mixture of exponential distributions with $p = 0.6$, $\lambda_F = 1/1482$, and $\lambda_S = 1/(10 * 365)$. We again found that the probability of a case being observed uncensored decreased as a quadratic function with survival time. Figure 3.15b compares simulated data from our fitted mixture of two exponential distributions with censoring to the Kaplan-Meier survival curve for the PLCO ovarian cancer data. For both cancer types, lung and ovarian, a mixture of two exponential

Figure 3.15: comparing the survival curve fit by our procedure with censoring added (red) to the Kaplan-Meier survival curve for the control arm clinical durations (black) in the PLCO.

(a) lung cancer survival times.



(b) ovarian cancer survival times.



distributions appears to be a good fit for the survival times from diagnosis, and our procedures have once again appropriately specified the parameters of these models.

3.6 Discussion

Section 3 began with the challenge of fitting a distribution to the survival times from diagnosis in a randomized controlled screening trial. When it became clear that a single common distribution could not provide a good approximation to the Kaplan-Meier curve calculated from the sample survival times, we proceeded to consider a mixture of two distributions to try to find an appropriate fit. In Section 3.1, we developed an iterative procedure to estimate the rate parameters and component proportions in a mixture of two exponential distributions. After the simulations in Section 3.2 illustrated that heavy

censoring led to larger relative errors in the parameter estimates, we devised another procedure in Section 3.3 to simulate censoring in a way that would mimic the pattern of censoring in the real data. The simulations in Section 3.4 showed the accuracy of our procedures to estimate the parameters in a mixture of two exponential distributions and estimate censoring probability as a function of survival time, so we applied them to the survival times from three historic screening trials with different cancer types in Section 3.5: the HIP trial (breast cancer), PLCO-Lung, and PLCO-Ovarian. In all three examples, a mixture of two exponential distributions provided a good fit to the distribution of clinical durations.

In most of the simulated examples, the relative errors of the parameter estimates by our procedure were fairly small, no more than 10-15% at most. However, our procedure did struggle to accurately estimate the mixture distribution parameters in the presence of heavy censoring. There was, at least, a pattern to these errors, as the exponential rate parameters were consistently underestimated when heavy censoring flattened out the Kaplan-Meier survival curve calculated from the sample data. With the nature of the error known, the parameter estimates from our procedure could be systematically increased accordingly so that the estimated survival curve would match more closely to the Kaplan-Meier curve for the sample data when plotted together. Moreover, when the censoring probability was low for shorter survival times and increased for longer survival times, only the rate parameter of the slow-growing cases λ_S seemed to be significantly underestimated; both other parameters, p and λ_F, were more accurately specified. Given that our procedure is so simple and quick to use, especially compared to the common approach for parameter estimation in a mixture distribution, the E-M algorithm, such easily

adjustable errors may be tolerated. Furthermore, the E-M algorithm is not without flaws of its own, some of which were highlighted by Seidel et al [44]; the predictable and correctable errors from our procedure may be preferable to those of E-M estimation.

While survival times in randomized controlled screening trials were the motivation for the development of the estimation procedures for a mixture of two exponential distibutions discussed in Section 3, this mixture distribution may be applicable to various other data types, and our procedures can still be used to estimate the parameters. An example we have considered is the reliability of a population of components, such as iPhones or machinery. A mixture of two exponential distributions may be useful if some components inexplicably fail rather quickly while others last for longer periods of time. Fitting a mixture of two exponential distributions to such failure times and using our procedure to estimate the parameters would allow for calculations about the reliability of the component, such as the mean failure time or the probability of failing within a certain amount of time. Additionally, if such failure time data has a lower rate of censoring, as the censoring rate of up to 50% in screening trials is rather high, then there would be minimal concern regarding the accuracy of the estimated distribution from our procedure, as illustrated by simulated examples 1-4 in Section 3.2 with little or no censoring.

Finally, because a mixture of exponential distributions is such a good fit to the control arm clinical durations in cancer screening data, and because of the plausible strong correlation between preclinical and clinical durations, we believe that the mixture of exponential distributions should also be an appropriate model for cancer preclinical durations. Preclinical durations are obviously unobservable, as they refer to a period of time in which cancer is growing undetected, so having an idea about the type of

distribution to fit to the preclinical durations is very important step. Moreover, there may be information in the pattern of diagnoses, such as the screen at which they occur and the prevalence of interval cases, that sheds some light on the parameters of the distribution of preclinical durations. Section 5 will discuss this idea further, as well as an idea for estimating these unobserved preclinical duration distribution parameters.

4. Modelling Screening Test Sensitivity

At the end of Section 3, we presented the idea of fitting a mixture of two exponential distributions to preclinical durations. This mixture distribution provided an appropriate fit to the clinical durations observed in three historic screening trials, and the plausible strong correlation between a subject's preclinical and clinical duration suggests that the same type of distribution could be fit to preclinical durations. We also proposed that the pattern of diagnoses in a screening trial may provide information about the parameters of the distribution of preclinical durations.

We wanted to design a simulation study to evaluate this idea accounting for several factors that arise in screening trials. First, even if we only focus on simulating diagnoses and ignore the subsequent post-diagnosis survival time, we still need to generate preclinical durations, define a screening program, and test each case at times specified by the screening program with some sensitivity, or probability of a successful diagnosis. While the screening program for any randomized controlled trial is known based on the study design, the test sensitivity, like the distribution of preclinical durations, is unknown. Moreover, the test sensitivity, like the distribution of preclinical durations, has an effect on the pattern of diagnoses in the trial. For example, a low mean lead time, which implies that screened subjects are not diagnosed much earlier than their unscreened counterparts, could be explained by either short preclinical durations or a low screening test sensitivity. In reality, there is probably an interactive effect between preclinical duration and test sensitivity on the pattern of diagnoses [36]. Thus, before trying to estimate preclinical duration parameters based on the diagnosis pattern in a screening trial, it is important to assess the extent to which the test sensitivity will also affect this pattern.

The literature, as discussed in Section 2, focuses primarily on estimating the overall sensitivity of a screening test. However, test sensitivity is unlikely to be constant; sensitivity of an individual screen may depend on several factors. While factors like age and comorbidities could plausibly affect sensitivity, a major factor is likely to be progression into the disease, as the test may be less sensitive very early in the preclinical phase and become more sensitive as the preclinical phase progresses to the point of clinical detection. This idea was discussed in detail by Heltshe et al [12]. For a given preclinical duration y, Heltshe et al allowed the test sensitivity to increase monotonically throughout y using either a uniform, beta, or Normal cdf. Sensitivity began at 50% at the start of the preclinical duration and converged to 100% roughly halfway through the preclinical duration. Sensitivity probably will increase as a function of elapsed preclinical duration, but the model from Heltshe et al that starts at 50% when the preclinical duration begins may be overly optimistic. A test sensitivity of 50% at the start of the preclinical duration suggests that half of cancers can be immediately detected by screening as soon as the first cancer cells begin to develop. Also, sensitivity that reaches 100% midway through the preclinical duration suggests a fabulous screening test that guarantees cancer detection in only half the time it would take for said cancer to be diagnosed by symptoms. The relationship between test sensitivity and stage of the preclinical duration specified by Heltshe et al may not be biologically plausible.

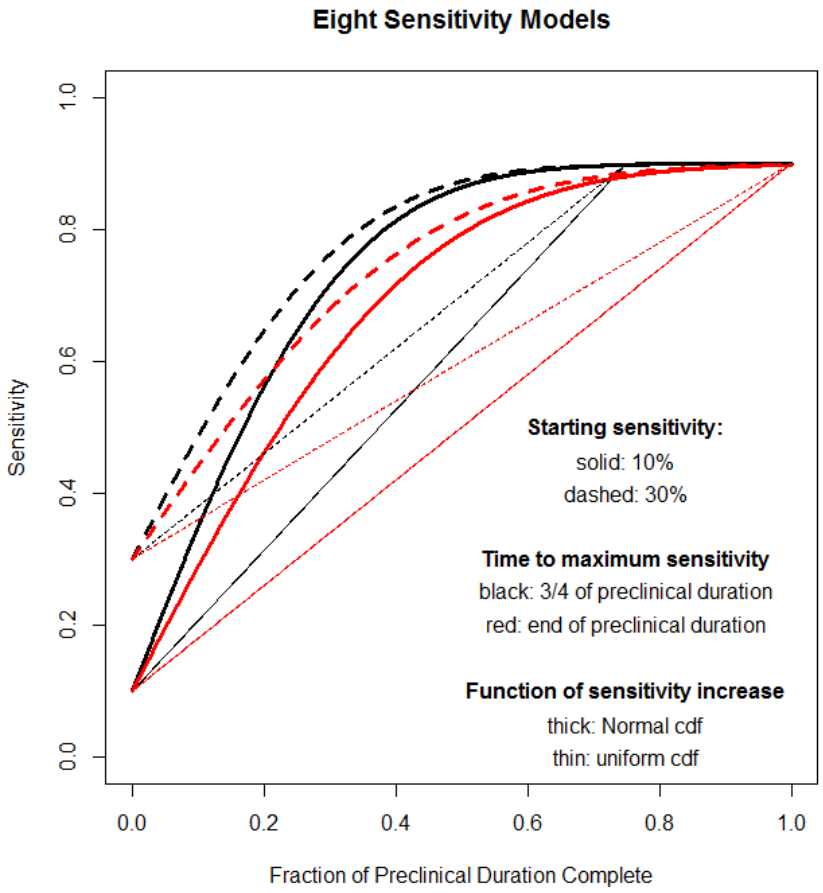
Because test sensitivity as a function of the preclinical growth cannot be estimated from sample data, simulations must be used to assess the effects of different sensitivity models on the pattern of diagnoses in a trial. In general, if we are assuming that test sensitivity is a function of the preclinical growth, the model for sensitivity should account

for four attributes: (i) the starting sensitivity at the initiation of the preclinical duration, (ii) the maximum achievable sensitivity of the test, (iii) the time at which this maximum sensitivity is achieved, and (iv) the functional form of the increase in test sensitivity from starting sensitivity to maximum. Of these four attributes, the maximum achievable sensitivity is the one that can be most accurately specified based on real data. Many studies have calculated the true positive rate for a screening test in detecting cancer among known cancer patients; for example, Cologuard© advertises a 92% sensitivity for detecting known cancers [22]. This suggests that the maximum achievable sensitivity in a sensitivity model should be fixed somewhere around 90%. However, the other three attributes – starting sensitivity, time at which the maximum sensitivity is achieved, and method of sensitivity increase – cannot be estimated without lab experiments to carefully evaluate test sensitivity dependence on preclinical growth. Therefore, we vary these three attributes in a simulation study to assess the effects of sensitivity on the pattern of diagnoses in a trial.

4.1 Simulation Study

We designed a simulation study to determine how a given model for test sensitivity affects the pattern of diagnoses in a randomized controlled screening trial. When modelling the test sensitivity as a function of the subject's preclinical growth, or the fraction of the preclinical duration completed at the time of the screening test, we varied three attributes of the model. The maximum achievable sensitivity could reasonably be fixed at 90%, but the starting sensitivity, time at which the maximum sensitivity is achieved, and method of sensitivity increase are not as easily assumed and need to be varied so that their effect on diagnosis times can be quantified.

Figure 4.1: plotting the eight sensitivity models in our 2^3 factorial experiment.



To assess the significance of these three variables, we designed a 2^3 factorial experiment. A 2^3 factorial experiment has three factors with two levels each, resulting in a total of eight runs per replicate. For our three attributes, the levels were as follows: starting sensitivity of either 10% or 30%, maximum sensitivity achieved at either three-quarters of the preclinical duration or the end of the preclinical duration, and sensitivity increasing by either a uniform or Normal cdf. The eight possible sensitivity models are plotted in Figure 4.1. When the method of sensitivity increase follows a uniform cdf, test sensitivity increases linearly as the preclinical duration progresses from starting sensitivity to 90% at the time at which maximum sensitivity is achieved. When the method of

sensitivity increase follows a Normal cdf, the Normal cdf function between the 50th and 97.5th percentiles is rescaled to fit the increase from starting sensitivity to 90% at the time at which maximum sensitivity is achieved. Because the Normal cdf is a concave down function in this range, this method of sensitivity increase suggests diminishing marginal returns of additional preclinical time passing on test sensitivity.

For each replicate of our 2^3 factorial experiment, preclinical durations were simulated from a predetermined mixture of two exponential distributions. The time between the initiations of two preclinical durations was generated as a Poisson process at a rate of 210 new cancers per year. In an actual trial, this rate would be determined by the study size and the population rate of diagnoses (i.e. m new cases each year per 100,000 at-risk people), but we found that changing this rate in a simulated trial has no effect on the study outcomes (results not pictured). Any case whose preclinical duration overlaps with our screening window is included in the data, which means that cases whose preclinical durations began before the screening window are included as long as the preclinical duration continues beyond the time of the first screen. For each simulated case, an identical observation was assigned to the treatment and control arm, presupposing no overdiagnosis. Because the preclinical durations and the times at which they began are both random, the number of simulated cases whose preclinical durations overlap with our screening program will not be exactly the same for each replicate, though the counts should be similar.

Our experiment was designed with four annual screens. Screening occurred at the start of the simulated screening trial, as well as at the end of the first, second, and third years. Our screening process followed the outline for the design of a computer simulation

of randomized controlled screening trial presented in Figure 2 of Kafadar and Prorok [46]. For any case j, our screening program began by checking whether this case was in the middle of its preclinical duration at the time of the first screen and, if so, testing for cancer with a sensitivity β specified by one of our eight sensitivity models. If a diagnosis were made, the time of the first screen was recorded as the time of diagnosis, and the lead time could be calculated as the difference between this time of diagnosis and the time at which the preclinical duration would have ended in the absence of screening. If no diagnosis was made, we proceeded to the next screen and repeated this process of testing for cancer. If the preclinical duration ended prior to this next screen, then the time at which the preclinical duration ended was recorded as the time of diagnosis, and the lead time was calculated to be zero, as this represents an interval case. This screening process was repeated eight times for case j so that each of our eight screening models could be used to carry out testing.

Our experiment was performed with various preclinical duration distribution parameters, as we expect that the preclinical duration and test sensitivity have an interactive effect on the pattern of diagnoses in a screening trial. We will present the results of three different example scenarios. Example 1 simulated 50% fast-developing cases with a mean preclinical duration of one year and 50% slow-developing cases with a mean preclinical duration of five years: $p = 0.5$, $\lambda_F = 1/365$, and $\lambda_S = 1/(5 * 365)$. Example 2 simulated a very slow distribution with $p = 0.25$, $\lambda_F = 1/(2 * 365)$, and $\lambda_S = 1/(7 * 365)$. Example 3 simulated a very fast distribution with $p = 0.75$, $\lambda_F = 1/(0.75 * 365)$, and $\lambda_S = 1/(3 * 365)$.

For each example scenario, we repeated the experiment 500 times. In each replicate, preclinical durations were generated from the specified distribution, and screening was simulated eight times based on our eight sensitivity models. The result of each replicate was a distribution of the times of diagnoses under each sensitivity model, as well as an estimate for the mean lead time, which is a useful proxy to measure how early the diagnoses were made under that sensitivity model. 500 replicates were performed to ensure the accuracy and precision of these estimates.

Figure 4.2 presents the mean lead time and distribution of diagnoses for example 1, averaged out across the 500 replicates. From these counts, it seems clear that the sensitivity model does have some effect on the pattern of diagnoses in a screening trial. A higher starting sensitivity, a shorter time to achieve the maximum sensitivity of 90%, and a more quickly increasing function (the Normal cdf) as the preclinical duration progresses all seem to result in a longer mean lead time and more diagnoses made from screening rather than interval and post-screening cases.

To determine the significance of these results for example 1, we performed an analysis of variance (ANOVA) on the mean lead time variable. The ANOVA model included the starting sensitivity, time at which the maximum sensitivity is achieved, and method of sensitivity increase, as well as all two- and three-way interactions. The replicate of the experiment corresponding to each mean lead time estimate was also included in the model as a blocking factor, as each of our eight sensitivity models was applied once per replicate. Figure 4.3 displays the ANOVA table for this analysis. From our simulation study under the preclinical duration conditions of example 1, all three attributes, as well as all two-way interactions and the replicate blocking factor, have a significant effect on the mean lead

Figure 4.2: Means and standard deviations for the number of diagnoses at various times in the simulated screening trial for example 1.

Sens. Start	Sens. Increase	Sens. Max.	Mean Lead Time	S1	Y1	S2	Y2	S3	Y3	S4	Post-Screen
0.1	Uniform	0.75	1.94 (0.11)	275 (16)	81 (9)	170 (13)	75 (9)	147 (12)	75 (9)	139 (11)	171 (13)
0.1	Uniform	End	1.74 (0.10)	228 (15)	96 (10)	159 (13)	88 (10)	138 (12)	87 (9)	129 (12)	207 (15)
0.3	Uniform	0.75	2.38 (0.12)	338 (18)	76 (9)	184 (13)	70 (8)	152 (12)	70 (9)	143 (11)	118 (11)
0.3	Uniform	End	2.27 (0.12)	283 (17)	88 (9)	179 (14)	79 (9)	150 (11)	78 (9)	138 (12)	137 (12)
0.1	Normal	0.75	2.19 (0.11)	309 (17)	77 (9)	175 (14)	71 (8)	149 (11)	71 (9)	142 (12)	138 (12)
0.1	Normal	End	2.01 (0.11)	272 (16)	86 (10)	171 (13)	78 (9)	144 (11)	78 (9)	138 (13)	165 (13)
0.3	Normal	0.75	2.52 (0.12)	346 (19)	73 (9)	182 (13)	66 (8)	153 (12)	67 (8)	146 (12)	98 (10)
0.3	Normal	End	2.42 (0.12)	318 (17)	80 (9)	182 (13)	72 (9)	152 (12)	72 (9)	142 (12)	115 (11)
0.1	Unif	0.75	1.94 (0.11)	275 (16)	81 (9)	170 (13)	75 (9)	147 (12)	75 (9)	139 (11)	171 (13)

Abbreviations: Sens., sensitivity; S1, first screen; Y1, first year interval cases; S2, second screen; Y2, second year interval cases; S3, third screen; Y3, third year interval cases; S4, fourth screen.

Figure 4.3: Analysis of variance results example 1.

Factor	df	SS	MS	F	p-value
Replicate	499	35.4	0.1	19.4	<0.0001
Sens. start	1	184.1	184.1	50,189.9	<0.0001
Sens. increase	1	39.3	39.3	10,712.4	<0.0001
Sens. max.	1	23.0	23.0	6,270.8	<0.0001
Sens. start ∗ sens. increase	1	3.7	3.7	1,005.3	<0.0001
Sens. start ∗ sens. max.	1	2.0	2.0	546.9	<0.0001
Sens. increase ∗ sens. max.	1	0.1	0.1	25.9	<0.0001
Sens. start ∗ sens. increase ∗ sens. max	1	0.0	0.0	0.0	0.925
Error	3,497	12.8	0.0		

Abbreviations: df, degrees of freedom; SS, sum of squares; MS, mean squares; F, F-statistic; Sens., sensitivity.

time of the simulated trial. Of the three attributes of the sensitivity model, starting sensitivity has the greatest effect on the run's mean lead time, followed by the method of increase and, finally, the time at which maximum sensitivity is achieved. One explanation for the significance of so many model effects is the extremely large sample size of this experiment; 500 replicates times eight runs per replicate (one for each sensitivity model) results in a total sample size of 4,000 simulated screening trials. Even with all possible interactions and the replicate blocking factor included in the ANOVA model, the residual degrees of freedom are still extremely high, resulting in low mean squared error, which makes smaller attribute effects appear more significant. We know that the replicate of our experiment should not have a significant effect on mean lead time, as the sample of preclinical durations are generated from the exact same distribution for each replicate, yet this blocking factor is still highly statistically significant. With the knowledge that the replicate of our experiment should not have a meaningful effect on mean lead time, perhaps we could conclude that any effect with a lower sum of squares than the blocking

Figure 4.4: Analysis of variance results for examples 2 and 3.

(a) example 2

Factor	df	SS	MS	F	p-value
Replicate	499	62.1	0.1	18.7	<0.0001
Sens. start	1	632.6	632.6	95,236.8	<0.0001
Sens. increase	1	112.1	112.1	16,890.2	<0.0001
Sens. max.	1	64.0	64.0	9,630.1	<0.0001
Sens. start ∗ sens. increase	1	12.7	12.7	1,905.0	<0.0001
Sens. start ∗ sens. max.	1	6.7	6.7	1,008.5	<0.0001
Sens. increase ∗ sens. max.	1	0.3	0.3	47.8	<0.0001
Sens. start ∗ sens. increase ∗ sens. max	1	0.1	0.1	20.0	<0.0001
Error	3,497	23.2	0.0		

(b) example 3

Factor	df	SS	MS	F	p-value
Replicate	499	9.8	0.0	21.5	<0.0001
Sens. start	1	21.8	21.8	23,909.5	<0.0001
Sens. increase	1	7.1	7.1	7,760.2	<0.0001
Sens. max.	1	4.1	4.1	4,505.3	<0.0001
Sens. start ∗ sens. increase	1	0.3	0.3	350.5	<0.0001
Sens. start ∗ sens. max.	1	0.2	0.2	265.0	<0.0001
Sens. increase ∗ sens. max.	1	0.1	0.1	6.6	0.01
Sens. start ∗ sens. increase ∗ sens. max	1	0.0	0.0	2.7	0.10
Error	3,497	3.2	0.0		

Abbreviations: df, degrees of freedom; SS, sum of squares; MS, mean squares; F, F-statistic; Sens., sensitivity.

factor is not really important when specifying a model for test sensitivity as a function of preclinical growth.

The same ANOVA model was applied to the experimental data for example 2, and the results are shown in Figure 4.4a. All three attributes, as well as all interactions and the replicate blocking factor, have a significant effect on the mean lead time of the simulated trial. Once again, of the attributes of the sensitivity model, starting sensitivity has the greatest effect on the run's mean lead time, followed by the method of increase and the

time at which maximum sensitivity is achieved. Compared to example 1, we see that the total sum of squares is much higher when the distribution of preclinical durations is slower, which means that there is greater variability in the mean lead times across simulated screening trials. Of this increased variability, some is explained by increased replicate-to-replicate differences in mean lead times, but the relative increase in sum of squares is much higher for the effects pertaining to the sensitivity model attributes. This suggests that, when the preclinical durations are long relative to the screening interval, the changes in test sensitivity have a much greater effect on the pattern of diagnoses in the trial.

Our ANOVA model was also applied to the experimental data for example 3, with the results displayed in Figure 4.4b. All three attributes, as well as a pair of two-way interactions and the replicate blocking factor, have a significant effect on the mean lead time of the simulated trial. Just as in the previous two example scenarios, starting sensitivity has the greatest effect on the run's mean lead time, followed by the method of increase and the time at which maximum sensitivity is achieved. However, compared to examples 1 and 2, the total sum of squares is much lower when the distribution of preclinical durations is faster, which means that there is less variability in the mean lead times across the 4,000 simulated screening trials. Moreover, in the results from example 3, the method of sensitivity increase and the time at which maximum sensitivity is achieved both explain less of the variability in mean lead time across runs than the replicate blocking factor, suggesting that these two attributes may not be important when specifying a model for test sensitivity as a function of preclinical growth if the preclinical durations are short.

4.2 Discussion

We concluded Section 3 with an idea about fitting a mixture of exponential distributions to preclinical durations and predicting the parameters for various cancers based on the pattern of diagnoses in a randomized controlled screening trial. Recognizing that the sensitivity of a test, in addition to the distribution of preclinical durations for that cancer type, also affects the pattern of diagnoses in a trial, we set out in Section 4 to design a simulation study to identify the attributes of a sensitivity model that impact diagnosis times. Defining test sensitivity as a function of preclinical growth, we designed in Section 4.1 a 2^3 factorial experiment that varied the starting test sensitivity, the time at which maximum test sensitivity of approximately 90% is achieved, and the functional form of sensitivity increase with preclinical growth. While all three attributes of the sensitivity model had at least some effect on the mean lead time of a cancer diagnosis, starting sensitivity by far had the greatest effect on the screening trial's mean lead time estimate, followed by the method of increase and, finally, the time at which maximum sensitivity is first achieved.

One of the important findings from our three experiments was that the relative significance of the sensitivity model attributes depended on the distribution of preclinical durations specified. When the preclinical durations were very fast, only the starting sensitivity had a greater effect on the mean lead times observed in the simulated screening trials than the replicate of the experiment. Because we know that the preclinical durations simulated come from the same distribution for each replicate, this blocking factor should not have a meaningful impact on the mean lead time, so any sensitivity model attribute with a lower sum of squares than the experimental replicate could be viewed as

unimportant. As the preclinical durations got slower, the functional form of sensitivity increase with preclinical growth gained in relative significance, as did the time to maximum sensitivity. However, this increasing relative significance of sensitivity model attributes may be achieved in part because the screening interval was held constant across all of our simulated experiments even as the distribution of preclinical durations changed. The mean preclinical duration was 5.75 years in example 2 and 1.3125 years in example 3, yet subjects were screened annually in both cases. In reality, we believe that a cancer as slow-developing as the one simulated in example 2 likely would not be screened annually, as there is no need to screen so frequently; slow-developing cancers – colorectal cancer is a familiar example – can be screened less frequently while still diagnosing the disease sufficiently early to see a treatment benefit. The two experiments for which annual screening makes more sense given the distribution of preclinical durations, examples 1 and 3, found that the starting sensitivity was the only sensitivity model attribute that had a highly important effect on the pattern of diagnoses.

Our results are fairly consistent with the findings of Heltshe et al. Recall that Heltshe et al allowed the test sensitivity to increase from 50% at the start to 100% roughly halfway through the preclinical duration, with this increase taking the form of either a uniform, beta, or Normal cdf. Although their study assessed the impact of sensitivity model on the length-biased sampling effect $E[Y^*]/E[Y]$ instead of its impact on the mean lead time, Heltshe et al found minimal difference between the results from the uniform and Normal sensitivity models [12]. Similarly, we found that the functional form of sensitivity's increase with preclinical growth had a minimal effect on the mean lead time observed in a simulated screening trial. Little-to-no change in the mean lead time between two trials

with the same distribution of preclinical durations suggests that the similar cases are being diagnosed at roughly the same times in both trials. Because we are comparing two trials with the same model for preclinical durations, the mean preclinical duration $E[Y]$ remains constant. Furthermore, if similar cases are being diagnosed at roughly the same times in both trials, then the mean screen-detected preclinical duration $E[Y^*]$ should also remain constant. Therefore, our result – that the method of sensitivity increase has minimal impact on mean lead time – intuitively implies Heltshe et al's result – that the method of sensitivity increase has minimal impact on the length-biased sampling effect. We will further assess the influence of sensitivity model on the length-biased sampling effect ourselves in Section 6.

The conclusions drawn from Section 4 will prove useful as we progress to predicting preclinical duration distribution parameters from the pattern of diagnoses, which is carried out in Section 5. Our results suggest that only the initial sensitivity at the start of the preclinical duration will have a notable effect on the pattern of diagnoses observed. Unfortunately, the starting test sensitivity and the preclinical duration parameters are confounded in the pattern of diagnoses. Because the true sensitivity model cannot be specified, the preclinical duration parameters cannot be specified with certainty either. For example, as we discussed at the beginning of Section 4, a low mean lead time could be explained by either short preclinical durations or low starting test sensitivity. Thus, when we predict the preclinical duration distribution parameters for a certain cancer type based on the pattern of diagnoses in a screening trial, we will have to specify a sensitivity model that we presume to be reasonably accurate, but still recognize that our parameter estimates could be erroneous due to a misspecification of the sensitivity model. This will

require us to ask two questions. First, how great of a mistake could we make in estimating these parameters by misspecifying the sensitivity model? Second, to what extent does this error in defining the preclinical duration parameters and sensitivity model impact the length-biased sampling effect on a given screening trial, as this is the ultimate effect we wish to estimate? These questions can be addressed through sensitivity analysis in Section 6.3.

5. Modelling Preclinical Durations

In Section 3, we discovered that a mixture of two exponential distributions provided a good fit to the clinical durations observed in three historic screening trials. Since a subject's preclinical and clinical durations are likely to be rather highly positively correlated, it makes sense that preclinical durations may also follow this same type of mixture distribution. Unfortunately, because the true length of a subject's preclinical duration is unknown, as it represents a period of cancer growth prior to diagnosis, we cannot plot the preclinical durations and estimate their distribution parameters as we did in Section 3. Thus, we need a different method of estimating the mixture distribution parameters for this unobservable data.

Our idea is that the preclinical duration distribution parameters may affect the pattern of diagnoses in a screening trial. When we say "the pattern of diagnoses," we mean the frequencies of diagnoses at general times in the screening trial: the first (initial) screen, the interval between the first and second screen, the second screen, the interval between the second and third screen, etc. Specifically, we hypothesize that the proportion of cases diagnosed at the first screen and the proportion of interval cases will be particularly informative about the distribution of preclinical durations. We also include the mean lead time of the trial along with the pattern of diagnoses, as we know this to also be related to the length of preclinical durations. Each of these three measures is discussed briefly below for its relationship to the preclinical duration distribution.

The proportion of cases diagnosed at the first screen records the fraction of treatment arm diagnoses that occurred at the initial screening test of the trial. This proportion is expected to be larger when the preclinical durations are longer, as cancers

dating back further from the start of the trial can be diagnosed at this first screen. For a simple illustration of this fact, consider two cancers: cancer A with all preclinical durations equal to one year and cancer B with all preclinical durations equal to two years. Assuming that the rate of new cancers is the same for both cancer A and cancer B, then it is expected that cancer B will have twice as many cases that are eligible for screen detection at the start of the trial. Even with imperfect test sensitivity, cancer B will be expected to have more diagnoses at the start of the trial than cancer A. Moreover, because the rate of new cancers is the same for both cancer A and B, the expected number of cases arising after the start of the trial should be the same for both, meaning that the proportion of diagnoses made at the start of the trial will be higher for cancer B. Thus, longer preclinical durations present more cases able to be detected by the first screening test of the trial, which will result in a larger proportion of the overall number of cases detected at this time.

The proportion of interval cases records the fraction of all cases diagnosed during the screening program that are interval diagnoses as opposed to screen detections. This proportion is expected to be larger when a cancer is faster-developing. Recall that an interval case arises when the entire preclinical duration falls between two screening tests or when a subject is diagnosed by symptoms after a previous false negative test. For the entire preclinical duration to fall between two screening tests, the preclinical duration can be no longer than the screening interval (time between screens). Moreover, the probability of a diagnosis by symptoms decreases as the number of screens increases, so, for a constant screening interval, the probability of an interval case is lower for longer preclinical durations. Both situations – the entire preclinical duration falling between two screens or diagnosis by symptoms after a previous false negative – favor faster-developing cancers

with shorter preclinical durations, meaning that the proportion of interval cases will increase for shorter preclinical durations.

The mean lead time measures the average amount of time earlier that cancer is diagnosed by screening compared to a diagnosis by clinical symptoms, and it can be estimated by the mean difference in time from entry to diagnosis between the treatment and control arms in the trial data. Presumably, longer lead times will result from longer preclinical durations. Consider two patients: patient A with a preclinical duration of one year and patient B with a preclinical duration of five years. Each will be screened annually starting at 0.5 years into the preclinical duration. If both patients are diagnosed with cancer at this first screen, patient A's lead time is 0.5 years, and patient B's lead time is 4.5 years. Although patient A is more likely to be diagnosed at this first screen due to a higher test sensitivity given that they are 50% of the way through their preclinical duration and patient B's is only 10% complete, patient B could be diagnosed at any of the three subsequent annual screens and still have a longer lead time than patient A diagnosed at the first screen. As such, longer preclinical durations present more opportunities for the screening test to diagnose cancer and will have longer lead times when such diagnoses occur.

Because the pattern of diagnoses and mean lead time in a screening trial are observable, we can use this information to infer about the preclinical duration distribution. While our previous discussion has focused on how three such measures – the proportion of cases diagnosed at the first screen, the proportion of interval cases, and the mean lead time – vary with the relative length of the preclinical durations, our hope is that changes to individual preclinical duration distribution parameters – p, λ_F, or λ_S – will affect each of the

measures differently, thus allowing us to reliably estimate all three mixture distribution parameters rather than just an overall mean. In Section 5.1, we will design an experiment to assess the effects of changes to preclinical duration distribution parameters (assuming a mixture of two exponential distributions) on the pattern of diagnoses and mean lead time in a simulated screening trial.

5.1 Simulation Study

We wanted to perform an experiment in which the three parameters in a mixture of two exponential distributions for preclinical durations were varied in order to evaluate their effects on the pattern of diagnoses and mean lead time in a screening trial. With three parameters to assess, we decided to design a 5^3 factorial experiment. The five levels for our three preclinical duration distribution parameters were as follows: $p = 0.1, 0.3, 0.5,$ 0.7, or 0.9; $\lambda_F = 1/(0.5 * 365), 1/(0.75 * 365), 1/(1 * 365), 1/(1.5 * 365),$ or $1/(2 * 365)$; and $\lambda_S = 1/(2 * 365), 1/(3 * 365), 1/(4 * 365), 1/(5 * 365),$ or $1/(7 * 365)$. For each combination of parameter values (125 possible scenarios), we would simulate 100 screening trials in which the preclinical durations were generated from the specified distribution.

Each simulated screening trial was carried out using the same approach as described in Section 4.1. In summary, preclinical durations were simulated from the mixture of two exponential distributions specified by the experimental variables, the time between the initiations of two preclinical durations was generated as a Poisson process at a rate of 210 new cancers per year, and any case whose preclinical duration overlapped with our screening window was included in both the treatment and control arm data

(presupposing no overdiagnosis). For this study, we used a screening program of four annual screens, so screening occurred at the start of the simulated trial, as well as at the end of the first, second, and third years. This screening program was selected to mirror the HIP trial [15] and the lung cancer screening arm of the PLCO [16]. A more detailed description of simulating the screening process can be reviewed in Section 4.1 or Figure 2 of Kafadar and Prorok [46]. Eight response variables were recorded for each simulated trial: the mean lead time, the proportion of cases diagnosed at the each of the four screens, and the proportion of interval cases in each of the three between-screen intervals.

We discovered in Section 4 that the distribution of preclinical durations and test sensitivity have an interactive effect on the mean lead time in a screening trial. Moreover, one of these models cannot be predicted without knowing the other, as the distribution of preclinical durations and test sensitivity are confounded in the screening trial's pattern of diagnoses. Therefore, we will have to assume a model for test sensitivity for our experiment without knowing the true sensitivities as a function of preclinical growth. We decided on a model where the initial sensitivity at the start of the preclinical duration was 20% and the sensitivity increased following the functional form of the Normal cdf to a maximum test sensitivity of 90% achieved three-quarters of the way through the preclinical duration. Recognizing that our sensitivity model is only an educated guess, we will have to evaluate the sensitivity of our preclinical duration parameter estimates to a misspecification of the test sensitivity function later.

After carrying out our experiment, we found that the proportion of interval cases was roughly the same in each of the three between-screen intervals for a given preclinical duration distribution scenario. The correlations between the proportions of interval cases

were 0.94 when comparing the first and second intervals, 0.94 when comparing the first and third intervals, and 0.95 when comparing the second and third intervals. Thus, we decided to combine these three responses into one variable: the total proportion of interval cases. This reduced our total number of response variables from eight to six.

To determine the significance of associations between the preclinical duration distribution parameters and our six response variables, we performed ANOVAs. In each case, the ANOVA model included the three preclinical duration distribution parameters, as well as all two- and three-way interactions. We treated each of the three preclinical duration distribution parameters as quantitative variables, rather than categorical, and estimated a linear effect for each; we had found little gain in sum of squares relative to the loss in degrees of freedom when treating these parameter levels as categorical variables. Figure 5.1 displays the ANOVA tables for this analysis. Overall, we found that all three preclinical duration distribution parameters and all two-way interactions between these parameters had statistically significant effects on each of our six response variables. This proves our original idea correct, that the mean lead time and pattern of diagnoses in a screening trial are affected by the distribution of preclinical durations.

Reviewing the ANOVA results one-by-one, we see some interesting associations between our preclinical duration distribution parameters and the six summary measures of the pattern of diagnoses in a trial. First, the R^2 values are particularly high for the response variables of mean lead time ($R^2 = 98\%$), the proportion of interval cases ($R^2 = 95\%$), and the proportion of cases diagnosed at the first screen ($R^2 = 95\%$). This suggests that most of the trial-to-trial variability in these measures can be explained by the distribution of preclinical durations, so using these measures from a real trial should allow

Figure 5.1: Analysis of variance results for the effect of preclinical duration distribution parameters on mean lead time and the pattern of diagnoses.

(a) Mean lead time

Factor	df	SS	MS	F	p-value
p	1	5,609	5,609	254,142.9	<0.0001
λ_F	1	212	212	9,609.5	<0.0001
λ_S	1	8,820	8,820	399,634.7	<0.0001
$p * \lambda_F$	1	132	132	5,980.2	<0.0001
$p * \lambda_S$	1	1,942	1,942	87,987.0	<0.0001
$\lambda_F * \lambda_S$	1	12	12	538.2	<0.0001
$p * \lambda_F * \lambda_S$	1	0	0	11.4	0.0007
Error	12,492	276	0		

(b) Proportion of interval cases

Factor	df	SS	MS	F	p-value
p	1	81.7	81.7	113,400.0	<0.0001
λ_F	1	39.9	39.9	55,330.0	<0.0001
λ_S	1	18.8	18.8	26,040.0	<0.0001
$p * \lambda_F$	1	18.4	18.4	25,510.0	<0.0001
$p * \lambda_S$	1	3.1	3.1	4,388.0	<0.0001
$\lambda_F * \lambda_S$	1	0.2	0.2	302.9	<0.0001
$p * \lambda_F * \lambda_S$	1	0.0	0.0	0.5	0.50
Error	12,492	9.0	0.0		

(c) Proportion of cases diagnosed at the first screen

Factor	df	SS	MS	F	p-value
p	1	32.9	32.9	128,100.0	<0.0001
λ_F	1	7.8	7.8	30,490.0	<0.0001
λ_S	1	14.9	14.9	58,030.0	<0.0001
$p * \lambda_F$	1	3.6	3.6	13,990.0	<0.0001
$p * \lambda_S$	1	3.9	3.9	15,090.0	<0.0001
$\lambda_F * \lambda_S$	1	0.1	0.1	185.0	<0.0001
$p * \lambda_F * \lambda_S$	1	0.0	0.0	0.4	0.51
Error	12,492	3.2	0.0		

(d) Proportion of cases diagnosed at the second screen

Factor	df	SS	MS	F	p-value
p	1	1.52	1.52	10,787.8	<0.0001
λ_F	1	0.95	0.95	6,725.8	<0.0001
λ_S	1	0.27	0.27	1,931.6	<0.0001
$p * \lambda_F$	1	0.50	0.50	3,520.8	<0.0001
$p * \lambda_S$	1	0.03	0.03	215.2	<0.0001
$\lambda_F * \lambda_S$	1	0.01	0.01	46.3	<0.0001
$p * \lambda_F * \lambda_S$	1	0.00	0.00	1.2	0.28
Error	12,492	1.76	0.00		

(e) Proportion of cases diagnosed at the third screen

Factor	df	SS	MS	F	p-value
p	1	0.09	0.09	755.7	<0.0001
λ_F	1	0.55	0.55	4,399.7	<0.0001
λ_S	1	0.14	0.14	1,120.6	<0.0001
$p * \lambda_F$	1	0.26	0.26	2,067.1	<0.0001
$p * \lambda_S$	1	0.08	0.08	666.1	<0.0001
$\lambda_F * \lambda_S$	1	0.00	0.00	28.7	<0.0001
$p * \lambda_F * \lambda_S$	1	0.00	0.00	0.5	0.49
Error	12,492	1.57	0.00		

(f) Proportion of cases diagnosed at the fourth screen

Factor	df	SS	MS	F	p-value
p	1	0.01	0.01	59.7	<0.0001
λ_F	1	0.54	0.54	4,344.5	<0.0001
λ_S	1	0.40	0.40	3,246.7	<0.0001
$p * \lambda_F$	1	0.24	0.24	1,931.6	<0.0001
$p * \lambda_S$	1	0.18	0.18	1,460.9	<0.0001
$\lambda_F * \lambda_S$	1	0.00	0.00	26.9	<0.0001
$p * \lambda_F * \lambda_S$	1	0.00	0.00	1.4	0.24
Error	12,492	1.54	0.00		

Abbreviations: df, degrees of freedom; SS, sum of squares; MS, mean squares; F, F-statistic; p, proportion of fast-developing cases; λ_F, exponential rate parameter for fast-developing cases; λ_S, exponential rate parameter for slow-developing cases.

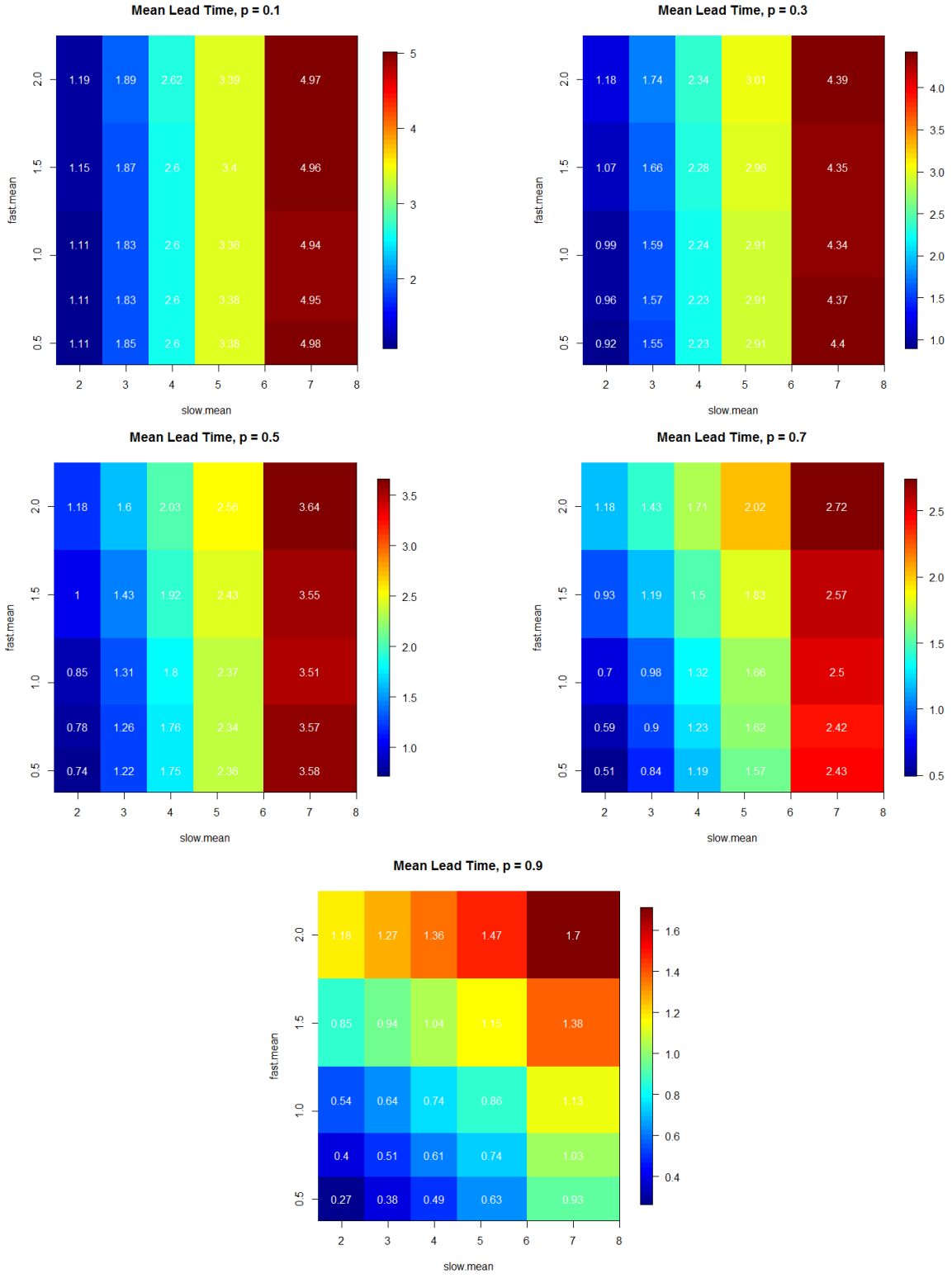
us to pinpoint the preclinical duration distribution parameters fairly accurately. Second, it seems that different preclinical parameters had the largest effect on different response variables. The mean of the slow-developing cases and their proportion of the mixture distribution were the greatest sources of variability in mean lead time and the proportion of cases diagnosed at the first screen (Figures 5.1a and 5.1c, respectively), while the mean of the fast-developing cases and their proportion in the mixture distribution were the greatest sources of variability in the proportion of interval cases (Figure 5.1b). Finally, regarding the proportions of cases diagnosed at the third and fourth screens (Figures 5.1e and 5.1f, respectively), the residual sums of squares are greater than the sums of squares accounted for by the preclinical duration parameters and their interactions. This suggests that most of the variability from trial-to-trial in these two measures is "random" and not explained by the distribution of preclinical durations, so the proportions of cases diagnosed at the third and fourth screens may not be as useful to predict the distribution parameters.

Figure 5.2 presents heatmaps of mean lead time, the proportion of interval cases, the proportion of cases diagnosed at the first screen, and the proportion of cases diagnosed at the second screen across the 125 experimental scenarios. Each individual heatmap pertains to a certain proportion of fast-developing cases p, showing changes in a response measure with the mean slow-developing preclinical duration in years $(1/(\lambda_S * 365))$ on the x-axis and the mean fast-developing preclinical duration in years $(1/(\lambda_F * 365))$ on the y-axis. Moving from left-to-right and top-to-bottom across the heatmaps shows the effect of an increase in the proportion of fast-developing cases p.

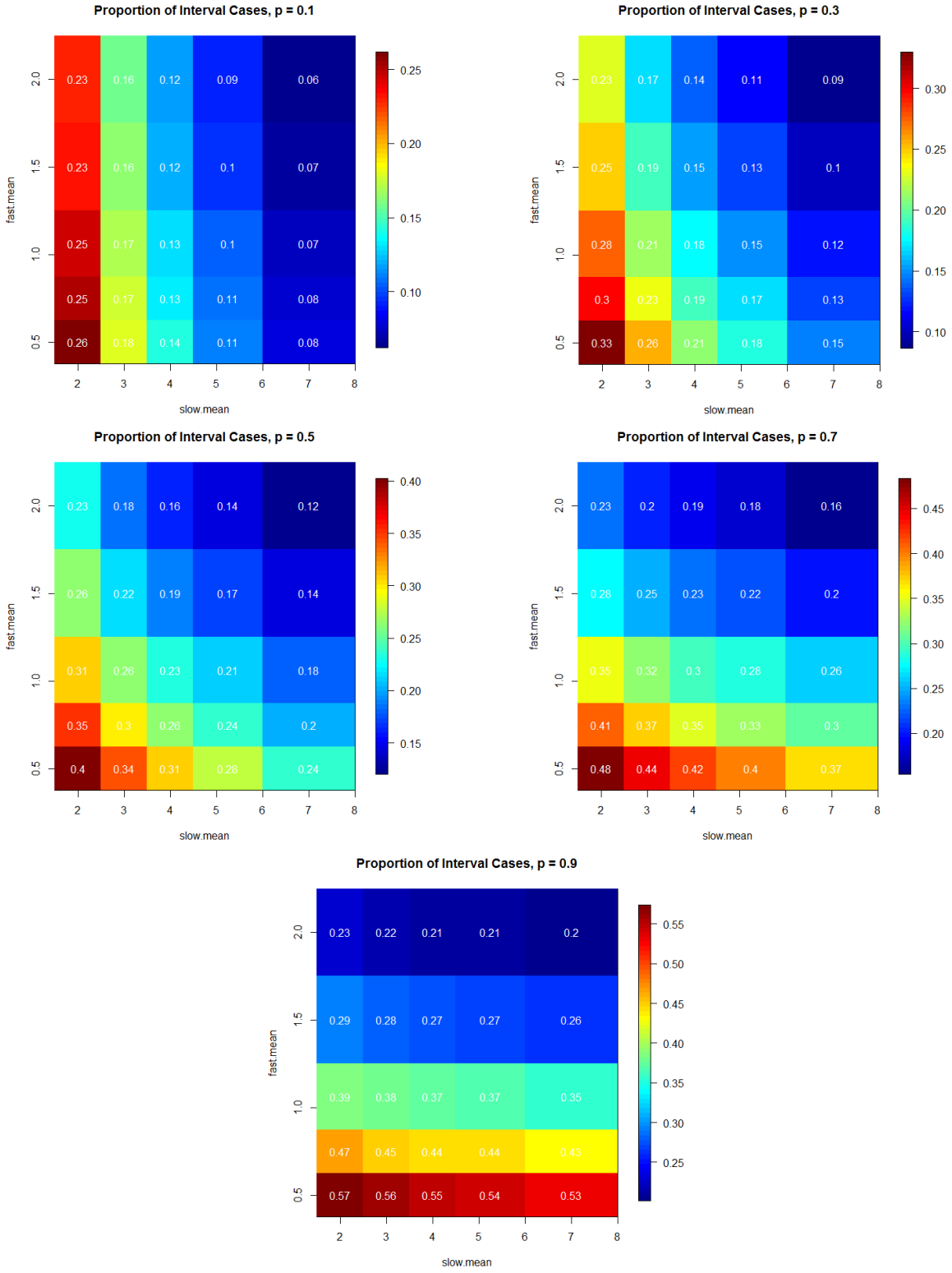
The heatmaps show several noteworthy trends. Overall, as the preclinical durations increased, either for the fast-developing group, the slow-developing group, or due to a

Figure 5.2: Heatmaps of results for different preclinical duration distribution parameters.

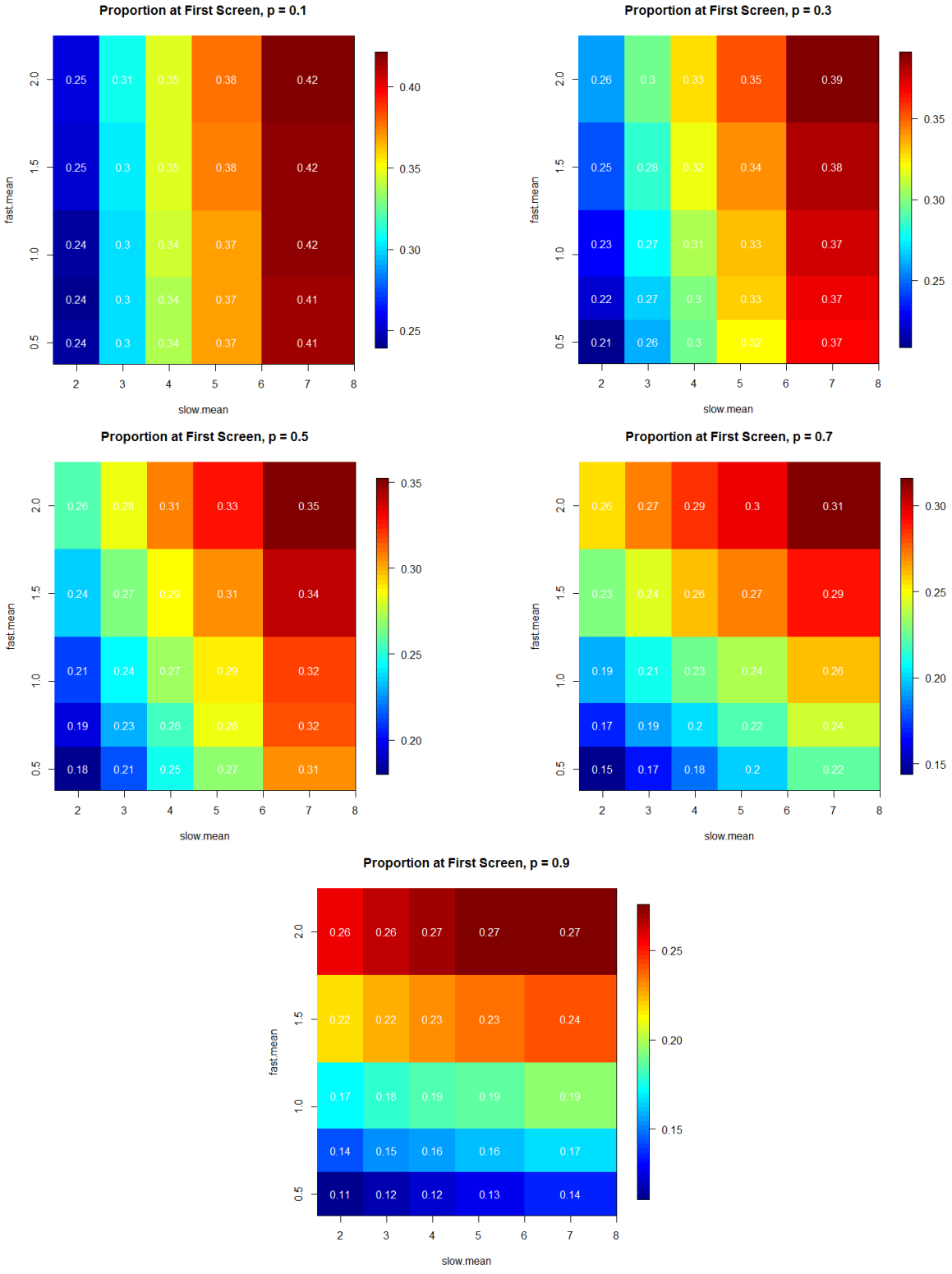
(a) Mean lead time



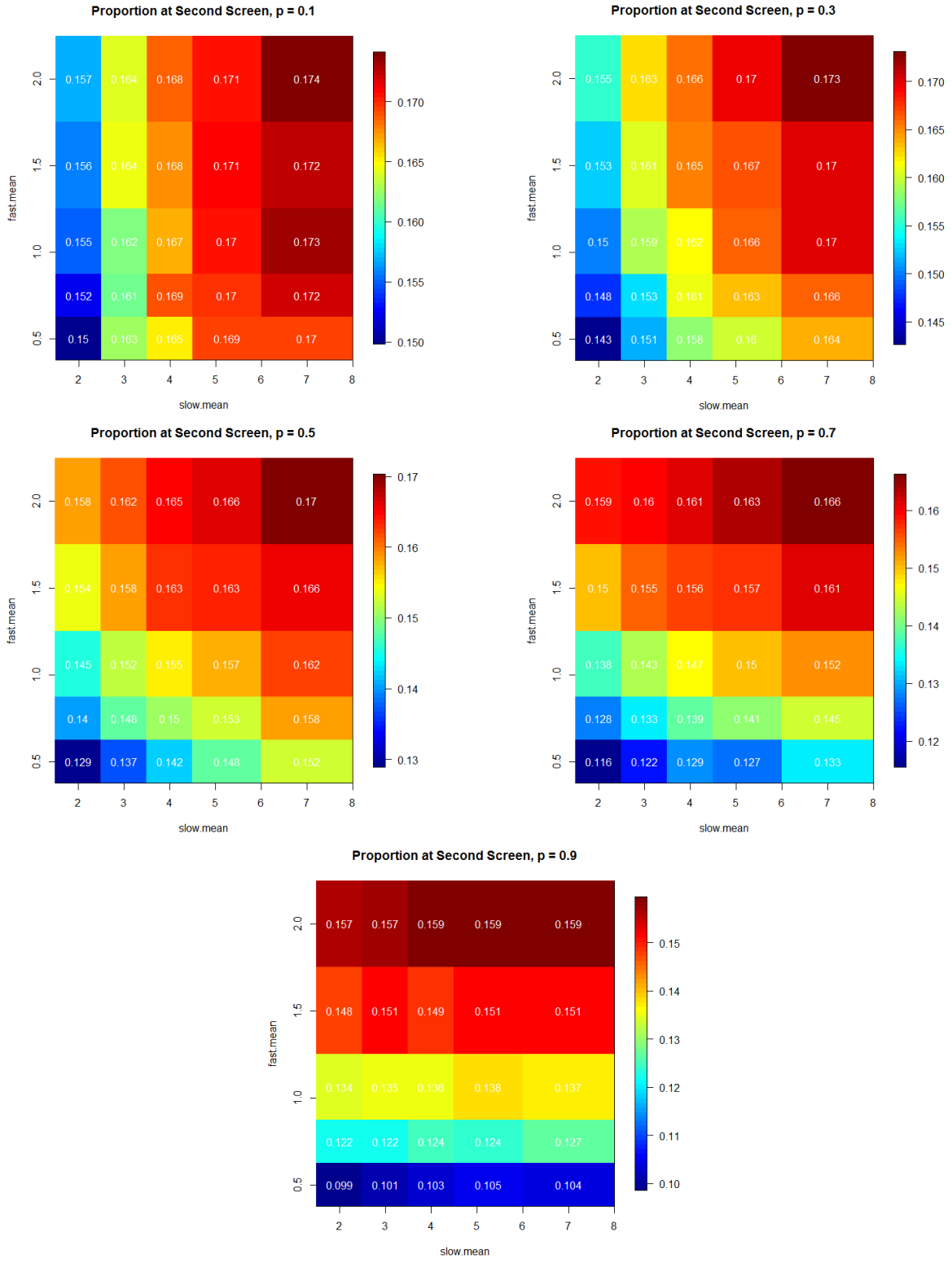
(b) Proportion of interval cases



(c) Proportion of cases diagnosed at the first screen



(d) Proportion of cases diagnosed at the second screen



change in the proportion of fast-developing cases, we saw that mean lead time increased, the proportion of interval cases decreased, and the proportion of cases diagnosed at both the first and second screens increased. We also observed interesting trends for the extreme preclinical duration distributions, where the proportion of fast-developing cases is either 10% or 90%. When only 10% of the cases are fast-developing, we see that the response measures are largely unaffected by a change in the mean preclinical duration for the fast-developing group; only the slow-developing group mean matters. For the mean lead time, it seems that the mean of the fast-developing group does not have much of an effect until over 50% of the cases are fast-developing. Furthermore, when 90% of the cases are fast-developing, we see that mean lead time is the only measure still affected whatsoever by the mean of the slow-developing cases; the proportion of interval cases, the proportion of cases diagnosed at the first screen, and the proportion of cases diagnosed at the second screen are only affected by the mean of the fast-developing cases. This suggests that we may struggle to identify the mean of the smaller component of the mixture distribution when the larger component is so dominant. Finally, when comparing trials in which the overall mean preclinical duration is roughly the same, we are able to differentiate between the distribution parameters; three examples are shown in Figure 5.3. In all three cases, the mean preclinical duration is approximately three years, but the mean lead time and pattern of diagnoses vary greatly across the three trial scenarios. This is further evidence suggesting that we should be able to separately identify all three preclinical duration distribution parameters from the mean lead time and pattern of diagnoses in a trial.

Figure 5.3: Comparing results in scenarios with a similar overall preclinical duration distribution mean.

p	$\frac{1}{\lambda_F * 365}$	$\frac{1}{\lambda_S * 365}$	μ	Mean lead time	Proportion of interval cases	Proportion diagnosed at screen 1	Proportion diagnosed at screen 2
0.5	2	4	3.0	2.03	0.16	0.31	0.165
0.5	1	5	3.0	2.37	0.21	0.29	0.157
0.7	0.75	5	3.025	1.62	0.33	0.22	0.141

5.2 Building a Predictive Model

The results from Section 5.1 suggested that the parameters of the distribution of preclinical durations could be inferred from the mean lead time and pattern of diagnoses in a trial. This is an important finding that can lead to valuable insights about the underlying disease process in the at-risk population covered by the screening trial. We further discuss the importance of these implications in Section 5.4. We wanted to use our 12,500 simulated screening trials to create a predictive model so that the unobservable preclinical duration parameters could be estimated from new trial data based on the trial's observed characteristics. The challenge of developing such a model is that we are trying to jointly estimate three variables – p, $1/(\lambda_F * 365)$, and $1/(\lambda_S * 365)$ – rather than a single response. To do so, we will use multivariate least-squares linear regression.

For multiple linear regression with a univariate response, we specify the relationship between explanatory and response variables in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}. \tag{5.1}$$

$\boldsymbol{y}$ is a $n * 1$ vector of responses. $\boldsymbol{X} = [\boldsymbol{1_n}, \boldsymbol{x_1}, \dots, \boldsymbol{x_p}]$ is a $n * (p + 1)$ matrix in which each of the final p columns corresponds to one of the explanatory variables used to predict $\boldsymbol{y}$ (the first column of 1's in $\boldsymbol{X}$ is included to estimate an intercept for $\boldsymbol{y}$); the rows correspond to the individual observations. $\boldsymbol{b} = [b_0, b_1, \dots, b_p]^T$ is a $(p + 1) * 1$ vector of coefficients specifying the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$, and $\boldsymbol{e}$ is a $n * 1$ vector of the prediction errors between $\boldsymbol{Xb}$ and $\boldsymbol{y}$. Using the least-squares method, $\boldsymbol{b}$ is estimated so that the sum of squared prediction errors, or $\boldsymbol{e}^T\boldsymbol{e}$, is minimized, i.e.

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{5.2}$$

Multivariate least-squares linear regression [47] expands this method from a one-dimensional response $\boldsymbol{y}$ to a k-dimensional response $\boldsymbol{Y} = [\boldsymbol{y_1}, \dots, \boldsymbol{y_k}]$. Because the same explanatory variables are used to predict this multivariate response as were used in the univariate case, $\boldsymbol{X}$ remains the same. However, $\boldsymbol{b}$ now expands to a $(p + 1) * k$ matrix $\boldsymbol{B} = [\boldsymbol{b_1}, \dots, \boldsymbol{b_k}]$, where each column $\boldsymbol{b_j}$ contains the $(p + 1)$ coefficients for a different response $\boldsymbol{y_j}$. The matrix of residuals $\boldsymbol{e}$ also expands to a $n * k$ matrix $\boldsymbol{E} = [\boldsymbol{e_1}, \dots, \boldsymbol{e_k}]$, as prediction errors are now being made for k responses per observation. The least-squares solution still remains the same with

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{5.3}$$

The result of multivariate least-squares regression is k independent regression equations, one for each response $\boldsymbol{y_j}$:

$$\hat{y}_{ij} = \hat{b}_{0j} + \sum_{m=1}^{p} \hat{b}_{mj} x_{im}. \quad (5.4)$$

We can apply the multivariate least-squares linear regression method to our $k = 3$ dimensional response $\boldsymbol{Y} = [\boldsymbol{p}, \mathbf{1}/(\boldsymbol{\lambda_F} * \mathbf{365}), \mathbf{1}/(\boldsymbol{\lambda_S} * \mathbf{365})]$. $\boldsymbol{X}$ will be a $12{,}500 * 7$ matrix, as we have 12,500 simulated trials with six measures recorded from each: the mean lead time, the proportion of interval cases, and the proportion of cases diagnosed at each of four screens. The resulting coefficient estimates are displayed in Figure 5.4. We found that adding interactions between our six measures did not improve the model fit, so these terms were excluded. The R^2 values of the individual regression models were 51% for the proportion of fast-developing cases, 41% for the mean of the fast-developing cases, and 66% for the mean of the slow-developing cases. Given some of the high R^2 values we calculated from the ANOVA tables in Figure 5.1, which suggested that the preclinical duration distribution parameters and their interactions explained over 90% of the variability in three of our measures (mean lead time, proportion of interval cases, and proportion of cases diagnosed at the first screen), the R^2 values for the regression models to predict the preclinical parameters using those very measures seem awfully low. Perhaps one explanation is that the three regression equations were estimated independently, meaning that they did not take into account the correlation between the response variables. For example, the proportion of fast-developing cases and the mean of the fast-developing cases will be estimated separately based on the mean lead time and pattern of diagnoses in a trial even though these two parameters are most definitely dependent on one another.

Figure 5.4: Coefficients for the multivariate least-squares regression model to predict preclinical duration distribution parameters.

Response	$\boldsymbol{p}$	$\boldsymbol{1/(\lambda_F * 365)}$	$\boldsymbol{1/(\lambda_S * 365)}$
Intercept	-0.49	3.81	10.02
Mean lead time	0.07	-0.23	2.07
Proportion of interval cases	1.75	-5.66	-3.84
Proportion diagnosed at first screen	-1.44	-2.49	-21.94
Proportion diagnosed at second screen	1.30	-1.68	-9.98
Proportion diagnosed at third screen	2.09	0.13	-6.66
Proportion diagnosed at fourth screen	2.56	0.68	-4.79

5.3 Evaluating the Predictive Model: Simulated Examples

To evaluate the accuracy of the multivariate regression model we developed in Section 5.2, we wanted to run simulations to assess how it performs for various preclinical duration distributions. We will simulate eight screening trials under a variety of preclinical duration distributions, following the same simulation procedure as described in Section 5.1. Our goal is to determine the prediction accuracy of our model for different combinations of small, moderate, and large values for p, $1/(\lambda_F * 365)$, and $1/(\lambda_S * 365)$. The eight scenarios are generalized as follows:

A. slow disease progression

B. most cases are slow-developing with a few extremely fast

C. half of cases are very slow-developing, rest are moderately fast

D. half of cases are very slow-developing, half are very fast-developing

E. moderate paced disease progression

F. half of cases are very fast-developing, rest are moderately slow

G. most cases are fast-developing with a few extremely slow

H. fast disease progression

Figure 5.5: Eight simulated scenarios to test our multivariate regression model.

Scenario	p	$\frac{1}{\lambda_F * 365}$	$\frac{1}{\lambda_S * 365}$	Mean Lead Time	Prop. Interval Cases	Prop. Screen 1	Prop. Screen 2	Prop. Screen 3	Prop. Screen 4
A	**0.3**	**1.5**	**5**	3.00	0.12	0.35	0.175	0.148	0.109
B	**0.1**	**0.5**	**5**	3.31	0.12	0.37	0.177	0.122	0.119
C	**0.5**	**1**	**7**	3.54	0.18	0.32	0.164	0.129	0.114
D	**0.5**	**0.5**	**7**	3.75	0.23	0.31	0.158	0.122	0.104
E	**0.5**	**1**	**4**	1.73	0.22	0.25	0.184	0.140	0.127
F	**0.5**	**0.5**	**4**	1.81	0.28	0.25	0.144	0.127	0.115
G	**0.9**	**0.75**	**7**	0.97	0.43	0.15	0.148	0.109	0.118
H	**0.7**	**0.75**	**3**	0.97	0.35	0.22	0.112	0.141	0.126

Figure 5.5 displays the parameters of the eight screening trial scenarios, as well as the mean lead time and pattern of diagnoses for each simulated trial. Before proceeding to predict the preclinical duration distribution parameters, we checked whether any of these individual trials deviated significantly from what is expected given the preclinical duration distribution parameters; in comparing these eight results to the mean results from 100 trials in our heatmaps in Figure 5.2, we found that there were no major differences. Thus, we would expect our model predictions to be fairly accurate, as none of our simulated trials produced an extreme result that deviated greatly from the sample data.

Figure 5.6 compares the true parameter values to the parameter values predicted by our multivariate regression model. Overall, the prediction errors made by our model were modest in size. However, when the mean of the fast-developing cases was set to 0.5 years, the smallest value of this parameter in our simulated data, the model would consistently overestimate this value. This same trend of considerable misestimation was generally not observed for extreme values of the mean of the slow-developing cases, and the model's predictions were much more accurate for the mean of the fast-developing cases when the

Figure 5.6: The prediction results for our eight simulated scenarios.

Scenario	p	$\frac{1}{\lambda_F * 365}$	$\frac{1}{\lambda_S * 365}$	$\hat{p}$	$\frac{1}{\hat{\lambda}_F * 365}$	$\frac{1}{\hat{\lambda}_S * 365}$
A	**0.3**	**1.5**	**5**	0.25	1.37	4.85
B	**0.1**	**0.5**	**5**	0.23	1.24	5.21
C	**0.5**	**1**	**7**	0.40	0.98	6.51
D	**0.5**	**0.5**	**7**	0.46	0.72	7.21
E	**0.5**	**1**	**4**	0.52	1.35	3.89
F	**0.5**	**0.5**	**4**	0.54	1.03	4.45
G	**0.9**	**0.75**	**7**	0.85	0.63	4.36
H	**0.7**	**0.75**	**3**	0.64	0.96	3.12

true mean was larger. Additionally, in the extremely imbalanced mixture distributions from Scenarios B and G, the model struggled to accurately predict the mean of the smaller component of the mixture distribution. In Scenario B, where just 10% of cases were fast-developing, the mean of this smaller component was grossly overestimated, as was the proportion of fast-developing cases. Similarly, in scenario G, where just 10% of cases were slow-developing, the mean of these cases was underestimated considerably. This follows a trend we saw in the heatmaps in which the mean of the smaller component of the mixture distribution had little-to-no effect on the trial's mean lead time and pattern of diagnoses, which would make it more difficult to estimate.

5.4 Discussion

Section 3 found that the observed clinical durations in three historic screening trials could be modelled using a mixture of two exponential distributions, which suggests that unobserved but correlated preclinical durations could feasibly be modelled by this same mixture distribution, as well. In Section 5, we set out to determine whether the parameters of this mixture distribution for preclinical durations could be estimated based on the

pattern of diagnoses in a trial. Section 5.1 carried out a factorial experiment to evaluate how changes to the three preclinical duration distribution parameters would affect the mean lead time, proportion of interval cases, and proportion of diagnoses at each screen in a trial. We found that there was a highly significant interactive effect between the parameters on all of our trial outcome measures. Section 5.2 used multivariate least-squares linear regression to develop a model to predict the preclinical duration parameters based on the simulated trial results from Section 5.1, and Section 5.3 evaluated the accuracy of the predictive model using new simulated trials under a variety of preclinical duration distributions. We found that our model was fairly accurate except in some extreme distribution scenarios.

To our knowledge, ours is the first study to predict the distribution of preclinical durations based on the observable pattern of diagnoses in trial data. Whether or not a screening test is beneficial, our study shows that the results of a screening trial can inform us about the nature of the underlying disease in the at-risk population covered by the trial. This finding is also important because it will allow real trials to be more accurately modelled in the simulated setting. Some attributes, such as the length-biased sampling effect on the screen-detections in a trial, which is highly dependent on the distribution of preclinical durations, can only be estimated through simulations. If the mean lead time and pattern of diagnoses from a real screening trial can be used to predict the distribution of preclinical durations, this distribution can then be used to estimate the length-biased sampling effect on the trial. The relationship between the distribution of preclinical durations and length-biased sampling effect will be explored further in Section 6.

Despite the novelty of our research, we did expect the predictive model developed in Section 5.2 to be more accurate. While the prediction errors in Section 5.3 for extreme scenarios – such as a highly imbalanced mixture distribution of preclinical durations – were reasonable given the trends observed in the simulation study heatmaps, the prediction errors for more moderate distributions of preclinical durations are not as easily understood. Because we found such minimal differences in the mean lead time and pattern of diagnoses between the individual simulated trials in Section 5.3 and the averages from 100 such trials in Section 5.1, it seems that a user-driven "trial-and-error" approach to match observed trial outcomes to a preclinical duration distribution could be more accurate than our predictive model. Perhaps the best explanation for the prediction errors is our choice of methodology to develop the predictive model. Multivariate least-squares linear regression estimates each parameter independently, but, as evidenced by the significant interactive effects between the parameters on trial outcomes in Figure 5.1, it is clear that dependencies exist between the three parameter specifications for a given mean lead time and pattern of diagnoses. To improve our predictive model, a method for correlated responses should be considered, which will be discussed further in Section 7.2.

Another potential cause of prediction error in estimating preclinical duration distribution parameters went unaddressed in our simulated examples in Section 5.3: the screening test sensitivity. The simulation study we designed in Section 5.1 assumed a certain model for sensitivity, and each of our simulated examples was carried out using the same function; thus, we did not evaluate how the prediction errors would change when the model for test sensitivity was misspecified. We discovered in Section 4 that, when test sensitivity is a function of preclinical growth, most attributes of the sensitivity model have

little-to-no effect on the mean lead time of a screening trial, which suggests that misspecification of one of these attributes would not have a detrimental effect on the prediction for the preclinical duration distribution. However, the initial test sensitivity at the start of the preclinical duration did significantly affect the mean lead time of the trial. While we did not explore the effect of a misspecified starting sensitivity on the predicted preclinical duration distribution parameters in Section 5, we will assess its impact on the length-biased sampling effect in Section 6.

Our prediction process was based on simulated trial data generated using the same screening program as the HIP trial and the lung cancer screening arm of the PLCO. Unfortunately, predictions for the distributions of preclinical durations for these trials are complicated by two additional factors: refusals and overdiagnosis. Both are complex to model in a simulation, as described below.

In treatment arm of the HIP trial, 120 of the women diagnosed with cancer, 29% of all treatment arm cases, refused screening. This is an exceedingly high rate, much higher than the 12% and 11% of refusals among lung and ovarian cancer treatment arm diagnoses in the PLCO, respectively. Because the predictive model built from our simulated trials assumed no refusals, we would need to assume how these subjects would have been diagnosed had they not refused screening. For example, if we assume that it was completely random for a subject to refuse screening, we could delete these 120 observations from the data. If we assume that the refusals would not have been screen-detected, we could categorize them as interval or post-screening cases based on the length of time between their enrollment in the trial and diagnosis. However, when the proportion of refusals is as high as it is in the HIP trial, any estimate of the mean lead time, proportion

of interval cases, or proportion of cases diagnosed at each screen would be highly dependent on how we classified the refusals. We decided that it was best to not make such conditional predictions from the HIP trial data. Even if we wanted to adjust the design of our simulation study to include a certain percentage of refusals, we would still need to make some assumption regarding the mechanism for why subjects are refusing treatment. Because such a high rate of refusals seems to be a unique issue to the HIP trial that does not persist in more recent trials like the PLCO, we feel that this does not necessarily warrant too much further investigation.

The lung cancer screening results in the PLCO had overdiagnoses from the chest X-ray diagnostic test. Recall that an overdiagnosis occurs when a subject is diagnosed by screening with a cancer that never would have progressed to an advanced enough stage to be diagnosed in the absence of screening. As a result, there is no matching subject in the control arm of the trial for the overdiagnosed screen-detected subject. At the end of the thirteen-year follow-up period in the PLCO, 1,801 subjects had been diagnosed with lung cancer in the treatment arm, while only 1,719 had been diagnosed in the control arm. This difference of eighty-two in cumulative cancer incidence can be attributed to overdiagnosis. Not only does overdiagnosis lead to a surplus of diagnoses in the treatment arm, but it also makes the trial's catch-up point unidentifiable. When the treatment and control arms are compared at the end of the follow-up period rather than at the catch-up point, there is an excess of post-screening detections in the treatment arm that dilute the benefit of screening. Thus, when a trial includes overdiagnoses, the mean lead time, mean benefit time, or even the pattern of diagnoses through catch-up cannot be accurately estimated, as these measures are affected by both the overdiagnoses and the high number of post-

screening detections in the treatment arm. This explains why we were unable to predict the parameters of the distribution of preclinical durations in the lung cancer screening arm of the PLCO.

The design of our simulation study in Section 5.1 could be modified to include a certain percentage of refusals or overdiagnoses. With respect to refusals, a useful direction might be to identify characteristics among the available covariates that could predict a participant's predilection to refuse treatment. For example, past studies may suggest that the rate varies with age, and a future simulation could properly incorporate a certain percentage of refusers per age group. Similarly, other covariates may inform the likelihood of a case being overdiagnosed. Preclinical duration is an obvious example, as slower-developing cases are more likely to be overdiagnosed than faster-developing cases by definition; generating overdiagnoses in simulations based on preclinical duration length is discussed in more detail in Section 7.2. Other factors, perhaps including a participant's weight or comorbidities, could also be useful predictors for creating a more realistic simulation of overdiagnoses. These ideas could be explored in future studies.

6. Estimating the Length-Biased Sampling Effect

In Section 5, we found that, assuming the preclinical durations of a certain cancer follow a mixture of two exponential distributions, we can predict the parameters of this mixture distribution based on the pattern of diagnosis times in screening trial data. Given the results from Section 3, that the clinical durations from three historic screening trials could all be modelled using this same mixture of two exponential distributions, we believe that adopting this mixture distribution for the preclinical durations is a reasonable assumption because of the plausibly strong correlation between a subject's preclinical and clinical durations. Thus, we can estimate two key components in the model for cancer growth – the distributions of preclinical and clinical durations – based on the pattern of diagnoses and survival times in real screening trial data, with just the correlation between these two components left to be estimated (a challenge which will be discussed further in Section 7.2).

These developments are especially impactful because both Kafadar and Prorok [29] and Heltshe et al [12] found that the joint distribution of preclinical and clinical durations had a great impact on the length-biased sampling effect in a randomized controlled screening trial. However, neither study modelled these durations using a mixture of two exponential distributions. The distributions used – the bivariate gamma and bivariate lognormal, respectively – seemed to be selected for simplicity or model flexibility, as neither study verified that their assumed distribution was actually a good fit for cancer growth periods. We believe that the simulation studies carried out by Kafadar and Prorok and Heltshe et al should be reassessed using our mixture distribution to generate cancer

growth periods, as our model has been proven to fit the distribution of these durations well.

Recall that there are two quantities of interest when assessing the length-biased sampling effect on a screening trial. $E[Y^*]/E[Y]$ measures how much longer the average preclinical duration is for a screen-detected case relative to the overall mean preclinical duration, and $E[Z^*]/E[Z]$ measures how much longer the average clinical duration is for a screen-detected case relative to the overall mean clinical duration. Unfortunately, without specifying a correlation between the preclinical and clinical durations, $E[Z^*]/E[Z]$ cannot be estimated, as it is unknown which clinical durations would match with the screen-detected preclinical durations. Because we have formulated a model for the preclinical and clinical durations based on real trial data, we do not want to hazard a speculative guess at the correlation structure between these two cancer growth periods. As a result, the simulation studies we design in Sections 6.1 and 6.2 will focus solely on $E[Y^*]/E[Y]$ and identifying the factors that augment this length-biased sampling effect.

6.1 Simulation Study: Preclinical Duration Distribution and Sensitivity on $E[Y^*]/E[Y]$

In our first simulation study, we assess the significance of preclinical duration distribution parameters and test sensitivity model on the length-biased sampling effect $E[Y^*]/E[Y]$. With regards to the preclinical duration distribution, we will assume that these times follow a mixture of two exponential distributions, so there will be three parameters varied in our study: p, the proportion of fast-developing cases; λ_F, the rate parameter for the fast-developing cases; and λ_S, the rate parameter for the slow-developing cases. For the test sensitivity model, we discussed in Section 4 that a realistic test sensitivity model would

allow sensitivity to increase as the preclinical duration progresses, and we discovered that the only attribute of such a model that meaningfully affected the pattern of diagnoses in a trial is the initial test sensitivity at the start of the preclinical duration. Hence, in our simulation study, the starting sensitivity will be the only attribute of the sensitivity model that we evaluate; across all scenarios, the maximum test sensitivity will be 90%, the maximum test sensitivity will be first achieved three-quarters of the way through the preclinical duration, and test sensitivity will increase with preclinical growth following a Normal cdf function.

We once again decided to use a factorial design for our simulation study, with four varied factors. For the preclinical duration distribution parameters, we defined five levels each: $p = 0.1, 0.3, 0.5, 0.7,$ or 0.9; $\lambda_F = 1/(0.5 * 365)$, $1/(0.75 * 365)$, $1/(1 * 365)$, $1/(1.5 * 365)$, or $1/(2 * 365)$; and $\lambda_S = 1/(2 * 365)$, $1/(3 * 365)$, $1/(4 * 365)$, $1/(5 * 365)$, or $1/(7 * 365)$. For the starting test sensitivity, we defined two levels: $\beta_0 = 0.1$ or 0.3. In total, this created 250 possible combinations of levels, and each scenario was simulated 100 times for replication of the results.

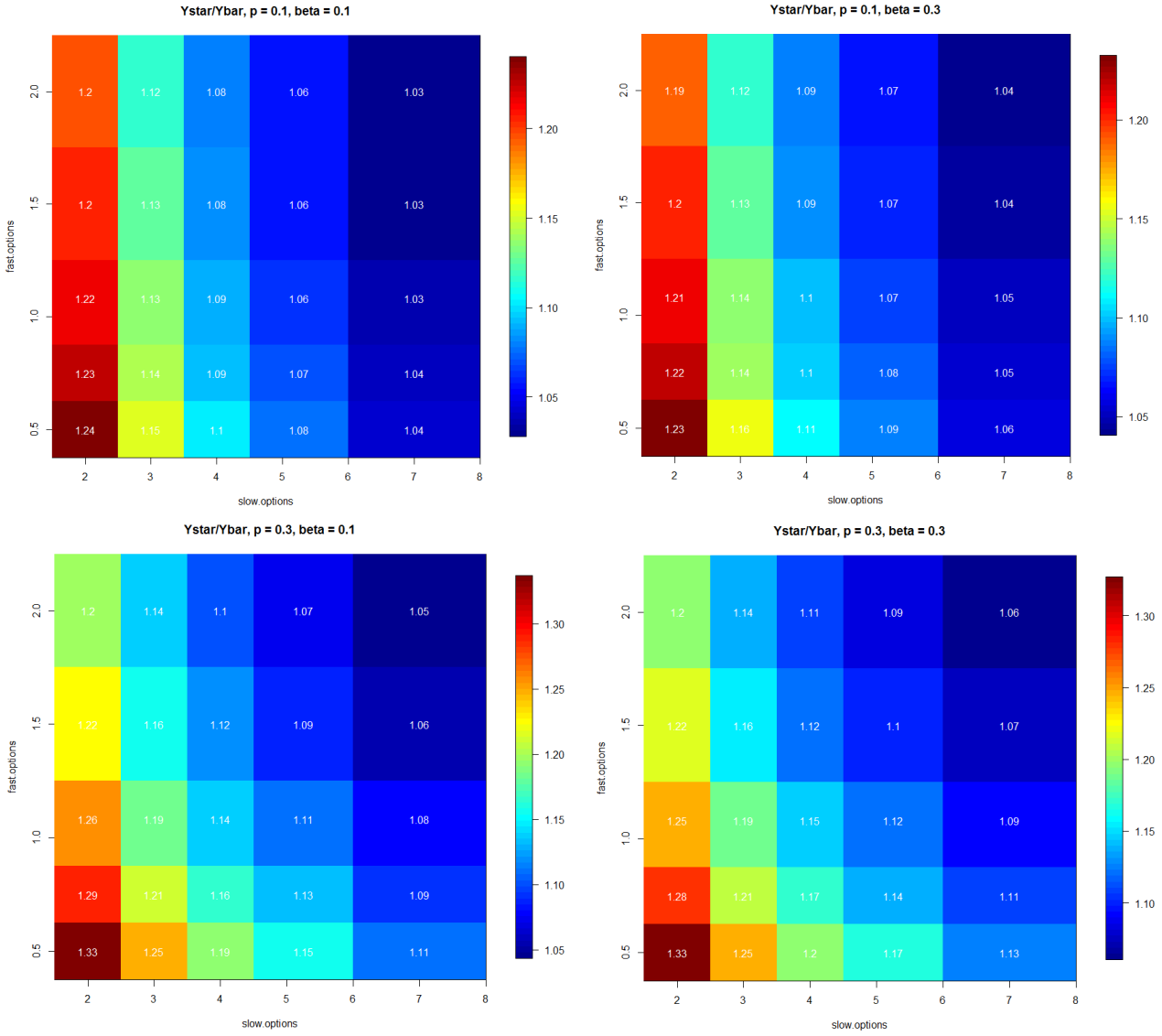
Each simulated screening trial was conducted using the same approach as described in Section 4.1. In summary, preclinical durations were simulated from the specified mixture of two exponential distributions, the time between the initiations of two preclinical durations was generated as a Poisson process at a rate of 210 new cancers per year, and any case whose preclinical duration overlapped with our screening window was included in both the treatment and control arm data (presupposing no overdiagnosis). For this study, we used a screening program of five annual screens, so screening occurred at the start of the simulated trial, as well as at the end of the first, second, third and fourth years.

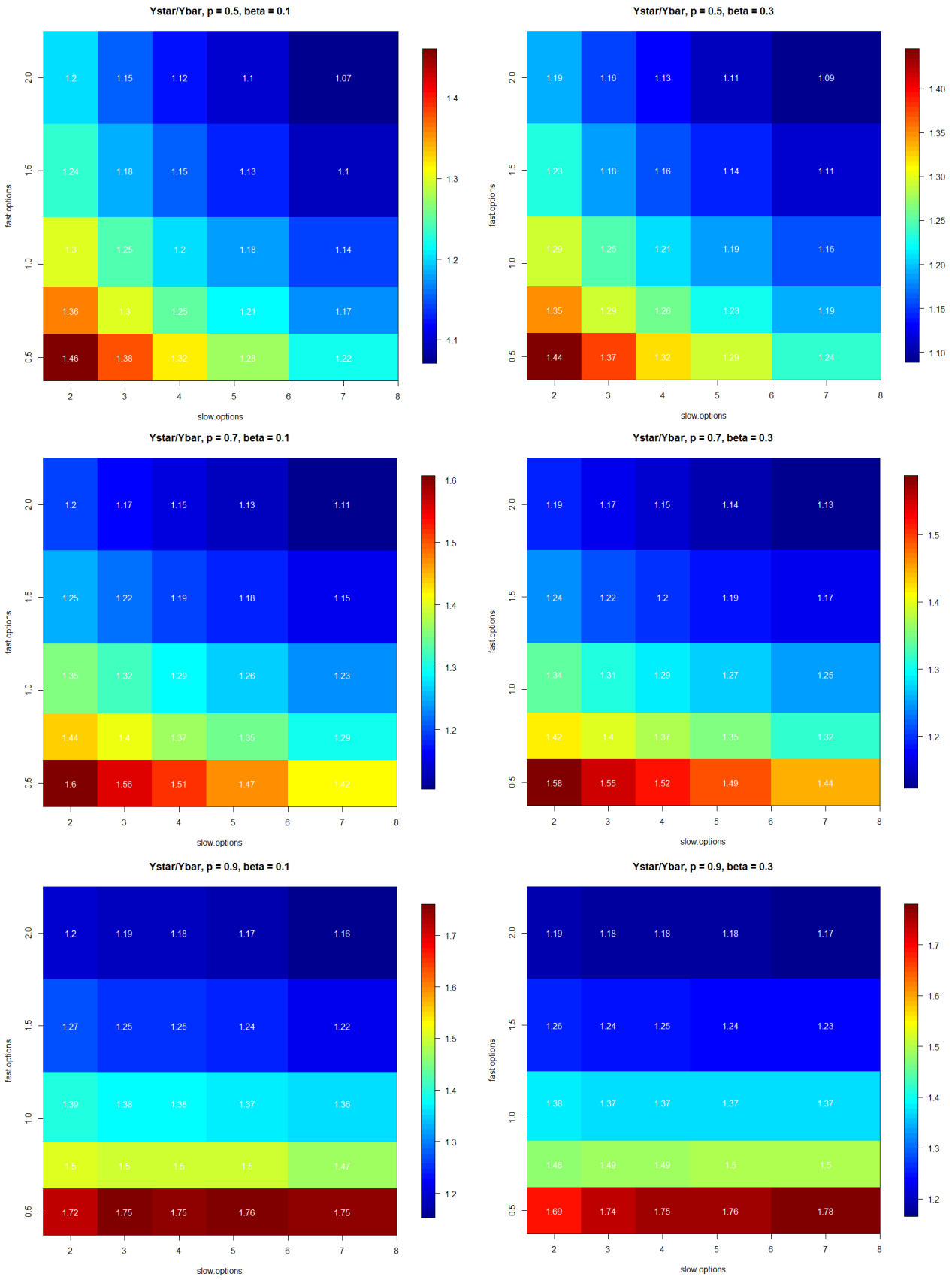
At each screen, cases were tested for cancer with sensitivity that was a function of preclinical growth starting at the specified initial sensitivity. A more detailed description of simulating the screening process can be reviewed in Section 4.1 or Figure 2 of Kafadar and Prorok [46]. The response variable for each simulated trial was $\bar{Y}^*/\bar{Y}$, the observed ratio of mean screen-detected preclinical duration to overall mean preclinical duration.

Figure 6.1 presents heatmaps of our study results. Each individual heatmap pertains to a certain combination of p and starting sensitivity, showing changes in $\bar{Y}^*/\bar{Y}$ with the mean slow-developing preclinical duration in years $(1/(\lambda_S * 365))$ on the x-axis and the mean fast-developing preclinical duration in years $(1/(\lambda_F * 365))$ on the y-axis. Moving from the left heatmap to the right heatmap shows the effect of an increase in starting test sensitivity from 10% to 30%, and moving down the rows of heatmaps shows the effect of an increase in the proportion of fast-developing cases p. Overall, $\bar{Y}^*/\bar{Y}$ ranges from 1.03 to 1.78. A low value of $\bar{Y}^*/\bar{Y}$ like 1.03 suggests that the screen-detected preclinical durations are not much longer than average, which means that the length-biased sampling effect is minimal. On the other hand, a value of $\bar{Y}^*/\bar{Y}$ as high as 1.78 suggests that the mean screen-detected preclinical duration is over 75% longer than the average preclinical duration, in which case the length-biased sampling effect is large.

There heatmaps showed several noteworthy trends. First, it appears that the length-biased sampling effect is more severe when the preclinical durations are shorter. $\bar{Y}^*/\bar{Y}$ increases when the proportion of fast-developing cases increases, the mean fast-developing preclinical duration decreases, or the mean slow-developing preclinical duration decreases. Second, it seems that there is an interactive effect between the preclinical duration distribution parameters on $\bar{Y}^*/\bar{Y}$. When p is low, changes to λ_S have a

<u>Figure 6.1</u>: Heatmaps of the length-biased sampling effect $\bar{Y}^*/\bar{Y}$ for different combinations of preclinical duration distribution parameters and starting test sensitivity.




Ystar/Ybar, p = 0.5, beta = 0.1
1.2 1.15 1.12 1.1 1.07
1.24 1.18 1.15 1.13 1.1
1.3 1.25 1.2 1.18 1.14
1.36 1.3 1.25 1.21 1.17
1.46 1.38 1.32 1.28 1.22
fast.options
slow.options
Ystar/Ybar, p = 0.5, beta = 0.3
1.19 1.16 1.13 1.11 1.09
1.23 1.18 1.16 1.14 1.11
1.29 1.25 1.21 1.19 1.16
1.35 1.29 1.26 1.23 1.19
1.44 1.37 1.32 1.29 1.24
fast.options
slow.options
Ystar/Ybar, p = 0.7, beta = 0.1
1.2 1.17 1.15 1.13 1.11
1.25 1.22 1.19 1.18 1.15
1.35 1.32 1.29 1.26 1.23
1.44 1.4 1.37 1.35 1.29
1.6 1.56 1.51 1.47 1.42
fast.options
slow.options
Ystar/Ybar, p = 0.7, beta = 0.3
1.19 1.17 1.15 1.14 1.13
1.24 1.22 1.2 1.19 1.17
1.34 1.31 1.29 1.27 1.25
1.42 1.4 1.37 1.35 1.32
1.58 1.55 1.52 1.49 1.44
fast.options
slow.options
Ystar/Ybar, p = 0.9, beta = 0.1
1.2 1.19 1.18 1.17 1.16
1.27 1.25 1.25 1.24 1.22
1.39 1.38 1.38 1.37 1.36
1.5 1.5 1.5 1.5 1.47
1.72 1.75 1.75 1.76 1.75
fast.options
slow.options
Ystar/Ybar, p = 0.9, beta = 0.3
1.19 1.18 1.18 1.18 1.17
1.26 1.24 1.25 1.24 1.23
1.38 1.37 1.37 1.37 1.37
1.48 1.49 1.49 1.5 1.5
1.69 1.74 1.75 1.76 1.78
fast.options
slow.options

large impact on the length-biased sampling effect, while changes to λ_F do not. When p is high, changes to λ_F have a large impact on the length-biased sampling effect, while changes to λ_S do not. Finally, varying the starting test sensitivity appears to result in a minimal change in $\bar{Y}^*/\bar{Y}$, which suggests that realistic changes to the sensitivity model do not have a major impact on the length-biased sampling effect.

To determine the significance of these associations viewed in the heatmaps, we performed an ANOVA on the variable $\bar{Y}^*/\bar{Y}$. Analysis of the log-ratio may have better interpretability in some scenarios, but the ANOVA results see minimal change. The ANOVA model included the three preclinical duration distribution parameters and starting test sensitivity, as well as all two-, three-, and four-way interactions. We treated each of the three preclinical duration distribution parameters as quantitative variables, rather than categorical, and estimated a linear effect for each; we had found that there was little gain in sum of squares relative to the loss in degrees of freedom when treating these parameter levels as categorical variables. Figure 6.2 displays the ANOVA table for this analysis. All four factors, as well as most of the two-way interactions and some of the three-way interactions, had a statistically significant effect on $\bar{Y}^*/\bar{Y}$, though the three preclinical duration distribution parameters had a much greater effect on $\bar{Y}^*/\bar{Y}$ than the starting test sensitivity. Of the preclinical parameters, the proportion of fast-developing cases p had the greatest effect, followed by the rate parameter for the fast-developing cases λ_F and, finally, the rate parameter for the slow-developing cases λ_S. The most significant interaction terms also corresponded to those two- and three-way interactions between preclinical parameters, confirming our observation from the heatmaps that there is an interactive effect between the preclinical duration distribution parameters on $\bar{Y}^*/\bar{Y}$.

Figure 6.2: Analysis of variance results for the effect of preclinical duration distribution parameters and starting test sensitivity on $\bar{Y}^*/\bar{Y}$.

Factor	df	SS	MS	F	p-value
Sensitivity	1	0.1	0.1	23.8	<0.0001
λ_F	1	173.0	173.0	56,109.8	<0.0001
λ_S	1	42.2	42.2	13,688.1	<0.0001
p	1	274.1	274.1	88,934.3	<0.0001
Sensitivity $*$ λ_F	1	0.0	0.0	0.9	0.34
Sensitivity $*$ λ_S	1	0.6	0.6	189.5	<0.0001
Sensitivity $*$ p	1	0.0	0.0	11.4	0.0007
$\lambda_F * \lambda_S$	1	0.6	0.6	199.3	<0.0001
$\lambda_F * p$	1	96.5	96.5	31,304.3	<0.0001
$\lambda_S * p$	1	7.7	7.7	2,493.4	<0.0001
Sensitivity $*$ $\lambda_F * \lambda_S$	1	0.0	0.0	6.5	0.01
Sensitivity $*$ $\lambda_F * p$	1	0.0	0.0	0.1	0.75
Sensitivity $*$ $\lambda_S * p$	1	0.0	0.0	9.1	0.003
$\lambda_F * \lambda_S * p$	1	0.2	0.2	67.2	<0.0001
Sensitivity $*$ $\lambda_F * \lambda_S * p$	1	0.0	0.0	3.7	0.05
Error	24,984	77.0	0.0		

Abbreviations: df, degrees of freedom; SS, sum of squares; MS, mean squares; F, F-statistic.

It is worth noting that some factors and their interactions appeared to be highly statistically significant despite explaining very little of the variability in the length-biased sampling effect. An explanation for this phenomenon is the extremely large sample size ofthis experiment; with 250 possible combinations of factor levels and 100 replicates of each scenario, our total sample size was 25,000 simulated screening trials. This drove the residual degrees of freedom very high, resulting in low mean squared error, which allows even the smallest of effects to appear statistically significant. This explains why a factor like starting sensitivity, which had minimal effect on $\bar{Y}^*/\bar{Y}$ in the heatmaps and a low sum of squares in the ANOVA, could be deemed significant.

6.2 Simulation Study: Preclinical Duration Distribution and Screening Program on $E[Y^*]/E[Y]$

We discovered in our first simulation study in Section 6.1 that changes to preclinical duration distribution parameters had a great impact on the length-biased sampling effect $E[Y^*]/E[Y]$, whereas realistic changes to the starting value of a function that models test sensitivity based on preclinical growth had minimal impact on $E[Y^*]/E[Y]$. These results suggest that, as long as the model for screening test sensitivity starts within a reasonable range, we can focus primarily on the preclinical duration parameters when quantifying the length-biased sampling effect. However, we know from Kafadar and Prorok [29] and Heltshe et al [12] that $E[Y^*]/E[Y]$ is also highly affected by the screening program applied in a trial. Because we want to reexamine the conclusions from these two studies when the preclinical durations are generated from a mixture of two exponential distributions, we now also want to assess the effect of the screening program on $E[Y^*]/E[Y]$.

In our second simulation study, the three preclinical duration distribution parameters (p, λ_F, and λ_S) and the screening program are varied to evaluate changes in the length-biased sampling effect. This second study was again conducted using a factorial design, just as the first. Each of the preclinical duration distribution parameters was allowed to vary between the same five levels as defined in Section 6.1. With regards to the screening program, each case would be screened for seven years, but the screening interval had three levels: $\Delta = 1$, 2, or 3. When $\Delta = 1$, subjects were screened at the start of the trial, as well as after years 1, 2, 3, 4, 5, and 6, for a total of seven screens; when $\Delta = 2$, subjects were screened at the start of the trial, as well as after years 2, 4, and 6, for a total of four screens; and, when $\Delta = 3$, subjects were screened at the start of the trial, as well as after

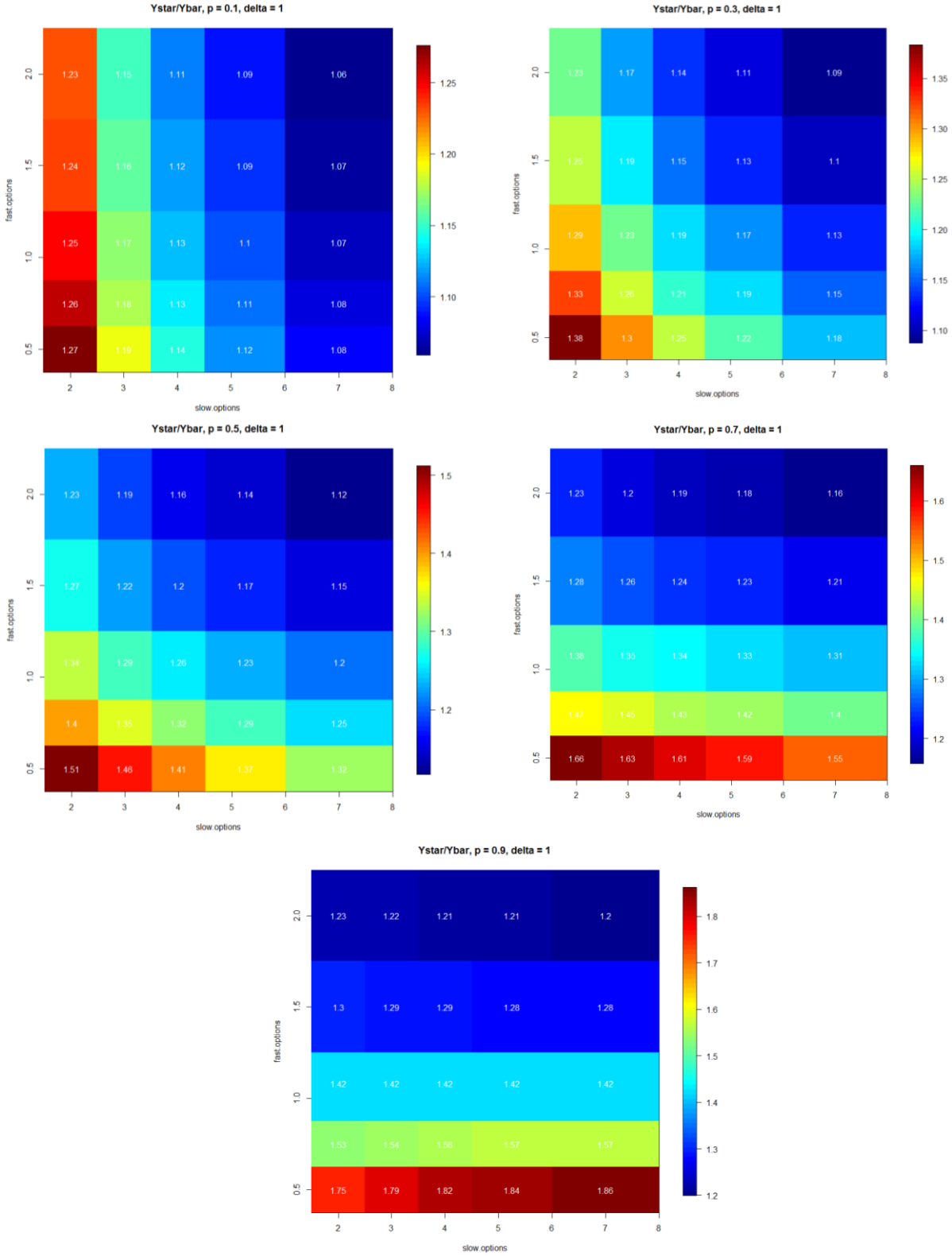
years 3 and 6, for a total of three screens. With five levels for each preclinical duration distribution parameter and three levels for the screening interval, we created 375 possible combinations, and each scenario was simulated 100 times for replication of the results.

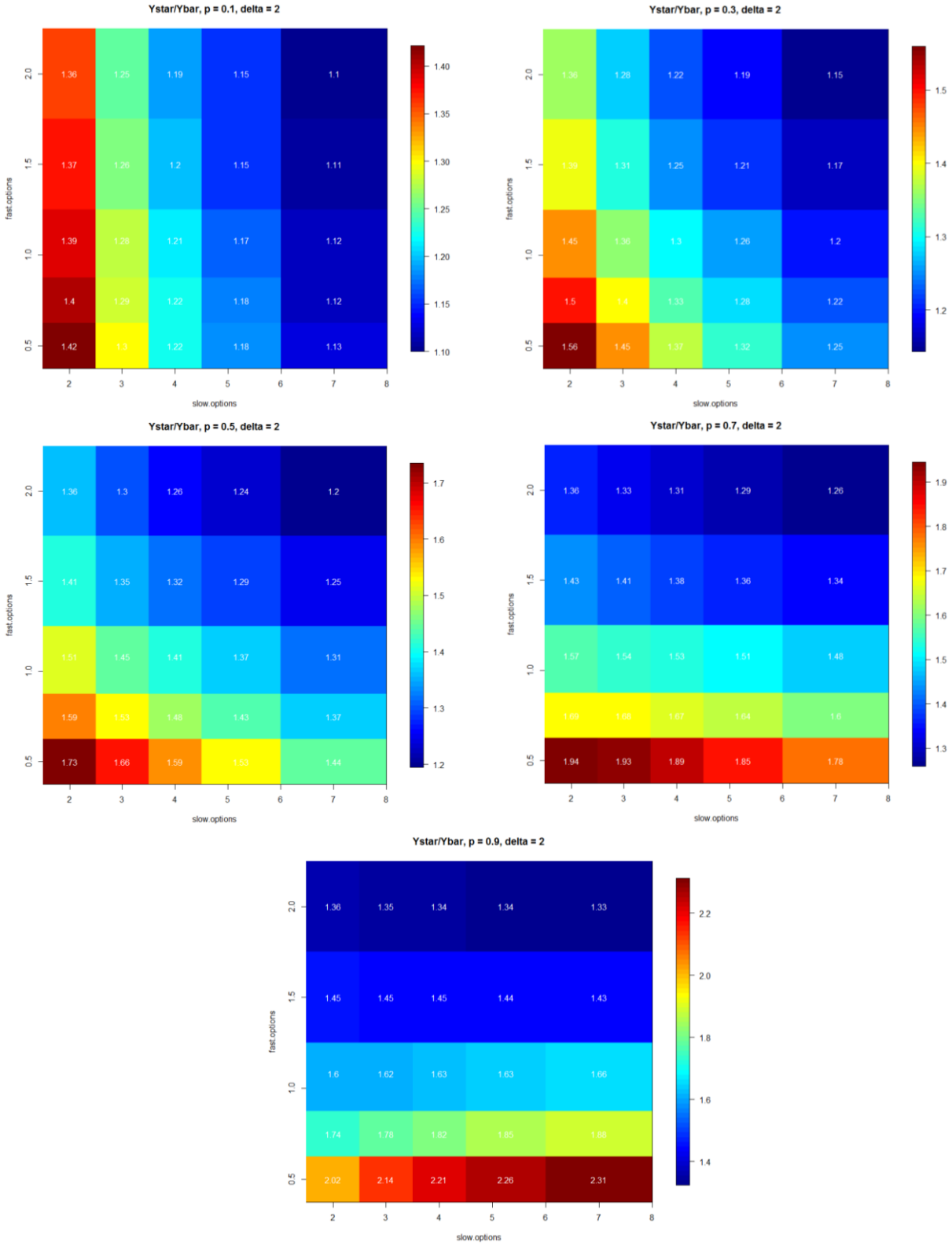
This second simulation study was carried out using the exact same approach as the first one described in Section 6.1. However, the screening program now varied while the function for test sensitivity remained constant. For this study, sensitivity begins at 20% at the start of a subject's preclinical duration and first reaches the maximum sensitivity of 90% three-quarters of the way through the preclinical duration; the increase in test sensitivity with preclinical growth was modelled using a Normal cdf function.

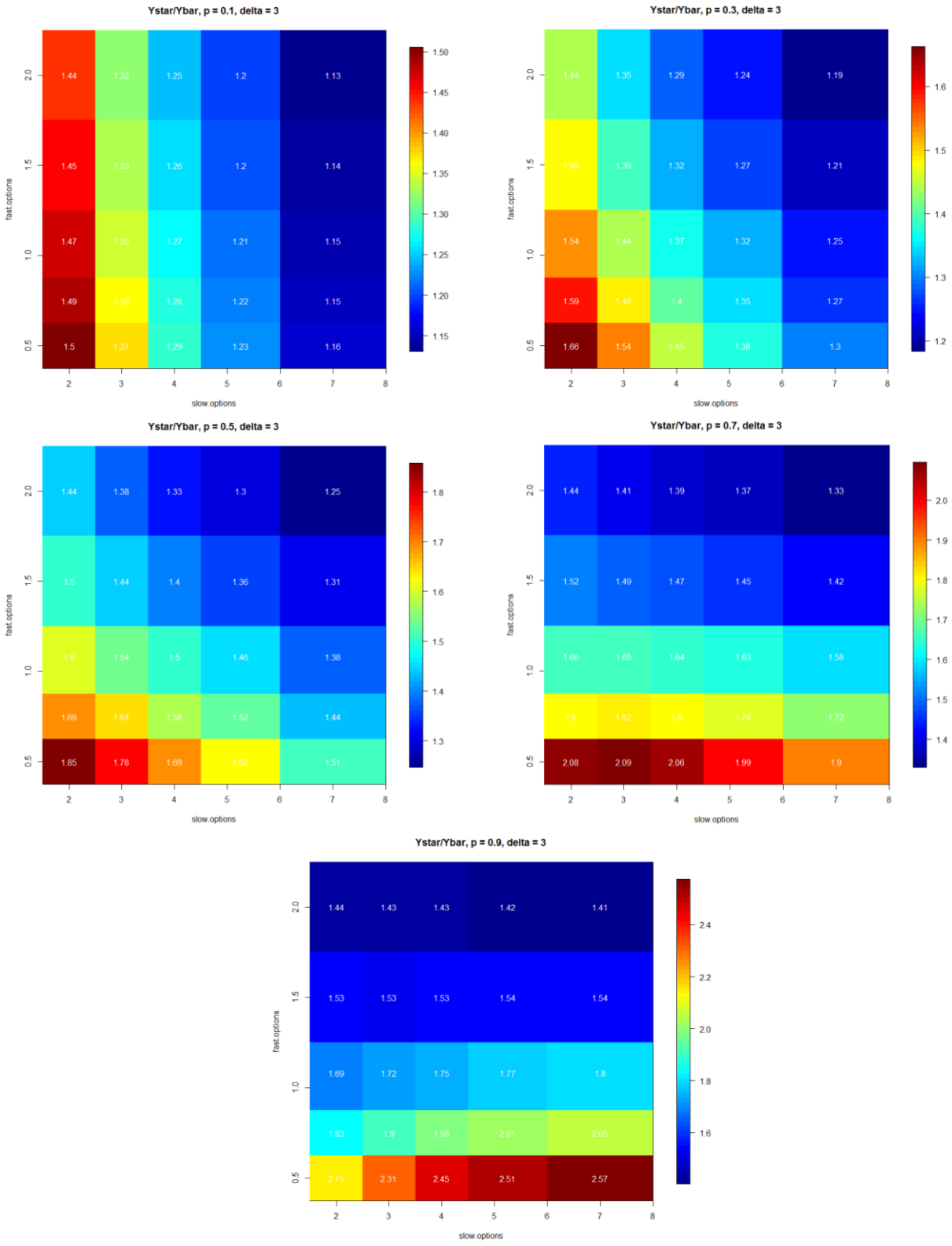
Figure 6.3 presents heatmaps of our study results. Each individual heatmap pertains to a certain combination of p and Δ, showing changes in $\bar{Y}^*/\bar{Y}$ with the mean slow-developing preclinical duration in years $(1/(\lambda_S * 365))$ on the x-axis and the mean fast-developing preclinical duration in years $(1/(\lambda_F * 365))$ on the y-axis. Moving from left-to-right and top-to-bottom on a page of heatmaps shows the effect of an increase in the proportion of fast-developing cases p, and moving from page-to-page of heatmaps shows the effect of an increase in the screening interval Δ. Overall, $\bar{Y}^*/\bar{Y}$ ranges from 1.06 to 2.57. While the smallest observed ratio is roughly the same as in the first simulation study, the largest is much higher now that Δ may be as large as three years. This suggests that the screening program can have a major impact on the length-biased sampling effect in a screening trial. A value of $\bar{Y}^*/\bar{Y}$ greater than 2 implies that the mean screen-detected preclinical duration is over twice as long as the average preclinical duration.

Generally, the scenarios with these extremely large length-biased sampling effects tend to match an unrealistically long screening interval with a very fast-growing cancer.

Figure 6.3: Heatmaps of the length-biased sampling effect $\bar{Y}^*/\bar{Y}$ for different combinations of preclinical duration distribution parameters and screening interval.




Ystar/Ybar, p = 0.1, delta = 2
fast.options
slow.options
2.0 1.5 1.0 0.5
2 3 4 5 6 7 8
1.36 1.25 1.19 1.15 1.1
1.37 1.26 1.2 1.15 1.11
1.39 1.28 1.21 1.17 1.12
1.4 1.29 1.22 1.18 1.12
1.42 1.3 1.22 1.18 1.13
1.40 1.35 1.30 1.25 1.20 1.15 1.10
Ystar/Ybar, p = 0.3, delta = 2
fast.options
slow.options
2.0 1.5 1.0 0.5
2 3 4 5 6 7 8
1.36 1.28 1.22 1.19 1.15
1.39 1.31 1.25 1.21 1.17
1.45 1.36 1.3 1.26 1.2
1.5 1.4 1.33 1.28 1.22
1.56 1.45 1.37 1.32 1.25
1.5 1.4 1.3 1.2
Ystar/Ybar, p = 0.5, delta = 2
fast.options
slow.options
2.0 1.5 1.0 0.5
2 3 4 5 6 7 8
1.36 1.3 1.26 1.24 1.2
1.41 1.35 1.32 1.29 1.25
1.51 1.45 1.41 1.37 1.31
1.59 1.53 1.48 1.43 1.37
1.73 1.66 1.59 1.53 1.44
1.7 1.6 1.5 1.4 1.3 1.2
Ystar/Ybar, p = 0.7, delta = 2
fast.options
slow.options
2.0 1.5 1.0 0.5
2 3 4 5 6 7 8
1.36 1.33 1.31 1.29 1.26
1.43 1.41 1.38 1.36 1.34
1.57 1.54 1.53 1.51 1.48
1.69 1.68 1.67 1.64 1.6
1.94 1.93 1.89 1.85 1.78
1.9 1.8 1.7 1.6 1.5 1.4 1.3
Ystar/Ybar, p = 0.9, delta = 2
fast.options
slow.options
2.0 1.5 1.0 0.5
2 3 4 5 6 7 8
1.36 1.35 1.34 1.34 1.33
1.45 1.45 1.45 1.44 1.43
1.6 1.62 1.63 1.63 1.66
1.74 1.78 1.82 1.85 1.88
2.02 2.14 2.21 2.26 2.31
2.2 2.0 1.8 1.6 1.4


Ystar/Ybar, p = 0.1, delta = 3
1.44 1.32 1.25 1.2 1.13
1.45 1.33 1.26 1.2 1.14
1.47 1.35 1.27 1.21 1.15
1.49 1.36 1.28 1.22 1.15
1.5 1.37 1.29 1.23 1.16
fast.options
slow.options
Ystar/Ybar, p = 0.3, delta = 3
1.44 1.35 1.29 1.24 1.19
1.48 1.39 1.32 1.27 1.21
1.54 1.44 1.37 1.32 1.25
1.59 1.49 1.4 1.35 1.27
1.66 1.54 1.45 1.38 1.3
fast.options
slow.options
Ystar/Ybar, p = 0.5, delta = 3
1.44 1.38 1.33 1.3 1.25
1.5 1.44 1.4 1.36 1.31
1.6 1.54 1.5 1.46 1.38
1.69 1.64 1.58 1.52 1.44
1.85 1.78 1.69 1.62 1.51
fast.options
slow.options
Ystar/Ybar, p = 0.7, delta = 3
1.44 1.41 1.39 1.37 1.33
1.52 1.49 1.47 1.45 1.42
1.66 1.65 1.64 1.63 1.58
1.8 1.82 1.8 1.78 1.72
2.08 2.09 2.06 1.99 1.9
fast.options
slow.options
Ystar/Ybar, p = 0.9, delta = 3
1.44 1.43 1.43 1.42 1.41
1.53 1.53 1.53 1.54 1.54
1.69 1.72 1.75 1.77 1.8
1.83 1.9 1.98 2.01 2.05
2.15 2.31 2.45 2.51 2.57
fast.options
slow.options

For example, the largest value of $\bar{Y}^*/\bar{Y}$ occurs with $p = 0.9$, $\lambda_F = 1/(0.5 * 365)$, $\lambda_s = 1/(7 * 365)$, and $\Delta = 3$; in this scenario, the mean preclinical duration is 1.15 years, yet patients are only being screened every three years. In reality, such a fast-developing cancer would have to be screened more often, as far too many subjects would be diagnosed in between screens with such a large interval Δ. When evaluating the length-biased sampling effects estimated by our simulations, it is important to recognize that not all scenarios match a reasonable screening interval to the distribution of preclinical durations.

As with the previous heatmaps, these heatmaps suggest several important trends, some familiar from the first simulation study and some new. First, it still appears that the length-biased sampling effect is more severe when the preclinical durations are shorter and that there is an interactive effect between the preclinical duration distribution parameters on $\bar{Y}^*/\bar{Y}$. Second, the length-biased sampling effect is also more severe when the screening interval is longer. With a greater time period in between screens, it is likely that fewer cases are being screen-detected and more interval cases are arising. As more fast-developing cases are diagnosed in between screens, the mean screen-detected preclinical duration $\bar{Y}^*$ will increase. Finally, it seems that there may be an interactive effect between preclinical duration distribution parameters and the screening interval, as well. The magnitude of changes $\bar{Y}^*/\bar{Y}$ resulting from changes in a preclinical parameter appears to be dependent on the value of Δ, such that small changes to a preclinical parameter have a greater effect on $\bar{Y}^*/\bar{Y}$ when Δ is larger.

We may also compare results in Figure 6.1 to similar scenarios with $\Delta = 1$ in Figure 6.3. If all preclinical duration distribution parameters are the same between two comparison scenarios, then the only difference is the starting sensitivity (10 or 30% for our

first simulation study vs. 20% for our second simulation study) and the number of annual screens (five for our first simulation study vs. seven for our second simulation study). Because we know that the starting sensitivity has a minimal effect on $\bar{Y}^*/\bar{Y}$, any differences we see are most likely explained by the change in the number of screens. For most cases, we see that adding additional screens increases the length-biased sampling effect, though this increase is usually very small, no larger than 0.05. However, in extreme scenarios where the preclinical durations are predominantly very fast with a small fraction of very slow cases – consider $p = 0.9$, $1/(\lambda_F * 365) = 0.5$, and $1/(\lambda_S * 365) = 7$ – the increase in number of screens leads to a much larger increase in $\bar{Y}^*/\bar{Y}$ greater than 0.1. In general, we can conclude that adding additional screens will augment the length-biased sampling effect in a trial, though this increase is usually trivial.

To determine the significance of these associations viewed in Figure 6.3, we performed an ANOVA on the variable $\bar{Y}^*/\bar{Y}$. Analysis of the log-ratio may have better interpretability in some scenarios, but the ANOVA results see minimal change. The ANOVA model included the three preclinical duration distribution parameters and screening interval, as well as all two-, three-, and four-way interactions. Similar to the ANOVA for our first simulation study in Section 6.1, we treated each of the four factors as quantitative variables, rather than categorical, and estimated a linear effect for each, as there was little gain in sum of squares relative to the loss in degrees of freedom when treating these variables as categorical. Figure 6.4 displays the ANOVA table for this analysis. All four factors, as well as the large majority of the interaction terms, had statistically significant effects on $\bar{Y}^*/\bar{Y}$. Of the four individual effects, the proportion of fast-developing cases p had the greatest impact, followed by the rate parameter for the fast-developing cases λ_F, the

Figure 6.4: Analysis of variance results for the effect of preclinical duration distribution parameters and screening interval on $\bar{Y}^*/\bar{Y}$.

Factor	df	SS	MS	F	p-value
Δ	1	360.4	360.4	55,800.0	<0.0001
λ_F	1	561.2	561.2	86,870.0	<0.0001
λ_S	1	84.3	84.3	13,060.0	<0.0001
p	1	885.9	885.9	137,100.0	<0.0001
$\Delta * \lambda_F$	1	16.1	16.1	2,497.0	<0.0001
$\Delta * \lambda_S$	1	2.6	2.6	403.2	<0.0001
$\Delta * p$	1	35.9	35.9	5,559.0	<0.0001
$\lambda_F * \lambda_S$	1	0.0	0.0	0.1	0.72
$\lambda_F * p$	1	298.4	298.4	46,190.0	<0.0001
$\lambda_S * p$	1	55.9	55.9	8,661.0	<0.0001
$\Delta * \lambda_F * \lambda_S$	1	0.1	0.1	9.4	0.002
$\Delta * \lambda_F * p$	1	10.7	10.7	1,663.0	<0.0001
$\Delta * \lambda_S * p$	1	6.0	6.0	926.6	<0.0001
$\lambda_F * \lambda_S * p$	1	4.6	4.6	708.1	<0.0001
$\Delta * \lambda_F * \lambda_S * p$	1	0.8	0.8	118.0	<0.0001
Error	37,484	242.1	0.0		

Abbreviations: df, degrees of freedom; SS, sum of squares; MS, mean squares; F, F-statistic.

screening interval Δ, and, finally, the rate parameter for the slow-developing cases λ_S. Moreover, as we hypothesized from the heatmaps, there were significant interactions between the preclinical parameters, as well as significant interactions between preclinical parameters and the screening interval. However, some of the statistically significant interaction terms actually had low sums of squares, suggesting that they explain very little of the variability in $\bar{Y}^*/\bar{Y}$; nonetheless, even the smallest effects would be determined statistically significant given our sample size of 37,500 simulated trials.

6.3 Sensitivity Analysis

In Section 6.2, we carried out a simulation study and generated heatmaps that could be used to estimate the length-biased sampling effect given a certain screening program and preclinical duration distribution. Because we derived a method for estimating preclinical duration distribution parameters using the pattern of diagnoses from real trial data in Section 5, we can now use such parameter estimates to quantify $E[Y^*]/E[Y]$ in said trials. However, we know that our method for estimating preclinical duration parameters is not exact, and the parameters can be misestimated in some cases. As such, it is important to assess the sensitivity of the estimate of the length-biased sampling effect to a misspecification of preclinical duration distribution parameters.

For this analysis, we will revisit the eight simulated examples from Section 5.3. For each scenario, we know the true preclinical duration distribution parameters, and these can be used to determine the true mean length-biased sampling effect $E[Y^*]/E[Y]$ for such a distribution. We also have the estimated parameters from each simulated trial scenario, and we can also use our heatmaps to determine the mean length-biased sampling effect for this estimated distribution. Comparing the two results will allow us to evaluate the sensitivity of $E[Y^*]/E[Y]$ to parameter misspecification. It is worth noting that the simulated examples in Section 5.3 were trials of four annual screens, whereas the estimates of $E[Y^*]/E[Y]$ from Section 6.2 were from trials of seven annual screens. Although $E[Y^*]/E[Y]$ would change with the number of screens, we will ignore this for now, as our primary concern is the sensitivity of $E[Y^*]/E[Y]$ to preclinical duration distribution parameter misspecification.

Figure 6.5: Sensitivity of $E[Y^*]/E[Y]$ to misspecification of the preclinical duration distribution parameters.

(a) Comparing the true parameters to their estimates for each scenario with differences highlighted in red

Scenario	True Parameters			Estimated Parameters		
	p	$\frac{1}{\lambda_F * 365}$	$\frac{1}{\lambda_S * 365}$	$\hat{p}$	$\frac{1}{\hat{\lambda}_F * 365}$	$\frac{1}{\hat{\lambda}_S * 365}$
A	**0.3**	**1.5**	**5**	0.3	1	5
B	**0.1**	**0.5**	**5**	0.3	1.5	5
C	**0.5**	**1**	**7**	0.3	1	7
D	**0.5**	**0.5**	**7**	0.5	0.75	7
E	**0.5**	**1**	**4**	0.5	1	3
F	**0.5**	**0.5**	**4**	0.5	1	5
G	**0.9**	**0.75**	**7**	0.9	0.75	4
H	**0.7**	**0.75**	**3**	0.7	1	3

(b) Comparing the true $E[Y^*]/E[Y]$ to the estimate under misspecified parameters with relative errors in parenthesis.

Scenario	True $E[Y^*]/E[Y]$	Estimated $E[Y^*]/E[Y]$
A	**1.13**	1.17 (4%)
B	**1.12**	1.13 (1%)
C	**1.20**	1.13 (6%)
D	**1.32**	1.25 (5%)
E	**1.26**	1.29 (2%)
F	**1.41**	1.23 (13%)
G	**1.57**	1.56 (1%)
H	**1.45**	1.35 (7%)

Because not all of our parameter estimates match with one of the five levels we tested for each parameter when calculating $E[Y^*]/E[Y]$ in Section 6.2, they were rounded to a nearby level for this sensitivity analysis. The true parameters and rounded parameter estimates are displayed in Figure 6.5a. We see that one parameter is misspecified in some scenarios, while multiple parameters may be misestimated in others. Compared to the actual parameter estimates in Figure 5.6 of Section 5.3, we see that our rounding generally worsens the magnitude of prediction errors for our parameters. This suggests that any

differences we see between true and estimated $E[Y^*]/E[Y]$ for our example scenarios are likely to be larger than the differences we would see in practice from reasonable parameter misspecification.

Figure 6.5b compares $E[Y^*]/E[Y]$ under the true and estimated parameter values for each scenario. We see that the effect of parameter misspecification is small in most scenarios with five of the eight relative errors less than 5%. The relative errors are larger for the misestimates in scenarios C, F, and H. The error in scenario C is driven by a misspecification of the proportion of fast-developing cases, which matches with our conclusion from Figure 6.4 that this parameter has the greatest effect on $E[Y^*]/E[Y]$. The proportion of fast-developing cases is also misspecified in scenario B, but this overestimation appears to be balanced out by simultaneously overestimating the mean of these fast-developing cases. In scenario F, both subgroup means are overestimated while the proportion of fast-developing cases is correctly specified, meaning that preclinical durations from the estimated distribution would be much longer than preclinical durations from the true distribution, which explains the underestimated length-biased sampling effect. Finally for scenario H, the mean of the fast-developing cases is misspecified, and, because the distribution is predominantly fast-developing, this has a larger effect on $E[Y^*]/E[Y]$. We conclude that $E[Y^*]/E[Y]$ is not particularly sensitive to minor preclinical duration distribution parameter misspecifications unless these errors pertain to the proportion of fast-developing cases or the mean of the larger component in the mixture distribution.

6.4 Discussion

Section 5 found that the parameters of an exponential mixture distribution of preclinical durations for a certain cancer could be estimated using the pattern of diagnoses in a screening trial. In Section 6, we set out to estimate the length-biased sampling effect on screen-detected preclinical durations in a trial based on their underlying distribution. The simulation study results in Sections 6.1 and 6.2 suggest than the preclinical duration distribution parameters and the screening interval all have a great impact on $E[Y^*]/E[Y]$, while reasonable changes to the model for screening test sensitivity have a minimal effect on the length-biased sampling effect. Furthermore, Section 6.3 found that major errors needed to be made in the specification of preclinical duration distribution parameters in order for the length-biased sampling effect to be significantly misestimated. Combining the results from Sections 6.1 and 6.3 suggests that misspecifying the test sensitivity has a minimal impact on the length-biased sampling effect, and any resulting misestimation of preclinical parameters due to misspecified sensitivity would also have a small effect on $E[Y^*]/E[Y]$ so long as the error is minor.

We compared our estimates for $E[Y^*]/E[Y]$ to the results from Kafadar and Prorok [29] in similar scenarios. Their study considered much more drastically different preclinical duration distributions, but their scenarios A, B, and E presented examples where the mean of the fast-developing cases, the mean of the slow-developing cases, and the proportion of fast-developing cases matched with a combination of parameter levels observed in our simulation studies. Comparing the results from Figure 5 in their paper to Figure 6.3 in ours shows that the differences in estimates for $E[Y^*]/E[Y]$ are moderate in most cases. One explanation for such differences could be that Kafadar and Prorok

assumed constant test sensitivity instead of determining sensitivity based on preclinical growth. However, the general proximity of the estimates between the two studies is particularly interesting because Kafadar and Prorok simulated preclinical durations from a mixture of gamma distributions, while we simulated preclinical durations from a mixture of exponential distributions. This suggests that the distribution type for preclinical durations may have a much lesser impact on the length-biased sampling effect than the mean(s) of said distribution. Although we did not consider a misspecification of the distribution type in our sensitivity analysis, it seems that such an error would have a minimal effect on $E[Y^*]/E[Y]$ so long as the distributions have similar means.

We also compared conclusions with Heltshe et al [12]. Heltshe et al found larger differences in the length-biased sampling effect when comparing trials with constant sensitivity to those where sensitivity was a function of preclinical growth, but smaller differences when minor realistic adjustments were made to this preclinical growth-related test sensitivity model. This is similar to our conclusion that adjusting the initial test sensitivity at the start of the preclinical duration had a minor effect on $E[Y^*]/E[Y]$. Additionally, we compared our estimates for $E[Y^*]/E[Y]$ to the results from Heltshe et al in similar scenarios. Coincidentally, scenarios A, B, and E again presented examples where the mean of the fast-developing cases, the mean of the slow-developing cases, and the proportion of fast-developing cases matched with a combination of parameter levels observed in our simulation studies. Comparing the results from Table 5 in their paper to Figure 6.3 in ours shows that the differences in estimates for $E[Y^*]/E[Y]$ are rather small. The minor differences we see could be attributed to Heltshe et al's more quickly increasing test sensitivity function or the fact that they simulated preclinical durations from a mixture

of lognormal distributions instead of exponentials. Regardless, the fact that these differences were so small again suggests that the distribution type for preclinical durations may have a much lesser impact on the length-biased sampling effect than the mean(s) of said distribution.

Because we did not specify the correlation between preclinical and clinical duration, we were unable to quantify the length-biased sampling effect on the screen-detected clinical durations, or $E[Z^*]/E[Z]$, as we do not know which clinical durations match with which preclinical durations. Remember that $E[Z^*]/E[Z]$ is particularly important for evaluating the effect of length-biased sampling on benefit time, as the benefit time compares survival from diagnosis between screened and unscreened subjects, and an estimate may be inflated if the screen-detected subjects have longer clinical durations than their unscreened counterparts. Focusing solely on simulating the diagnosis times from screening in the preclinical durations, we were able to estimate $E[Y^*]/E[Y]$, length-biased sampling effect on the screen-detected preclinical durations, which is informative for the effect of length-biased sampling on lead time. Lead time compares differences in time to diagnosis between screened and unscreened subjects, and an estimate may be inflated if the screen-detected subjects have longer preclinical durations than their unscreened counterparts. If we believe benefit time to be correlated with lead time, then knowing $E[Y^*]/E[Y]$ could still give an indication of $E[Z^*]/E[Z]$. However, this idea should be studied further, as several other factors, including the correlation between preclinical and clinical durations, likely affect the relationship between $E[Y^*]/E[Y]$ and $E[Z^*]/E[Z]$.

7. Conclusions and Future Research

7.1 Conclusions

Before a cancer screening test is recommended to the general at-risk population, it is evaluated through a randomized controlled screening trial, where estimates of the mean lead and benefit times from screening may be calculated. However, cancers with longer preclinical durations are more likely to be screen-detected than faster-developing cases, so the screen-detected cancers in a trial are a length-biased sample of all cancers in the population, which causes the aforementioned measures to be overestimated. Furthermore, we know that the severity of the length-biased sampling effect is highly dependent on the joint distribution of preclinical and clinical durations from previous simulation studies [12, 29]. Thus, we set out to develop better methods of estimating these distributions so that the length-biased sampling effect on a real trial can be more accurately assessed.

In Section 3, we analyzed clinical durations from the control arms of three historic screening trials – the HIP trial and the lung and ovarian cancer screening arms of the PLCO – and found that no common survival distribution approximated the data well. This suggested that clinical durations may be modelled better by a mixture distribution, where one group of cases is faster-developing and the other is slower-developing. We developed an iterative procedure to estimate the three parameters in a mixture of two exponential distributions and recommended common sense adjustments to make in the presence of heavy censoring. Our method would be best described as exploratory and user-driven, and it is involves no likelihood maximization like the E-M algorithm commonly used to specify mixture distribution parameters, but it is much simpler to use. Applying our procedure, we

found that a mixture of two exponential distributions approximated the clinical durations well for all three historic screening trials, and we believe that our method could also be applied to other survival time data in the future.

Discovering that clinical durations for multiple cancer types could be modelled using a mixture of two exponential distributions was a key result for our research, as we leveraged this finding to generate preclinical durations from the same type of distribution. The seemingly strong correlation between a subject's preclinical and clinical duration made us feel comfortable with this distribution assumption. Although preclinical durations are unobservable, making parameter estimation more difficult than it is for clinical durations, we believed that the mean lead time and pattern of diagnoses in a trial would allude to the preclinical parameter values. Our simulation study in Section 5 illustrated a highly significant interactive effect between the three parameters of a mixture of two exponential distributions on trial outcomes. We used these simulated trials to train a multivariate regression model to predict the preclinical duration distribution parameters based on the mean lead time, proportion of interval cases, and proportion of cases diagnosed at each screen in a trial. Our model was able to predict the preclinical duration distribution parameters fairly accurately in most scenarios, though it did struggle somewhat understandably when estimating the mean of the smaller component of the mixture distribution in extremely imbalanced scenarios.

Returning to the primary goal of our research, we wanted to assess the length-biased sampling effect under a variety of preclinical duration distribution scenarios in Section 6. We found that changes to this distribution had a significant effect on the mean length of screen-detected preclinical durations relative to the average preclinical duration.

Specifically, when the preclinical durations become shorter due to a change in any distribution parameter, the length-biased sampling effect becomes more severe. Moreover, the length-biased sampling effect becomes more severe for a given preclinical duration distribution when the screening interval is increased. While these findings are similar to previous studies, the advantage of our approach is that the length-biased sampling effect for a real trial could be estimated from the distribution of preclinical durations in that trial, which we can predict using the model from Section 5.

We recognized that our model predictions for preclinical duration distribution parameters in Section 5 were not exact, and we showed in Section 4 that realistic changes to the assumed test sensitivity function could also have an impact on the trial outcomes we used to predict these parameters. To evaluate the impact of such misspecifications on the estimated length-biased sampling effect, we carried out a sensitivity analysis in Section 6.3. Test sensitivity misspecification had a trivial impact on the length-biased sampling effect, and we found that larger errors needed to be made in estimating the mixing proportion or the mean of the larger mixture distribution component in order for the length-biased sampling effect to be significantly misestimated for a given preclinical duration distribution. Overall, our results from the limited scenarios we considered show promise towards being able to quantify the length-biased sampling effect on a real screening trial.

7.2 Challenges and Future Research

Although our research has produced many novel and significant results, the project was not without its challenges, and we want to conclude with some areas for improvement and further study.

Perhaps the biggest highlight of our research came in Section 5 when we discovered that the parameters of the distribution of preclinical durations could be predicted based on the mean lead time and pattern of diagnoses in a trial. To our knowledge, we are the first to address this association and attempt to develop a predictive model for preclinical parameters that can be used on real trial data. However, the prediction accuracy of our model was not as high as we would have expected. Multivariate least-squares linear regression fit three separate regression equations – one for each parameter to be estimated – such that each equation minimized the sum of squared residuals across our 12,500 simulated trials for the parameter in question. As a result, when using new data to predict preclinical duration distributions as we did in Section 5.3, each parameter was estimated independently. We know that, conditional on the mean lead time and pattern of diagnoses, there are dependencies, and perhaps strong ones, between the three preclinical duration parameters. For example, if the estimate for the proportion of fast-developing cases were to increase, we would expect the mean of the fast-developing cases would have to increase as well if the mean lead time and pattern of diagnoses were to be maintained. This suggests that a method for correlated responses would create a more accurate predictive model than multivariate least-squares linear regression. Moreover, we believe that the functional relationship between the preclinical duration distribution parameters Y and the mean lead time and pattern of diagnoses in a trial X should be examined more closely. Fitting a nonlinear function to X could also improve the prediction accuracy of the model. In all, we know that our discovery of a link between the distribution of preclinical durations and the mean lead time and pattern of diagnoses in a trial is a step in the right direction,

but further work is needed to improve on a model to predict these preclinical parameters from trial outcomes.

We also failed to thoroughly consider other mixture distributions besides the mixture of two exponentials when modelling the cancer growth periods. Discovering in Section 3 that a mixture of two exponential distributions fit the clinical durations from three historic screening trials well was an important result for our project, as we proceeded to assume that the same distribution type could be used to model the unobservable preclinical durations. The mixture of two exponential distributions approximated very well the clinical durations from the HIP trial, PLCO-Lung, and PLCO-Ovarian, but of course mixtures of other distributions would involve more parameters to be estimated and hence may yield better fits. In Section 6, we evaluated the sensitivity of the length-biased sampling effect $E[Y^*]/E[Y]$ to misspecification of the initial test sensitivity or preclinical duration distribution parameters, but we did not consider a misspecification of the form of this distribution. When comparing our estimates for $E[Y^*]/E[Y]$ to the results from similar scenarios in Kafadar and Prorok [29] and Heltshe et al [12], we found similar estimates despite these studies modelling the preclinical durations using the gamma and lognormal distributions, respectively. These parallels suggest that other attributes of the distribution of preclinical durations, such as the proportion of fast-developing cases or the subgroup means, are driving the length-biased sampling effect $E[Y^*]/E[Y]$ rather than the form of the distribution, which would mean that modelling the preclinical durations with a different mixture distribution would change very little in our results and conclusions. Nonetheless, other mixture distributions to model preclinical and clinical durations should be explored in future work.

Another limitation of our project is that we were not able to fully specify the joint distribution of preclinical and clinical durations. We developed an exploratory method in Section 3 to estimate the parameters of a mixture of two exponential distributions fit to observed clinical durations, and we created a predictive model using simulated trials in Section 5 to estimate the same parameters for preclinical durations based on the pattern of diagnoses in a trial. However, we did not attempt to estimate the correlation between preclinical and clinical durations necessary to complete the joint distribution. The correlation helps to determine which preclinical durations match with which clinical durations. Without specifying the correlation between preclinical and clinical durations in Section 6, we were unable to estimate the length-biased sampling effect on the clinical durations, $E[Z^*]/E[Z]$, which is particularly informative for the bias of mean benefit time estimates. The simulation studies from Kafadar and Prorok [29] and Heltshe et al [12] found that changes to the assumed correlation had a significant effect on $E[Z^*]/E[Z]$.

A dissertation committee member has suggested using a transition matrix to help specify the correlation between preclinical and clinical durations. Assuming that there are two states to both the preclinical and clinical phases – fast-developing (F) or slow-developing (S) – a transition matrix between the phases would look something like

$$P = \begin{bmatrix} P_{FF} & P_{FS} \\ P_{SF} & P_{SS} \end{bmatrix}. \tag{7.1}$$

Starting with those slow-developing cases in the preclinical phase, we would expect P_{SF}, the probability of transitioning from slow preclinical growth to fast clinical growth, to be very low and, consequently, P_{SS}, the probability of transitioning from slow preclinical

growth to slow clinical growth, to be very high; perhaps we may even set $P_{SF} = 0$ and $P_{SS} = 1$. For fast-developing cases in the preclinical phase, it makes sense that the clinical phase could be either short or long; a fast-developing cancer could continue to progress quickly after diagnosis, or its development could slow down due to an effective treatment. Thus, P_{FF} and P_{FS} would need to be selected based on the likelihood of effective treatment for a fast-progressing disease. Knowledge of the proportion of fast-developing cases in both the preclinical and clinical phases, estimates of which we discussed in Sections 5 and 3, respectively, could also assist in specifying P_{FF} and P_{FS} for the transition matrix.

Additionally, we considered avoiding specification of the correlation between preclinical and clinical duration entirely, instead estimating the functional relationship between the mean lead time and mean benefit time. Lead time is a function of only the screening program (including test sensitivity) and the preclinical duration, meaning that, once the effect of length-biased sampling on the screen-detected preclinical durations is estimated, a debiased estimate of mean lead time could be derived, perhaps by reweighting each observation. If lead time and benefit time were associated, then the debiased estimate of mean benefit time could potentially be predicted from this debiased mean lead time. Intuitively, this association makes sense, as a longer lead time means that the cancer was detected earlier at a less-advanced stage, so there is a greater likelihood of successful treatment and a long benefit time. To evaluate this idea, we would need to graphically compare lead and benefit times from real trial data. However, because an individual lead or benefit time is unobservable, we would need to assess the relationship between mean lead and benefit times. Moreover, because the lead and benefit times are also dependent on the cancer type and the screening program, we could not compare mean lead and

benefit times across trials. We considered using Bootstrap resampling on an individual screening trial's data to generate many pairs of estimated mean lead and benefit times, from which we could assess the relationship between these two measures. The flaw in this approach is that Bootstrap resampling does not ensure comparable case groups for estimation, as there is no guarantee that the matching cases from the treatment and control arm will both be selected; any relationship between mean lead and benefit time could be attributed to sampling bias rather than a true association between the variables. Regrettably, we were unable to devise another approach that would avoid bias in assessing the relationship between mean lead and benefit time.

Finally, while we were able to quantify the effects of length-biased sampling on the screen-detected preclinical durations using data-driven estimates for the distribution of preclinical durations, we largely ignored the second source of bias in screening trials, overdiagnosis. While some screening trials seem to present no cases of overdiagnosis, such as the HIP trial [10, 33] and National Lung Screening Trial (NLST) [48], overdiagnosis is an issue in almost all other trials, including the lung and ovarian cancer screening arms of the PLCO [19,41]. When a trial includes overdiagnoses, the catch-up point cannot be identified, so reliable estimates cannot be calculated for the mean lead time, mean benefit time, or even the pattern of diagnoses, all of which are limited by both the overdiagnoses themselves and the high number of post-screening detections in the treatment arm when the catch-up point cannot be determined. As a result, in Sections 5 and 6 of this report, we were unable to make predictions regarding the distribution of preclinical durations or the length-biased sampling effect for lung and ovarian cancers in the PLCO.

None of our simulation studies included overdiagnoses because we ensured that every screen-detected cancer had its unscreened match included in the control arm of our simulated trial. This ensured that cancer incidence would be the same between both study arms. However, we have an idea for how to generate overdiagnoses in a simulated screening trial. When simulating preclinical durations, we would classify any duration longer than some time point m to be an overdiagnosis. The challenge with our idea is the choice of m. It would make sense to choose a very large value for m so that, had the overdiagnosed case not been screen-detected, the preclinical duration would have outlasted the follow-up period of the trial, as a real-world overdiagnosed case would never complete its preclinical duration. Furthermore, m also could be selected so that the relative frequency of overdiagnoses matches with a comparison real trial, where the number of overdiagnoses is estimated by the difference in cancer incidence between the treatment and control arms. The overdiagnosed cases would be included in the simulated treatment arm but not the control arm. We could assess the effect of these overdiagnoses on mean lead and benefit time similarly to how the length-biased sampling effect is currently studied. Such simulated data could also be used to develop more thoughtful methods for identifying overdiagnosed cases in real trial data. Given the significance of the effect of overdiagnoses on trial outcomes and the prevalence of overdiagnoses in real screening trials, this is an area of study that merits great consideration.

References

1. Cancer Statistics. National Cancer Institute: National Institutes of Health. 2018. Available from: https://www.cancer.gov/about-cancer/understanding/statistics.

2. Cuomo MI. *A world without cancer: the making of a new cure and the real promise of prevention*. Rodale Books: Emmaus, PA, 2012.

3. Screening Tests. National Cancer Institute: National Institutes of Health. 2019. Available from: https://www.cancer.gov/about-cancer/screening/screening-tests.

4. Breast Cancer: Screening. U.S. Preventative Services Task Force. 2016. Available from: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening.

5. Colorectal Cancer: Screening. U.S. Preventative Services Task Force. 2016. Available from: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/colorectal-cancer-screening.

6. 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.

7. Kafadar K, Prorok PC. Computational methods in medical decision making: to screen or not to screen? *Statist. Med*. 2005; 24: 569-581.

8. Prorok PC, Connor RJ, Baker SG. Statistical considerations in cancer screening programs. *Urologic Clinics of North America.* 1990; 17: 699 –708.

9. Connor RJ, Prorok PC. Issues in the mortality analyses of randomized controlled trials of cancer screening. *Controlled Clinical Trials.* 1994; 15: 81–99.

10. Kafadar K, Prorok PC. Alternative definitions of comparable case groups and estimates of lead time and benefit time in randomized cancer screening trials. *Statist. Med*. 2003; 22: 83-111.

11. Prorok PC, Kramer BS, Miller AB. Study designs for determining and comparing sensitivities of disease screening tests. *J Med Screen*. 2015; 22(4): 213-220.

12. Heltshe SL, Kafadar K, Prorok PC. Quantification of length-bias in screening trials with covariate-dependent test sensitivity. *Biometrical Journal*. 2015; 0: 1-20.
13. Lafata JE, Simpkins J, Lamerato L, Poisson L, Divine G, Johnson CC. The economic impact of false-positive cancer screens. *Cancer Epidemiol Biomarkers Prev.* 2004; 13: 2126-2132.

14. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute*. 1982; 69(2): 349-355.

15. Shapiro S. Periodic screening for breast cancer: the HIP randomized controlled trial. *JCNI Monographs*. 1997; 1997(22): 27-30.

16. Gohagan JK, Prorok PC, Hayes RB, Kramer BS. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials*. 2000; 21(6S): 251S-272S.

17. Miller EA, Pinsky PF, Schoen RE, Prorok PC, Church TR. Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: long-term follow-up of the randomised US PLCO cancer screening trial. *Lancet Gastroenterol Hepatol*. 2019; 4(2): 101-110.

18. Pinsky PF, Prorok PC, Yu K, Kramer BS, Black A, Gohagan JK, Crawford ED, Grubb RL, Andriole GL. Extended mortality results for prostate cancer screening in the PLCO trial with median follow-up of 15 years. *Cancer*. 2017; 123(4): 592-599.

19. Oken MM, Hocking WG, Kvale PA, Andriole GL, Buys SS, Church TR, Crawford ED, Fouad MN, Isaacs C, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Rathmell JM, Riley TL, Wright P, Caparaso N, Hu P, Izmirlian G, Pinsky PF, Prorok PC, Kramer BS, Miller AB, Gohagan JK, Berg CD. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. *JAMA*. 2011; 306(17): 1865-1873.

20. Pinsky PF, Yu K, Kramer BS, Black A, Buys SS, Partridge E, Gohagan J, Berg CD, Prorok PC. Extended mortality results for ovarian cancer screening in the PLCO trial with median 15 years follow-up. *Gynecological Oncology*. 2016; 143(2): 270-275.

21. Kaminski MF, Bretthauer M, Zauber AG, Kuipers EJ, Adami HO, van Ballegooijen M, Regula J, van Leerdam M, Stefansson T, Påhlman L, Dekker E, Hernán MA, Garborg K, Hoff G. The NordICC Study: rationale and design of a randomized trial on colonoscopy screening for colorectal cancer. *Endoscopy*. 2012; 44(7): 695-702.

22. Cologuard. 2017. Trademark of Exact Sciences Corporation, Madison, WI. https://www.cologuardtest.com/

23. ClinicalTrials.gov. Bethesda, MD: National Library of Medicine (US). 2014. Identifier NCT02078804, Colonoscopy and FIT as Colorectal Cancer Screening Test in the Average Risk Population. Available from: https://clinicaltrials.gov/ct2/show/NCT02078804.

24. Quintero E, Castells A, Bujanda L, Cubiella J, Salas D, Lanas A, Andreu M, Carballo F, Morillas JD, Hernandez C, Jover R, Montalvo I. Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. *N Engl J Med*. 2012; 366: 697-706.

25. ClinicalTrials.gov. Bethesda, MD: National Library of Medicine (US). 2009. Identifier NCT00906997, Colorectal Cancer Screening in Average-risk Population: Immunochemical Fecal Occult Blood Testing Versus Colonoscopy. Available from: https://clinicaltrials.gov/ct2/show/NCT00906997.

26. Dominitz JA, Robertson DJ, Ahnen DJ, Allison JE, Antonelli M, Boardman KD, Ciarleglio M, Del Curto BJ, Huang GD, Imperiale TF, Larson MF, Lieberman D, O'Connor T, O'Leary TJ, Peduzzi P, Provenzale D, Shaukat A, Sultan S, Voorhees A, Wallace R, Guarino PD. Colonoscopy vs. Fecal Immunochemical Test in Reducing Mortality From Colorectal Cancer (CONFIRM): rationale for study design. *Am J Gastroenterol*. 2017; 112(11): 1736-1746.

27. ClinicalTrials.gov. Bethesda, MD: National Library of Medicine (US). 2010. Identifier NCT01239082, Colonoscopy Versus Fecal Immunochemical Test in Reducing Mortality From Colorectal Cancer (CONFIRM). Available from: https://clinicaltrials.gov/ct2/show/NCT01239082.

28. Blumenthal S. Proportional sampling in life length studies. *Technometrics.* 1967; 9: 205-218.

29. Kafadar K, Prorok PC. Effect of length biased sampling of unobserved sojourn times on the survival distribution when disease is screen detected. *Statist. Med*. 2009; 28: 2116-2146.

30. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer*. 2013; 108(11): 2205-2240.

31. Kafadar K, Prorok PC. A data-analytic approach for estimating lead time and screening benefit based on survival curves in randomized cancer screening trials. *Statist. Med.* 1994; 13: 569-586.

32. 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.

33. Aron JL, Prorok PC. An analysis of the mortality effect in a breast cancer screening study. *International Journal of Epidemiology.* 1986; 15: 36-43.

34. Etzioni R, Self SD. On the catch-up time method for analyzing cancer screening trials. *Biometrics*. 1995; 51: 31–43.

35. Gulati R, Feuer EJ, Etzioni R. Conditions for valid empirical estimates of cancer overdiagnosis in randomized trials and population studies. *Am J Epidemiol*. 2016; 184(2): 140-147.

36. Day NE, Walter SD. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. *Biometrics*. 1984; 40(1): 1-13.

37. Day NE. Estimating the sensitivity of a screening test. *Journal of Epidemiology and Community Health*. 1985; 39: 364-366.

38. Duffy SW, Chen H, Tabar L, Day NE. Estimation of mean sojourn time in breast cancer screening using Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine*. 1995; 14: 1531-1543.

39. Straatman H, Peer PGM, Verbeek ALM. Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. *Biometrics*. 1997; 53(1): 217-229.

40. Hakama M, Auvinen A, Day NE, Miller AB. Sensitivity in cancer screening. *J Med Screen*. 2007; 14: 174-177.

41. Buys SS, Partridge E, Black A. Effect of screening on ovarian cancer mortality. *JAMA*. 2011; 305(22): 2295-2303.

42. Kafadar K, Prorok PC. Estimating average durations of survival time components in screen-detected cases accounting for overdiagnosis. In progress.

43. McLachlan G, Peel D. Application of EM algorithm for mixture models. In *Finite Mixture Models*. Cressie N, Fisher N, Johnstone I, Kadane JB, Scott D, Silverman B, Smith A, Teugels J, Barnett V, Bradley R, Hunter JS, Kendall G (eds). John Wiley & Sons, Inc., 2000: 47-50.

44. Seidel W, Mosler K, Alker M. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*. 2000; 52(3): 481-487.

45. Verbelen R, Gong L, Antonio K, Badescu A. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *Journal of the International Actuarial Association*. 2015; 45(3): 729-758.

46. Kafadar K, Prorok PC. Computer simulation of randomized cancer screening trials to compare methods of estimating lead time and benefit time. *Computational Statistics & Data Analysis*. 1996; 23: 263-291.

47. Johnson RA, Wichern DW. Multivariate linear regression models. In *Appliced Multivariate Statistical Analysis*. Pearson Education, Inc., 2007: 360-429.

48. National Lung Screening Trial Research Team. Lung cancer incidence and mortality with extended follow-up in the National Lung Screening Trial. *Journal of Thoracic Oncology*. 2019; 14(10): 1732-1742.