8-2017

Estimation of the three key parameters and the lead time distribution in lung cancer screening.

Ruiqi Liu
University of Louisville

ESTIMATION OF THE THREE KEY PARAMETERS AND THE LEAD TIME DISTRIBUTION IN LUNG CANCER SCREENING

By

Ruiqi Liu
B.S., China Jiliang University, 2010
M.S., University of Louisville, 2013

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2017

ESTIMATION OF THE THREE KEY PARAMETERS AND THE LEAD TIME DISTRIBUTION IN LUNG CANCER SCREENING

By

Ruiqi Liu
B.S., China Jiliang University, 2010
M.S., University of Louisville, 2013

A Dissertation Approved on

July 25, 2017

by the following Dissertation Committee:

Dongfeng Wu, Ph.D., Dissertation Director

Shesh Rai, Ph.D.

Goetz Kloecker, M.D.

Qi Zheng, Ph.D.

Jeremy Gaskins, Ph.D.

Ritendranath Mitra, Ph.D.

DEDICATION

This dissertation is dedicated to my parents who have given me invaluable educational opportunities. Their affection, love and encouragement make me able to complete this work.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Dongfeng Wu, for her excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I would like to thank all the other committee members, Dr. Shesh Rai, Dr. Goetz Kloecker, Dr. Jeremy Gaskins, Dr. Qi Zheng and Dr. Ritendranath Mitra, for their assistance and thoughtful comments regarding my research.

I would also like to thank the Department of Bioinformatics and Biostatistics and the Department of Pediatrics at University of Louisville for providing the financial support of my Ph.D study.

ABSTRACT

ESTIMATION OF THE THREE KEY PARAMETERS AND THE LEAD TIME DISTRIBUTION IN LUNG CANCER SCREENING

Ruiqi Liu

July 25, 2017

This dissertation contains three research projects on cancer screening probability modeling. Cancer screening is the primary technique for early detection. The goal of screening is to catch the disease early before clinical symptoms appear. In these projects, the three key parameters and lead time distribution were estimated to provide a statistical point of view on the effectiveness of cancer screening programs.

In the first project, cancer screening probability model was used to analyze the computed tomography (CT) scan group in the National Lung Screening Trial (NLST) data. Three key parameters were estimated using Bayesian approach and Markov Chain Monte Carlo simulations. The NLST CT arm data have been used for the estimation. The sensitivity for lung cancer screening using CT scan is much higher than those screening using X-ray. The transition probability from disease-free to preclinical state has a peak around age 70 for both genders. The posterior mean sojourn time is around 1.5 years for all groups.

The second project is dealing with lead time distribution estimation. Since the lead time is unobservable, the effectiveness of screening exams regarding the survival benefits becomes a major concern. In this study, the estimates for the projected lead time was presented by using the NLST CT arm data. Simulation results show

that the probability of no-early-detection increases monotonically when the screening interval increases for both genders. The mean lead time appears longer for women than for men.

In previous study, it was assumed that a person has no screening history before entering the study. However, the participants of the screening programs are usually aged population and they may already have at least one prior screening exam in the past and look healthy. In the third project, we extended the previously developed lead time distribution to consider an individual's screening history and to see how much this history will affect the lead time. We did simulation for each combination of initial screening ages, sensitivities, mean sojourn times, current ages and screening schedules in the past and in the future. We also applied the newly developed lead time distribution to the NLST data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This dissertation consists of three interconnected research projects (Chapters 2-4) on cancer screening probability modeling. Three key parameters in cancer screening are estimated using Bayesian approach in the first project (Chapter 2). In the second project (Chapter 3), the lead time distribution is estimated based on the results obtained in first project. The last project deals with developing lead time distribution for individuals with screening history.

This chapter is a review of methods used to estimate the three key parameters and the lead time distribution, which follows Liu et al. (2017).

1.1 Cancer Screening Overview

Lung cancer is the most life-threatening cancer for both men and women in the U.S. and worldwide. The cause of lung cancer remains unknown. However, about 80% of lung cancer deaths are caused by smoking and secondhand smoke exposure. It is clear that tobacco smoking is one of the strongest risk factors for lung cancer. The advanced lung cancer is often hard to treat, the 5-year survival rate for patients with early stage lung cancer is around 50%, but it is less than 5% for patients with stage IV lung cancer (NCI, 2015). The disease may not show any signs or symptoms during the early stage lung cancer, thus it is very possible that the disease unknowingly moves to the late stage without any intervention. The fact is, nearly 70% of lung cancers are diagnosed at advanced stages, and the general prognosis of lung cancer is poor

(Molina et al., 2008).

Cancer screening, as the primary technique for early detection, has been carried out since 1960s. The goal of screening is to catch the disease early before clinical symptoms appear. The United States Preventive Services Task Force (USPSTF) has recommended screening schedules for almost all of the most prevalent cancers (USPSTF, 2016), such as breast, lung, colon, cervical cancer, etc. Although different cancer sites have their specific characteristics and developmental stages, they all share some common features as well.

I will first outline the commonly followed progressive model used in cancer screening and its parameters. A cohort of apparently healthy individuals are enrolled in a screening program to detect the presence of a specific disease. The disease progressive stochastic model was first proposed by Zelen and Feinleib (1969) and has been used since then. In this model, the disease develops by progressing through 3 states: $S_0 \rightarrow S_p \rightarrow S_c$ (See Figure 1.1). S_0 refers to the disease-free state or the state in which the disease can not be detected; S_p refers to the preclinical disease state, in which an asymptomatic individual unknowingly has the disease that a screening exam can detect; and S_c refers to the disease state at which the disease manifests itself in clinical symptoms. The progressive disease model describes the natural history of lesions detected by screening for cancer. The goal of screening programs is to detect the cancer in the preclinical state (S_p), so that it may be treated before adverse symptoms arise.

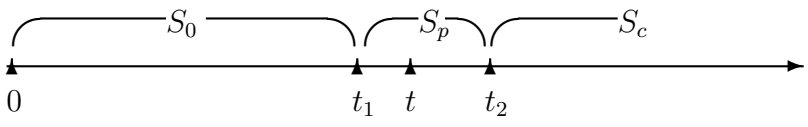

Figure 1.1: Disease progressive states and the lead time

Sensitivity is the probability that an screening exam result is positive, given that an individual is in the preclinical state S_p. More specifically, a binary variable

D represents the true disease status of an individual, that is, D takes value one when an individual has the disease, and zero otherwise. The binary variable X represents test result from a screening exam, with $X = 1$ indicating that the test is positive. The sensitivity is the probability of correctly identifying those who have the disease, that is, $\beta = P(X = 1|D = 1)$. *Specificity* is the probability of correctly identifying those who do *not* have the disease, that is, $\alpha = P(X = 0|D = 0)$. Ideally, we desire the test to have both a sensitivity and specificity of 100%, but in reality this is unachievable. In fact, both sensitivity and specificity cannot be estimated directly from data summary in a mass screening. To see why, suppose there are n people taking part in one screening exam, according to their true disease status and the screening results, they can be classified into four categories as in Table 1.1.

Table 1.1: True disease status and test result in one mass screening

		Disease Status	
		Diseased: $D = 1$	*Not diseased:* $D = 0$
Test Result	+	True positive (n_{11})	False positive (n_{12})
	−	False negative (n_{21})	True negative (n_{22})

From Table 1.1, the sensitivity $\beta = n_{11}/(n_{11} + n_{21})$, and the specificity $\alpha = n_{22}/(n_{12} + n_{22})$, where n_{11} and n_{12} can be obtained by a follow-up exam, such as a biopsy after a positive screening result to confirm either the finding is cancerous or not. However, for those screened negative individuals (who are the majority in a mass screening), confirmation of the true disease status is not cost effective, nor ethical. Therefore n_{21} and n_{22} are usually unknown, hence the β and the α cannot be obtained from data directly. Also, a screened negative individual who has been followed and found to be positive later may involve two cases: either it is a false negative on the previous screening exams, or it is a newly developed case. However, sensitivity can be estimated by the likelihood method and collected mass screening data (Shen and Zelen, 1999; Wu et al., 2005a,b).

Sojourn time is the time from when the disease first develops to the manifestation of clinical symptoms. If one enters the preclinical state (S_p) at age t_1, and becomes clinically incident (S_c) later at age t_2, then $(t_2 - t_1)$ is the *sojourn time*, see Figure 1.1. The nature of data collection in a screening program make the exact observation of time of onset of either S_p or S_c impossible. Therefore, estimation of the sojourn time distribution is difficult. However, this information can be obtained under model assumptions. For example, it is believed that the preclinical state of breast cancer may last from 1 to 4 years (Shen and Zelen, 1999; Shen et al., 2001; Wu et al., 2005a,b), and it may last longer for colorectal cancer (Wu et al., 2009b). Hence, there is a good chance that cancer could be detected in its preclinical stage, which is the goal of implementing a screening program.

The transition density from the disease free state (S_0) to the preclinical state (S_p) is the probability density function (PDF) of the time duration in the disease-free state S_0, i.e., t_1 in Figure 1.1. It is commonly assumed that the sojourn time and the transition time are independent (Wu et al., 2005a,b). Due to the imperfect sensitivity of the test and the interval-censored nature of the data, the transition density is typically estimated by relying on common parametric models or interval-constant assumptions.

Lead time is the length of time that the diagnosis is advanced by screening. In Figure 1.1, if one is offered a screening exam at time t within the time interval (t_1, t_2), and cancer is diagnosed, then the length of the time $(t_2 - t)$ is the *lead time*. An individual with a longer lead time usually has a better prognosis than one with a shorter lead time. For a particular case detected by the screening, the lead time is unobservable, due to the fact that once cancer was diagnosed, it will be treated immediately, making it impossible to observe the onset of clinical state S_c.

The three key parameters in screening are the sensitivity, the sojourn time and the transition density. They are the key parameters due to the fact that all

other estimates are functions of these three key parameters, including the lead time. We will briefly review the existing statistical methods used to estimate the three key parameters in cancer screening, as well as the methods for estimating the lead time.

We first introduce some notation used in the remainder of the dissertation. Consider a group of initially asymptomatic individuals scheduled with K ordered screening exams $t_0 < t_1 < \ldots < t_{K-1}$, where t_{i-1} represents a person's age when receiving the ith screen, $i = 1, \ldots, K$. For an annual screening program, $t_i = t_0 + i$. We define β as the sensitivity of the screening exam, $\beta = \beta(t_i)$ if it is age-dependent. The function $w(t)$ describes the time duration in S_0; note that it is a sub-PDF due to the fact that someone may stay in the state S_0 during their lifetime. Finally, $q(\cdot)$ is the probability density function of the sojourn time in S_p, with the survival function $Q(z) = \int_z^\infty q(x)\,dx$.

The mass screening data used in these methods usually consist of three parts from each screening cycle: n_i is the total number of individuals examined at ith screen (at age t_{i-1}); s_i denotes the number of individuals diagnosed by the ith screening exam, that is, the number of screen-detected cases; r_i is the number of individuals found in the clinical state (S_c) within the ith screening interval (t_{i-1}, t_i), that is, the number of interval cases. Table 1.2 shows the data format for a mass screening program with K scheduled exams, where t_0 is the age at the first exam, and the triplets (n_i, s_i, r_i) stratified by the initial age are the data we use.

Table 1.2: A sample of mass cancer screening data

Age (t_0)	n_1	s_1	r_1	n_2	s_2	r_2	...	n_K	s_K	r_K
⋮										
60	1946	16	3	1847	13	1	...	1797	17	0
61	1786	18	0	1678	14	1	...	1659	11	3
62	1548	11	1	1452	8	2	...	1408	12	0
⋮										

1.2 Estimation of the Three Key Parameters

1.2.1 Likelihood function in stable and nonstable disease models

Shen and Zelen (1999) proposed a likelihood function to estimate the screening sensitivity and the mean sojourn time under the assumptions of a stable and nonstable disease model. The stable model means that the transition density $w(t) = w$, is uniformly distributed over all ages, and the nonstable model allows the probability of transitioning $w(t)$ to depend on t. In their approach, they take $w(t)$ to be a step function of age with discontinuities every five years. The sojourn time was assumed to follow an exponential(μ) distribution in both stable and nonstable models, i.e., $Q(x) = \exp(-x/\mu)$. The estimated parameters are the sensitivity β, the mean sojourn time μ and the transition density w.

Consider the ith screening interval $[t_{i-1}, t_i)$ of a fixed age strata. Let D_i be the probability of an preclinical individual diagnosed at the ith screen given at age t_{i-1}. It can be calculated by

$$D_i = \begin{cases} \beta w \mu \left[1 - \beta \sum_{j=1}^{i-1} (1-\beta)^{i-j-1} Q(t_{i-1} - t_{j-1})\right] & (i > 1), \\ \beta w \mu & (i = 1). \end{cases} \tag{1.1}$$

Let I_i be the probability of an individual being incident in the ith interval, it is given by

$$I_i = w\mu \left[\frac{t_i - t_{i-1}}{\mu} - \beta \sum_{j=0}^{i-1} (1-\beta)^{i-j-1} \{Q(t_{i-1} - t_j) - Q(t_i - t_j)\} \right]. \tag{1.2}$$

Thus, the full likelihood function was derived as

$$L_i = {D_i}^{s_i} {I_i}^{r_i} \{1 - D_i - I_i\}^{n_i - s_i - r_i} \prod_{j=1}^{3} \left(\frac{\alpha_j}{\beta} \right)^{s_{ij}}, \tag{1.3}$$

where the likelihood functions only depend on sensitivities for different modalities α_j and the parameter vector of the sojourn time distribution. The overall sensitivity, $\beta = \alpha_1 + \alpha_2 + \alpha_3$, is applied to the case of using two screening modalities simultaneous in each exam, such as using mammogram and physical exam in breast cancer, or using chest X-ray and sputum cytology in lung cancer, with $\beta_1 = \alpha_1 + \alpha_3$ and $\beta_2 = \alpha_2 + \alpha_3$ represent sensitivity of each modality (See Shen et al. (2001) for details). And $s_{i1} + s_{i2} + s_{i3} = s_i$ denotes the number of cases detected by modality 1 only, by modality 2 only and by both.

By treating r_i and s_i as approximately Poisson, they develop a simplified conditional likelihood function

$$L_i = \frac{{I_i}^{r_i} {D_i}^{s_i}}{\{I_i + D_i\}^{(r_i+s_i)}} \prod_{j=1}^{3} \left(\frac{\alpha_j}{\beta}\right)^{s_{ij}}. \tag{1.4}$$

In both papers (Shen and Zelen, 1999; Shen et al., 2001), the data was not stratified by age, which means, Table 1.2 could be collapsed into a vector. Two breast cancer screening datasets: the Health Insurance Plan (HIP) study and Canadian National Breast Screening studies were used in both the stable and non-stable model. In the non-stable model, estimates of the transition constant w in every five years can be achieved by using the incidence data from the SEERs database. The innovation of this study is that a likelihood function was developed to estimate the sensitivity and the mean sojourn time.

1.2.2 Estimation of age-dependent sensitivity and transition probability

Wu et al. (2005a) developed statistical inference procedures to estimate the sojourn time, the age-dependent sensitivity, and the age-dependent transition density from the disease-free state to the preclinical state. Both maximum likelihood estimate (MLE) and Bayesian posterior estimates were used to estimate the parameters. The

age was considered to be a covariate of the sensitivity and the transition probability density.

Consider a cohort of initially asymptomatic individuals who enter the screening program at age t_0. There are K ordered screening exams that will occur at age $t_0 < t_1 \cdots < t_{K-1}$. $T = t_K$ is the follow-up time after the last exam, during which an incident case may be detected. Let $(n_{i,t_0},\ s_{i,t_0},\ r_{i,t_0})$ be the data for the ith interval as defined for the strata with starting age t_0. Then the likelihood for the individuals aged t_0 at study entry is proportional to

$$L(\cdot|t_0) = \prod_{k=1}^{K} D_{k,t_0}^{s_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1 - D_{k,t_0} - I_{k,t_0})^{n_{k,t_0} - s_{k,t_0} - r_{k,t_0}}, \tag{1.5}$$

where D_{k,t_0} is the probability that an individual will be detected by the kth screening exam (at age t_{k-1}) given this person is in the state S_p. When $k = 1, 2, \ldots, K$, D_{k,t_0} can be calculated by

$$D_{1,t_0} = \beta(t_0) \int_0^{t_0} w(x) Q(t_0 - x)\, dx, \tag{1.6}$$

$$\begin{aligned} D_{k,t_0} =& \beta(t_{k-1}) \left\{ \sum_{i=1}^{k-2} \left\{ [1 - \beta(t_i)] \cdots [1 - \beta(t_{k-2})] \int_{t_{i-1}}^{t_i} w(x) Q(t_{k-1} - x)\, dx \right\} \right. \\ &\left. + \int_{t_{k-2}}^{t_{k-1}} w(x) Q(t_{k-1} - x)\, dx \right\}, \text{ for } k = 2, \ldots, K. \end{aligned} \tag{1.7}$$

The likelihood also depends on I_{k,t_0}, the probability of an individual being incident during the kth interval (t_{k-1}, t_k), it can be calculated by

$$\begin{aligned} I_{k,t_0} =& \sum_{i=0}^{k-1} \left\{ [1 - \beta(t_i)] \cdots [1 - \beta(t_{k-1})] \int_{t_{i-1}}^{t_i} w(x) [Q(t_{k-1} - x) - Q(t_k - x)]\, dx \right\} \\ &+ \int_{t_{k-1}}^{t_k} w(x) [1 - Q(t_k - x)]\, dx, \text{ for } k = 1, \ldots, K. \end{aligned} \tag{1.8}$$

For one screening study, the likelihood for all initial age groups is proportional to

$$L = \prod_{t_0} L(\cdot|t_0). \tag{1.9}$$

We can clearly see the likelihood is a function of the three key parameters $\beta(t)$, $w(t)$ and $q(x)$. The parametric models for the three key parameters were carefully chosen as following:

$$\beta(t) = \frac{1}{1+\exp\{-b_0 - b_1(t-\bar{t})\}}, \tag{1.10}$$

$$w(t) = w_{max} \cdot \frac{1}{\sqrt{2\pi}\sigma t} \exp\{-(\log t - \mu)^2/(2\sigma^2)\}, \quad t > 0, \tag{1.11}$$

$$q(x) = \frac{\kappa x^{\kappa-1}\rho^{\kappa}}{(1+(x\rho)^{\kappa})^2}, \tag{1.12}$$

where $\bar{t}$ is the average age at entry in the study group. The sensitivity $\beta(t)$ was associated with age t by a logistic link. The log-normal distribution was used for the transition probability $w(t)$. As the integral of $w(t)$ over all ages is the lifetime risk of developing a cancer and should always be less than 1, $w(t)$ is in fact a sub-PDF. Hence, the upper limit was set to $w_{max} = \int w(t)dt$. For breast cancer, the upper limit was set to be 0.2 (Wu et al., 2005a). For the sojourn time, the log-logistic distribution was adopted, in part due to its convenient survival function $Q(x) = [1+(\rho x)^{\kappa}]^{-1}$. The unknown parameters $\theta = (b_0, b_1, \mu, \sigma^2, \kappa, \rho)$ were estimated from the likelihood function described above. Simulations were carried to evaluate the reliability of the proposed likelihood, and the detailed procedure can be found in Wu et al. (2005b). Both Markov Chain Monte Carlo (MCMC) estimates and MLEs were obtained. They applied their model to the HIP female breast cancer study and obtained estimates for age-dependent sensitivity and transition probability along with the sojourn time.

1.2.3 Key parameters estimation when sensitivity depends on sojourn time

Wu et al. (2008) argued that the screening sensitivity should be a function of age at diagnosis and the amount of time spent in the preclinical state, rather than only depend on the age at diagnosis. Intuitively, as the cancer gets closer to progression from the preclinical state to the clinical state, it should be easier to catch by a screening exam than it was previously.

In this way, the sensitivity is modeled as $\beta = \beta(t, s|S)$, where t represents an individual's age at the screening exam, s is the time duration a person has already spent in the preclinical state, and S is the sojourn time in S_p ($s < S$). The probability that an individual will be diagnosed by the kth screening exam (at age t_{k-1}) given that this person is in the state S_p with initial age t_0 becomes

$$D_{1,t_0} = \int_0^{t_0} w(x) \int_{t_0-x}^{\infty} q(t)\beta(t_0, t_0 - x|t)\, dt\, dx, \tag{1.13}$$

$$\begin{aligned} D_{k,t_0} = & \\ & \sum_{i=0}^{k-2} \left\{ \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{\infty} q(t) \left(\prod_{j=i}^{k-2} [1 - \beta(t_j, t_j - x|t)] \right) \beta(t_{k-1}, t_{k-1} - x|t)\, dt\, dx \right\} \\ & + \int_{t_{k-2}}^{t_{k-1}} w(x) \int_{t_{k-1}-x}^{\infty} q(t)\beta(t_{k-1}, t_{k-1} - x|t)\, dt\, dx, \text{ for } k = 2, \ldots, K. \end{aligned} \tag{1.14}$$

The probability that an individual is an incident case during the kth interval (t_{k-1}, t_k) with initial age t_0 becomes

$$\begin{aligned} I_{k,t_0} = & \sum_{i=0}^{k-1} \left\{ \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{t_k-x} q(t) \left(\prod_{j=i}^{k-1} [1 - \beta(t_j, t_j - x|t)] \right) dt\, dx \right\} \\ & + \int_{t_{k-1}}^{t_k} w(x)[1 - Q(t_k - x)]\, dx, \text{ for } k = 2, \ldots, K. \end{aligned} \tag{1.15}$$

The sensitivity associated with age, time spent in S_p and sojourn time is

$$\beta(t, s|S) = \frac{1}{1 + exp[-b_0 - b_1(t - \bar{t})]} \times \frac{s}{S}, \tag{1.16}$$

where $\bar{t}$ is the average age at entry for the entire study group, S is the sojourn time, and s is the time a person already spent in preclinical state S_p, $s \in [0, S]$. Clearly, the sensitivity is increasing in s where the maximum sensitivity is achieved at $s = S$, that is, the moment the cancer transitions from preclinical to clinical. When $b_1 > 0$, the sensitivity is a monotonic increasing function of age t. This method was applied to breast cancer data, such as HIP (Wu et al., 2008).

Motivated by the fact that age seems to have little effect on the screening sensitivity in lung cancer, Kim and Wu (2016) treated the sensitivity as a function of time spent in the preclinical state and the sojourn time for further inference. The sensitivity was modeled as a ratio of time spent in the preclinical state s to the sojourn time S, given by

$$\beta(s|S) = \frac{1}{1 + \tau} \left(\frac{s}{S}\right)^{\gamma}, \quad \tau, \gamma \geq 0, \tag{1.17}$$

where τ is a parameter added to control the overall sensitivity. The parameter γ reflects the changing rate of sensitivity: when s/S is close to zero, the sensitivity increases rapidly if $\gamma < 1$, while it increases slowly if $\gamma > 1$.

The probabilities D_{k,t_0} and I_{k,t_0} are the same with Equations 1.13, 1.14 and 1.15. This method combined with the likelihood in Equation 1.5 was applied to the Johns Hopkins Lung Project data in Kim and Wu (2016).

1.3 Estimation of the lead time distribution

Lead time is the length of time that the diagnosis is advanced by screening. It can serve as a surrogate measurement on how effective a screening program is. In the case of cancer, survival time is measured from the time of diagnosis. Hence, an

earlier detection of the tumor due to screening will cause the patient's survival to appear long, even if there is no real effect on mortality. When survival benefit is compared between the study and the control group, the lead time must be adjusted by the study group, so accurate estimation of the lead time is necessary.

Many researchers have proposed methods to estimate the lead time (Kafadar and Prorok, 1994, 1996, 2003; Straatman et al., 1997). Most of these methods assumed that the sojourn time follows an exponential distribution, and due to the memoryless nature of the exponential random variable, the lead time will follow the same exponential distribution as well. These publications have provided estimates of the mean and variance of the lead time under the exponential assumption. We will focus on three major methods in this section.

1.3.1 Local lead time distribution for the screen-detected cases

Prorok (1982) made a major contribution by deriving the conditional probability distribution of the lead time, given that one was detected at the i-th screening exam. Consider a screening program with a total of K screening exams. If an individual enters the preclinical state S_p during the time interval $(t_{i-1}, t_i]$, $i = 0, 1, \ldots, K-1$, this person is a member of the ith generation, where $t_{-1} = 0$. Prorok (1982) argued that the lead time distribution at a given screen, say $(j+1)$th screen, is a weighted average of the lead time distributions for all generations potentially detectable at it. The local lead time PDF for individuals detected in S_p by the $(j+1)$th (at time t_j) can be defined by

$$f_{Dj}(l) = \frac{\sum_{i=0}^{j} D_{ij} f_{ij}(l)}{\sum_{i=0}^{j} D_{ij}}, \quad l \geq 0, \quad j = 0, 1, \ldots, K-1, \tag{1.18}$$

where $f_{ij}(l)$ is the lead time distribution for ith generation who are detected at $(j+1)$th screen but not before. This distribution can be interpreted as a weighted-average

of the lead time distributions for each generation i, with mixing weights D_{ij}. The ith generation lead time distribution can be calculated by

$$f_{ij}(l) = \frac{\int_0^{t_i - t_{i-1}} w_i(t_i - u) Q_i(l + u + t_j - t_i)\, du}{\int_0^{t_i - t_{i-1}} w_i(t_i - u) Q_i(u + t_j - t_i)\, du}, l \geq 0, i = 0, 1, \ldots, K-1, j \geq i, \tag{1.19}$$

where $w_i(\cdot)$ and $Q_i(\cdot)$ are the transition density from S_0 to S_p and survival function of sojourn time for the ith generation, respectively. The u represents the time length that a person stays in S_p till t_i, a random variable.

The weighting factor D_{ij} is the probability that an individual is detected at $(j+1)$th screen given the person belongs to the ith generation. It can be obtained by

$$D_{ij} = P(E_i) P(t_i) Q_{vi}(t_j - t_i) f(\beta_{ij}), \quad j \geq i, \tag{1.20}$$

where $P(E_i)$ is the probability that an individual belongs to the ith generation. $P(t_i)$ is the probability that an ith generation individual is in S_p at time t_i. $Q_{vi}(t_j - t_i)$ is the probability that the time length of $(\tau - t_i)$ for an ith-generation individual is not less than $t_j - t_i$, where τ represents the time point this individual enters S_c. The term $f(\beta_{ij})$ takes account of the sensitivities of screens. The derivation of these probabilities can be found in Prorok (1982) and Prorok (1976).

Simulations were conducted to explore the lead time properties based on the derived lead time distribution. In the simulation, the sojourn time is assumed to follow the generalized gamma distribution with the same mean at 2 years and three different variances, corresponding to the cases of the coefficient of variation to be larger, smaller and equal to one. Simulation results showed that the local lead time for the ith screen-detected cases will not change after a certain number (four or five) of screens, given the screening interval was fixed at 1 year. This suggested a possible stopping rule when designing the screening programs, since it tended to not yield any additional benefit in continued screenings. However, this study only focuses on the

analysis of screen-detected cases whose lead time is positive, and ignored the interval cases whose lead time is zero.

1.3.2 Global lead time distribution when lifetime is fixed

Wu et al. (2007) rigorously evaluated the lead time distribution based on model parameters for the whole cohort participating in the screening program, including both the screen-detected and the interval incident cases. In this way, the proportion of patients whose lead time is zero can be estimated, together with the distribution of time of those patients who were detected early by screening. Thus, the lead time distribution is a mixture of a point mass at zero and a probability density function of a positive continuous random variable.

Let us consider an initially asymptomatic individual with no history of cancer, he or she is assumed to take K screening exams at ages $t_0 < t_1 < \cdots < t_{K-1}$, and T represents the lifetime, a fixed value. Let D represent true disease status, with $D = 1$ indicating having cancer and $D = 0$ indicating no clinical disease in one's lifetime. Let L represent the lead time of an individual. The distribution of lead time is a mixture of the conditional probability $P(L = 0|D = 1)$ and the conditional density function $f_L(z|D = 1)$, for $z \in (0, T - t_0)$

$$P(L = 0|D = 1) = \frac{P(L = 0, D = 1)}{P(D = 1)}, \tag{1.21}$$

$$f_L(z|D = 1) = \frac{f_L(z, D = 1)}{P(D = 1)}, \tag{1.22}$$

where $P(D = 1)$ is the probability of developing cancer after age t_0, and

$$P(D = 1) = \int_0^{t_0} w(x)[Q(t_0 - x) - Q(T - x)]\, dx + \int_{t_0}^{T} w(x)[1 - Q(T - x)]\, dx. \tag{1.23}$$

$P(L = 0, D = 1)$ is the probability that the lead time is zero, i.e., the collective

probability of being an interval case,

$$
\begin{aligned}
&P(L=0, D=1) \\
=&\sum_{j=1}^{K}\left\{\sum_{i=0}^{j-1}(1-\beta(t_i))\cdots(1-\beta(t_{j-1}))\int_{t_{i-1}}^{t_i} w(x)[Q(t_{j-1}-x)-Q(t_j-x)]\,dx\right. \\
&\left.+\int_{t_{j-1}}^{t_j} w(x)[1-Q(t_j-x)]\,dx\right\}.
\end{aligned} \tag{1.24}
$$

The joint probability density function $f_L(z, D=1)$ when $z \in (0, T-t_0)$ is

$$
f_L(z, D=1) = \beta(t_0)\int_0^{t_0} w(x)q(t_0+z-x)\,dx, \text{ if } T-t_1 < z \le T-t_0, \tag{1.25}
$$

$$
\begin{aligned}
&f_L(z, D=1) = \\
&\quad\sum_{i=1}^{j-1}\beta(t_i)\left\{\sum_{r=0}^{i-1}(1-\beta(t_r))\cdots(1-\beta(t_{i-1}))\int_{t_{r-1}}^{t_r} w(x)q(t_i+z-x)\,dx\right. \\
&\qquad\left.+\int_{t_{i-1}}^{t_i} w(x)q(t_i+z-x)\,dx\right\}+\beta(t_0)\int_0^{t_0} w(x)q(t_0+z-x)\,dx, \\
&\qquad\qquad \text{if } T-t_j < z \le T-t_{j-1}, \text{ for } j=2,3,\ldots,K.
\end{aligned} \tag{1.26}
$$

The validity of the probability calculation can be proved by

$$
P(L=0|D=1)+\int_0^{T-t_0} f_L(z|D=1)dz = 1. \tag{1.27}
$$

It is clear that the lead time distribution depends on the three key parameters: the sensitivity $\beta(\cdot)$, the transition probability $w(\cdot)$ and the distribution of sojourn time $q(\cdot)$. The method was applied to the HIP study and the posterior predictive distribution of the lead time was estimated using MCMC posterior samples. Bayesian inference was made to explore the lead time properties with different screening intervals (6, 9, 12, 18 and 24 months), given the initial screening age t_0=50 and lifetime T=80. Later, this method was applied to various cancer screening studies, including breast, lung, and colon cancer (Wu et al., 2007, 2011, 2009a).

1.3.3 Global lead time distribution when lifetime is a random variable

Wu et al. (2012) extended the lead time distribution by allowing the lifetime T to be a random variable, which is more realistic. The lead time distribution when T is a random variable can be obtained by

$$P(L = 0|D = 1, T \geq t_0) = \int_{t_0}^{\infty} P(L = 0|D = 1, T = t) f_T(t|T \geq t_0)\, dt, \tag{1.28}$$

$$f_L(z|D = 1, T \geq t_0) = \int_{t_0+z}^{\infty} f_L(z|D = 1, T = t) f_T(t|T \geq t_0)\, dt,\ z \in (0, \infty), \tag{1.29}$$

where $P(L = 0|D = 1, T = t)$ and $f_L(z|D = 1, T = t)$ can be calculated by Equations 1.21 and 1.22, and $f_T(t|T \geq t_0) = f_T(t)/P(T \geq t_0)$ is the conditional lifetime distribution. The validity of this mixed probability distribution can be proved by

$$P(L = 0|D = 1, T \geq t_0) + \int_0^{\infty} f_L(z|D = 1, T \geq t_0) dz = 1. \tag{1.30}$$

The actuarial life table from the United States Social Security Administration was used to estimate the lifetime distribution $f_T(t|T \geq t_0)$ (see `http://ssa.gov/OACT/STATS/table4c6.html`). The life table provides the conditional probability of death within one year from age 0 to age 119, denoted as $b_N = P(T < N + 1|T \geq N), N = 0, 1, \ldots, 119$. The conditional density can be approximated by

$$f_T(t = t_0 + N|T \geq t_0) = (1 - a_{t_0+N}) \prod_{i=1}^{N} a_{t_0+i-1},\ \forall N = 1, 2, \ldots, 120 - t_0, \tag{1.31}$$

where $a_N = 1 - b_N$. The final lifetime distribution was approximated by a step function, $f_T(t|T \geq t_0) \approx f_T(t = t_0 + N|T \geq t_0)$, for any $t \in (N, N + 1)$.

Because the lifetime T is random, the number of screening exams $K = \lceil (T - t_0)/\Delta \rceil$ is a function of T, hence it is also a random variable, with Δ as the screening

interval. We can see the final distribution of the lead time is a weighted average of different lengths of lifetimes. Additional simulations were done in Kendrick et al. (2015).

CHAPTER 2

ESTIMATION OF THE THREE KEY PARAMETERS

2.1 The National Lung Screening Trial Study

In 2002, the National Cancer Institute launched the National Lung Screening Trial (NLST), a randomized clinical trial that screened a high-risk population with either low-dose helical (spiral) computed tomography (CT) or standard chest X-ray (X-ray). The purpose of the study was to evaluate whether low-dose CT screening reduces lung cancer mortality comparing to chest radiography among high-risk individuals. NLST enrolled approximately 54,000 male and female current or former heavy smokers (with a smoking history of at least 30 pack-years, and at most 15 years since quitting if former smokers) aged 55 to 74 years between August 2002 and April 2004. Participants were randomized to two study arms in equal proportions: CT or X-ray. Participants were offered screening exams annually for 3 years, with the first screening performed soon after study entry. 15,537 male and 10,769 female participants in CT arm and 15,396 male and 10,634 female participants in X-ray arm had first screening exam. If any of the exam results was abnormal, then the screen was considered positive and more diagnostic tests should be done, such as biopsy. The median follow-up time was 6.5 years, the final results revealed participants in CT arm had a 15 to 20 percent lower lung cancer mortality than participants who received standard chest X-rays. The data we used are with the same format as shown in Table 1.2, including the

This chapter estimated the three key parameters using NLST data, which follows Liu et al. (2015).

number of participants in each screening exam, the number of screening detected and confirmed cancer cases, and the number of interval-incident cases, stratified by initial age.

Our study will focus on evaluating lung cancer screening using only CT arm data in NLST. The reason is 1). CT screening is the most current screening modality, commonly known with higher sensitivity. 2). There is little literature on the estimation of the three key parameters and lead time distribution in CT scan.

2.2 Method

Let the time variable t represents the participant's age. Then let $\beta(t)$ represents the sensitivity of the screening. Define $w(t)dt$ as the probability of a transition from S_0 to S_p during $(t, t+dt)$. Let $q(x)$ be the probability density function of the sojourn time in S_p, and let $Q(z) = \int_z^\infty q(x)\,dx$ be the survival function of the sojourn time in the preclinical state S_p.

For an initially asymptomatic heavy smoker of age t_0, who has no history of lung cancer, and suppose that the person plans to undergo K screening exams at ages $t_0 < t_1 < \cdots < t_{K-1}$, where $t_i = t_0 + i$ for annual screening exams in NLST study. Define the ith screening interval as the time interval between the ith and the $(i+1)$th screening exams, that is, (t_{i-1}, t_i), $i = 1, 2, , K-1$. We let $t_{-1} \equiv 0$. For each screening exam, let n_{i,t_0} be the total number of individuals in this cohort examined at the ith screening, s_{i,t_0} is the number of cases detected at the ith screening exam, and r_{i,t_0} is the number of cases diagnosed in the clinical state S_c within the interval (t_{i-1}, t_i), which is the interval cases.

For NLST data, the age of participants enrolled was between 55 to 74 at the study entry and three annual screening exams were offered ($K = 3$). Hence, the

likelihood function for all groups based on Equations 1.5 and 1.9 is

$$L = \prod_{t_0=55}^{74} \prod_{k=1}^{3} D_{k,t_0}^{s_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1 - D_{k,t_0} - I_{k,t_0})^{n_{k,t_0} - s_{k,t_0} - r_{k,t_0}}, \tag{2.1}$$

where D_{k,t_0} is the probability that an individual will be detected by the kth screening exam (at age t_{k-1}) given this person is in the state S_p. I_{k,t_0} is the probability of an individual being incident during the kth interval (t_{k-1}, t_k). These two probabilities can be calculated as in Equations 1.6, 1.7 and 1.8, respectively.

The parametric models for the three key parameters were chosen as following:

$$\beta(t|b_0, b_1) = \frac{1}{1 + \exp\{-b_0 - b_1(t - \bar{t})\}}, \tag{2.2}$$

$$w(t|\mu, \sigma^2) = 0.3 \cdot \frac{1}{\sqrt{2\pi}\sigma t} \exp\{-(\log t - \mu)^2/(2\sigma^2)\}, \quad t > 0, \sigma > 0, \tag{2.3}$$

$$q(x|\lambda, \alpha) = \alpha\lambda x^{\alpha-1} \exp(-\lambda x^{\alpha}), \quad \lambda > 0, \alpha > 0, \tag{2.4}$$

$$Q(x|\lambda, \alpha) = \exp(-\lambda x^{\alpha}), \quad \lambda > 0, \alpha > 0, \tag{2.5}$$

where $\bar{t}$ is the average age at entry in the study group, in this data, $\bar{t} = 61.4$ years. We also associate the sensitivity β with age t by a logistic link. As mentioned before, if $b_1 > 0$, then $\beta(t)$ will be a monotone increasing function of age t. The log-normal distribution was chosen for $w(t)$ with an upper limit of 30%. According to the NIH SEER database, the lifetime risk of lung cancer for the general population is about 7% for both genders (NCI, 2015). Since participants in NLST were heavy smokers, the risk would be higher than that, besides the fact that not all people in the preclinical state will progress into clinical cancer. This research proposes 30% as a reasonable upper limit for $w(t)$. A more detailed description of the parametric models can be found in Wu et al. (2005a, 2011). We choose a different sojourn time distribution than Wu et al. (2005a), where the previous research used log-logistic, and we use Weibull distribution here, both share the same property of mathematical simplicity,

and both are stable with two parameters. However, Weibull is more flexible in that the nth moments always exist.

2.3 Application

The six unknown parameters $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ were estimated based on CT arm data. The data were split into three groups: male, female and overall. Theoretically, the parameters have a domain of either $(-\infty, +\infty)$ or $(0, +\infty)$. The practical meaning of these parameters will limit them to a finite range. As was described in Wu et al. (2005a), the range for each parameter can be identified as: $0 < b_0 < 4$, $-0.1 < b_1 < 0.1$, $4.0 < \mu < 4.5$, $0.01 < \sigma^2 < 0.05$, $0.01 < \lambda < 0.5$, and $1.5 < \alpha < 4.0$.

Markov Chain Monte Carlo (MCMC) was used to draw posterior samples with non-informative Uniform priors. We partitioned the posterior simulation into three sub-chains, sampling the posterior for (b_0, b_1), (μ, σ^2) and (λ, α) separately. That is, there are three sampling steps for updating (b_0, b_1), (μ, σ^2) and (λ, α) in each iteration. Two simulations were carried out with different initial values that were over dispersed with respect to the target distribution. Each simulation was run for 130,000 iterations, with 30,000 burn-in steps, and after the burn-in steps, the posteriors were sampled every 200 steps, providing 500 posterior samples for the parameter vector θ. The 500 posterior samples from each of the two chains were pooled for the analysis, giving a total of 1000 posterior samples for θ. The MCMC trace and the posterior density of θ are plotted using the final 1000 posterior samples for θ of 3 groups: overall, male and female groups. Figure 2.1 shows the MCMC trace for of overall group, the MCMC trace for males and females are similar to Figure 2.1 and we omit here. Figures 2.2, 2.3 and 2.4 show the density plots for three groups, respectively. Bayesian output diagnosis showed that the chains had converged. The posterior estimates for six parameters and the standard deviations are listed in Table 2.1.

The age-dependent Bayesian estimates of the sensitivity β and the transition

density $w(t)$ for each group are listed in Table 2.2. Figures 2.5, 2.6 and 2.7 show posterior quantiles of sensitivity and transition probability for each group. From Equation 2.2, we can see $\beta(t)$ will be monotonic increasing with age t if $b_1 > 0$. In our cases, b_1 is greater than but is also closed to 0 in all cases. We did a Bayes hypothesis test for $H_0 : b_1 \leq 0$ versus $H_1 : b_1 > 0$. For the overall group which includes both genders, the posterior probability of a positive slope is $P(b_1 > 0|Data) = 0.532$; For males group, this posterior probability is $P(b_1 > 0|Data) = 0.513$; for females, this posterior probability is 0.651. Hence, the evidence of age effect is not significant in all groups. The age-dependent transition probability is a sub-PDF from our model construction. The posterior density curve of the transition probability could be seen from Figures 2.5, 2.6 and 2.7. The transition probability is not a monotone function of age, having a single maximum around age 70 for both males and females. The posterior mean sojourn time is 1.48 years for CT overall, 1.44 years for CT male and 1.62 years for CT female, with a posterior median of 1.47 years for CT overall, 1.41 years for CT male and 1.58 years for CT female, respectively. The 95% highest posterior density interval is (1.22, 1.77) for overall, (1.11, 1.78) for males and (1.21, 2.04) for females. The standard error for the sojourn time is 0.144 for CT overall, 0.185 for CT male and 0.221 for CT female.

Table 2.1: Bayesian posterior estimates for the 6 parameters in NLST data CT arm

	Mean	SD	2.5%	50%	97.5%
			Overall		
b_0	3.263	0.503	2.154	3.339	3.963
b_1	0.002	0.053	-0.094	0.005	0.094
μ	4.271	0.008	4.255	4.270	4.288
σ^2	0.022	0.002	0.018	0.022	0.027
λ	0.270	0.053	0.163	0.275	0.370
α	2.703	0.496	1.899	2.643	3.822
			Male		
b_0	2.923	0.622	1.705	2.950	3.939
b_1	0.002	0.058	-0.095	0.003	0.096
μ	4.268	0.010	4.249	4.268	4.288
σ^2	0.021	0.003	0.016	0.020	0.026
λ	0.306	0.079	0.140	0.312	0.452
α	2.713	0.601	1.715	2.672	3.903
			Female		
b_0	3.247	0.516	2.182	3.330	3.968
b_1	0.017	0.054	-0.091	0.026	0.096
μ	4.276	0.014	4.248	4.275	4.303
σ^2	0.026	0.004	0.019	0.026	0.034
λ	0.194	0.059	0.090	0.189	0.330
α	2.983	0.562	1.945	2.948	3.934

Table 2.2: Bayesian posterior estimates of β and $w(t)$ for each group

Age	Sensitivity β			Transition density $w(t)$		
	Median	Mean	SD	Median	Mean	SD
			Overall			
55	0.9642	0.9551	0.0306	0.0030	0.0030	3.07×10^{-4}
60	0.9657	0.9581	0.0238	0.0066	0.0066	3.86×10^{-4}
65	0.9642	0.9587	0.0220	0.0101	0.0100	5.63×10^{-4}
70	0.9616	0.9570	0.0256	0.0114	0.0114	6.01×10^{-4}
75	0.9613	0.9529	0.0343	0.0102	0.0102	3.96×10^{-4}
			Male			
55	0.9484	0.9360	0.0461	0.0029	0.0029	4.01×10^{-4}
60	0.9496	0.9398	0.0369	0.0067	0.0067	4.95×10^{-4}
65	0.9497	0.9396	0.0385	0.0104	0.0104	7.16×10^{-4}
70	0.9495	0.9355	0.0497	0.0118	0.0118	7.74×10^{-4}
75	0.9506	0.9274	0.0678	0.0105	0.0105	5.07×10^{-4}
			Female			
55	0.9601	0.9499	0.0337	0.0034	0.0034	4.98×10^{-4}
60	0.9641	0.9563	0.0255	0.0065	0.0065	5.76×10^{-4}
65	0.9665	0.9599	0.0228	0.0094	0.0094	7.75×10^{-4}
70	0.9666	0.9610	0.0256	0.0104	0.0104	8.31×10^{-4}
75	0.9710	0.9596	0.0332	0.0096	0.0095	6.00×10^{-4}

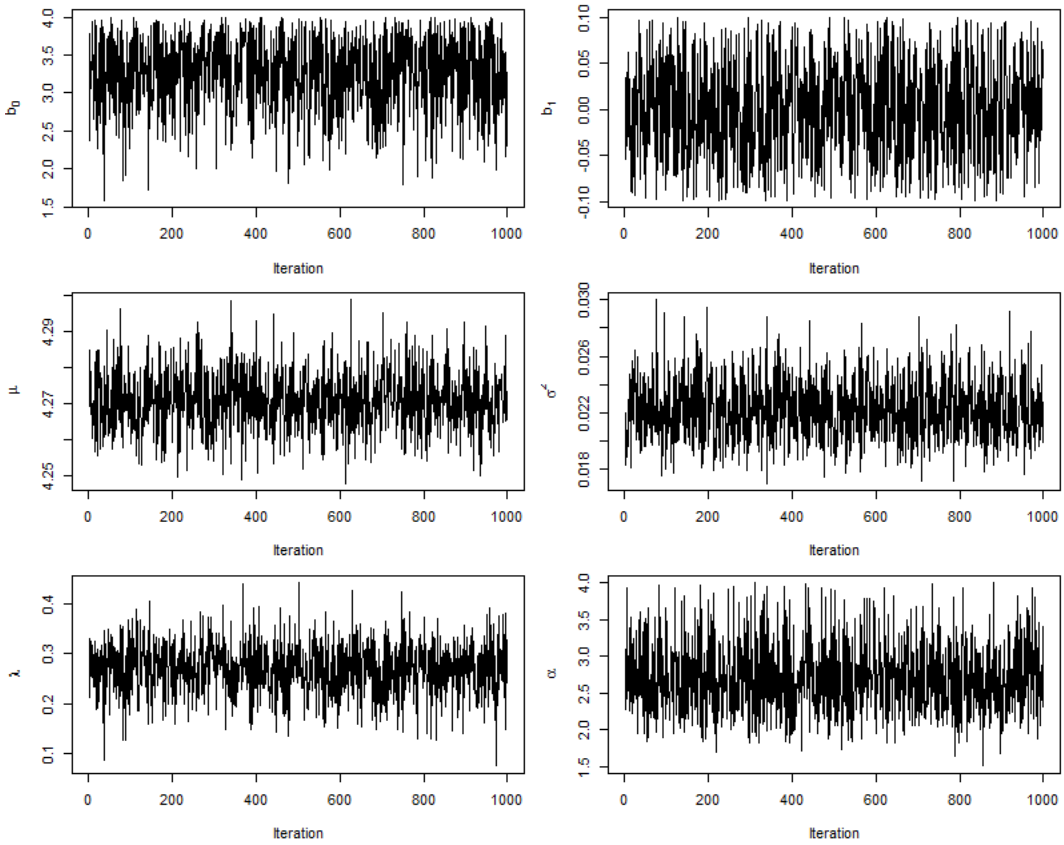

Figure 2.1: The MCMC trace plots of the parameters $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm overall group in NLST data

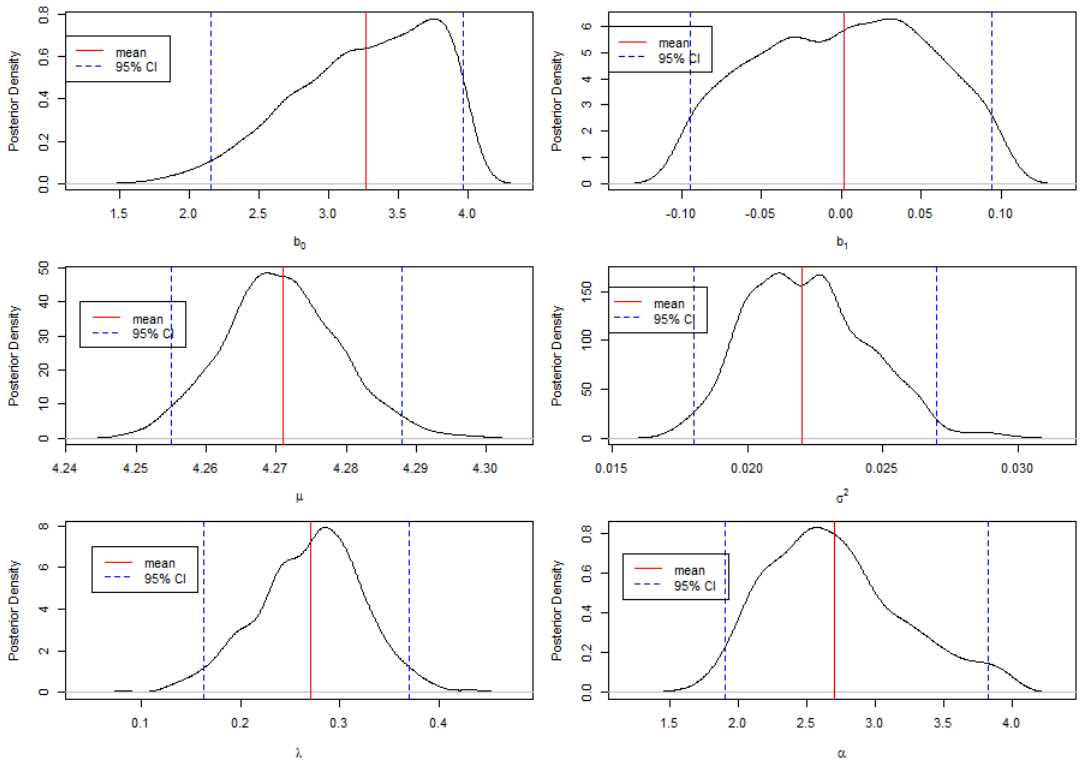

Figure 2.2: The posterior density plots of the parameters $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm overall group in NLST data

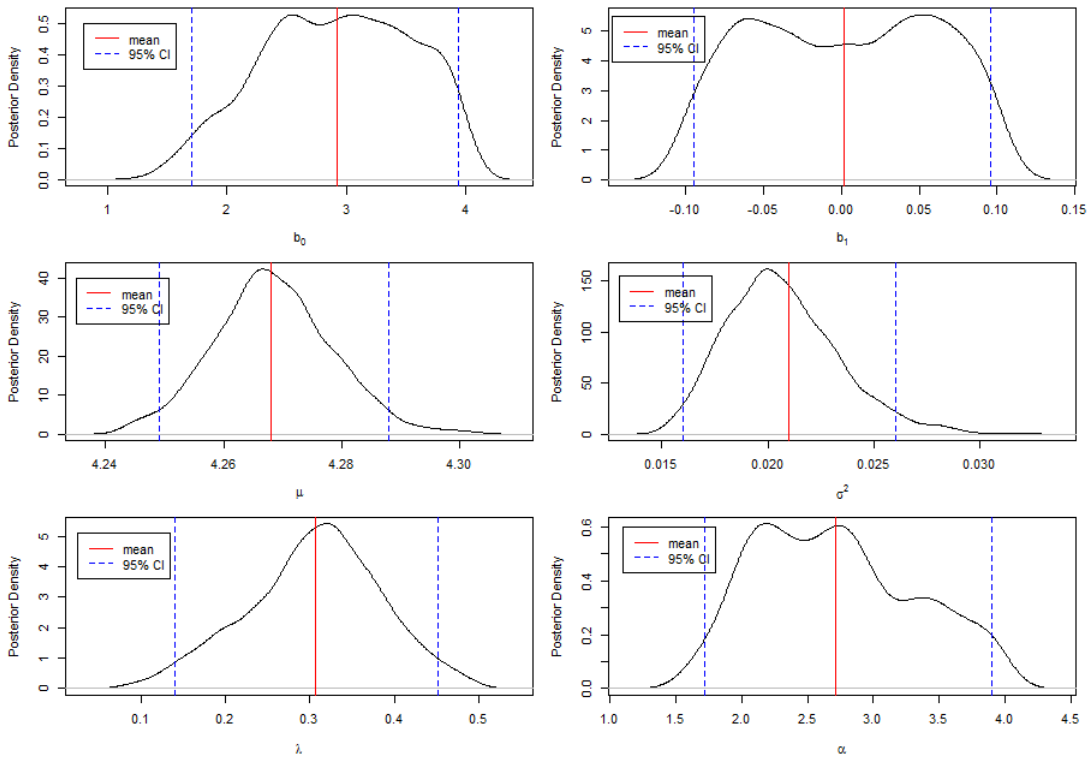

Figure 2.3: The posterior density plots of the parameters $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm male group in NLST data

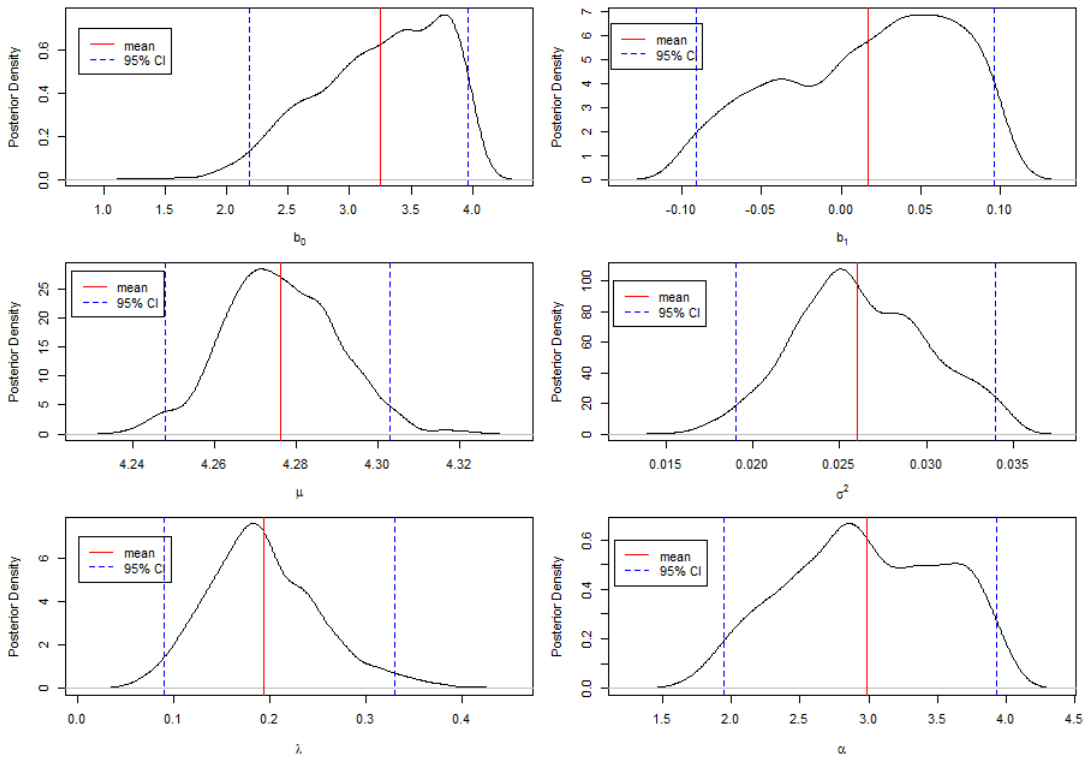

Figure 2.4: The posterior density plots of the parameters $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm female group in NLST data

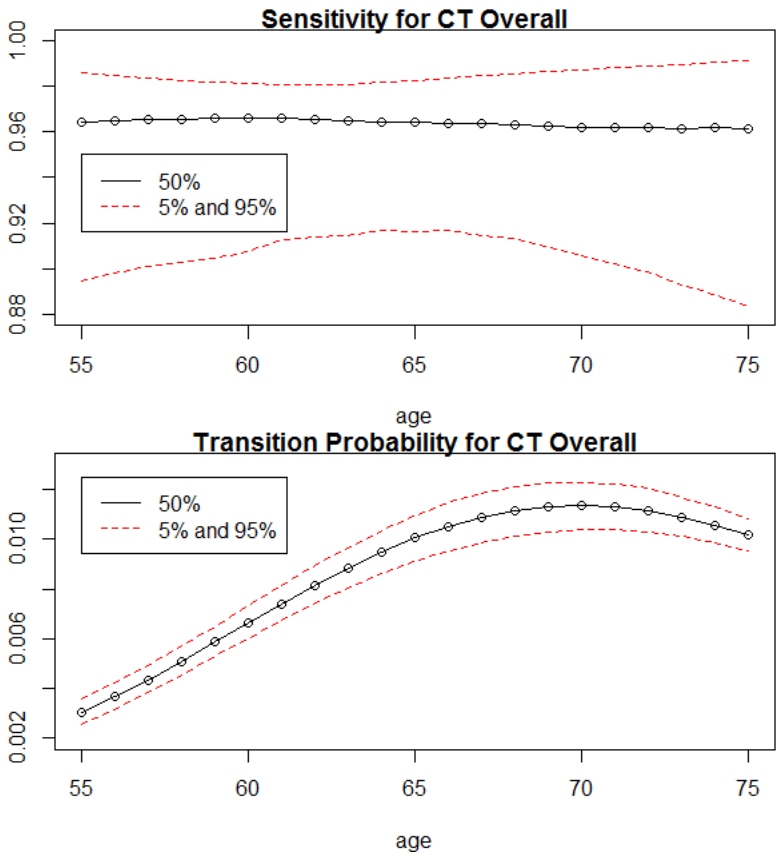

Figure 2.5: Posterior quantiles (5%, 50% and 95%) of sensitivity and transition probability for CT overall group

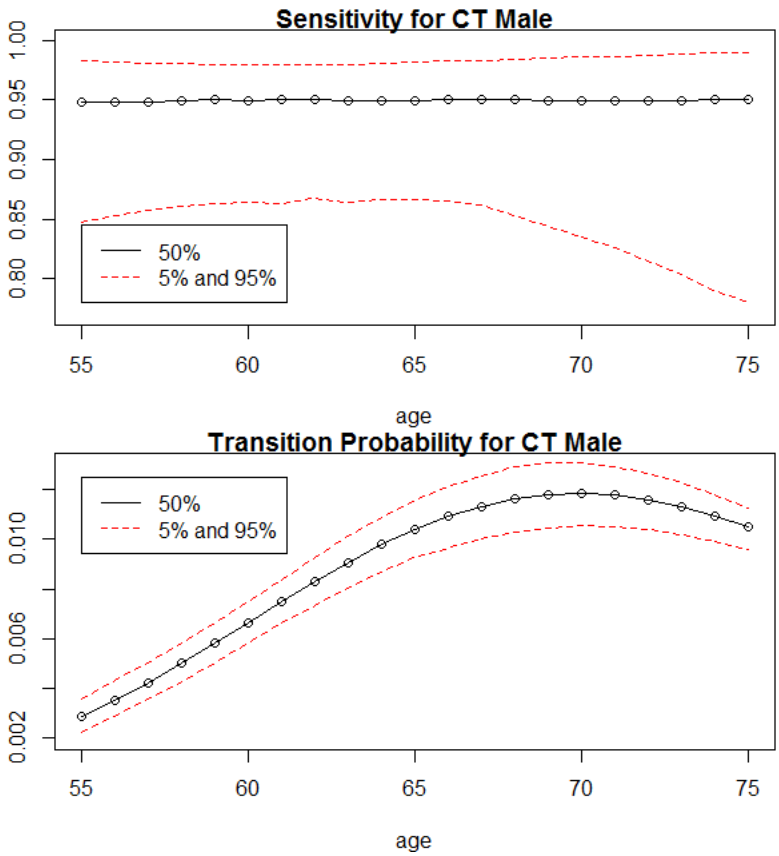

Figure 2.6: Posterior quantiles (5%, 50% and 95%) of sensitivity and transition probability for CT male group

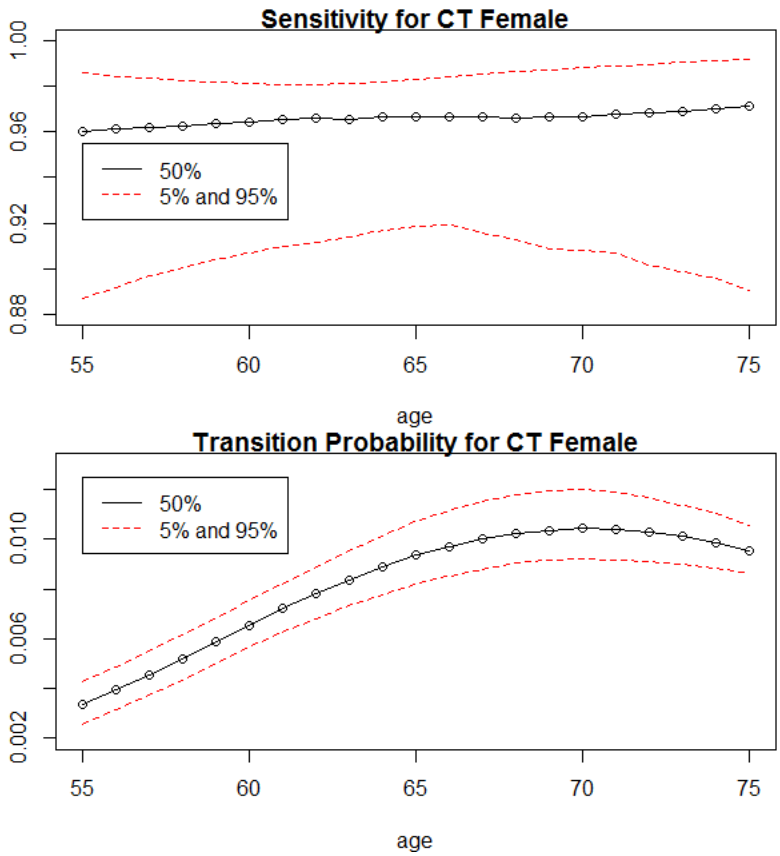

Figure 2.7: Posterior quantiles (5%, 50% and 95%) of sensitivity and transition probability for CT female group

2.4 Discussion

In this project, the three key parameters (screening sensitivity, the transition probability density and the sojourn time distribution) were estimated using Bayesian approach. The estimation was based on the NLST CT arm data.

For lung cancer screening, little research has been done in estimating the three key parameters. For instance, Jang et al. (2013) estimated the three key parameters using the Johns Hopkins Lung Project (JHLP) control group data. The control group was only administered X-ray screenings, and the estimated sensitivity was 56.8%. Kim and Erwin (2012) estimated the sensitivity as 79.9% using the JHLP study group data, in which both X-ray and sputum cytology were used. By using

Mayo Lung Project (MLP) male heavy smokers data, Wu et al. (2011) estimated the sensitivity of exams combined X-rays and sputum cytology was 89.4%. The screening sensitivity of sputum cytology as a supplement to the chest X-ray was 86.64%, which was estimated using the Lung Cancer Screening Program at the Memorial Sloan-Kettering Cancer Center (MSKC-LCSP) data (Chen et al., 2014).

Compared with these previous results, the sensitivity estimated in this study is around 95% for all the groups, which is much larger. This confirms that CT scan improves the lung cancer screening sensitivity compares to X-rays. In addition, it seems that the sensitivity of lung cancer screening using CT scan does not depend on the age of patients. Pinsky et al. (2015) and Aberle et al. (2013) reported the sensitivity was 93.5% and 94.4% for the NLST CT arm, respectively, which is also close to our estimation of sensitivity.

The transition probability from disease-free to preclinical state has a peak around age 70 for both men and women. The transition probability also has a single maximum around age 70 in Chen et al. (2014). The "SEER Fast Fact Stats" (NCI, 2015) shows that the highest percent of new lung cancer cases is in 65-74 age group. Our results are consistent with that fact.

In the MLP study, the mean sojourn time was 2.2 years (Wu et al., 2011), the mean sojourn time for male heavy smokers in MSKC-LCSP data was about 3.35 years. The posterior mean sojourn in this study is around 1.5 years for both gender groups in this study. The sojourn time is relatively short compares to other studies, it may be caused by the fact that the estimated sensitivity and sojourn time are correlated using this likelihood.

In summary, this project focuses on the estimation of the three key parameters: sensitivity, sojourn time distribution and transition probability density from the disease-free to the preclinical state. It lays a foundation for the estimation of other interesting terms, such as lead time, over diagnosis, long term outcomes in the future,

because all these interesting terms can be expressed as a function of the three key parameters.

CHAPTER 3

ESTIMATION OF THE LEAD TIME DISTRIBUTION

3.1 Method

The lead time distribution when human lifetime is fixed and considered as a random variable were derived in Wu et al. (2007) and Wu et al. (2012), respectively. They are both functions of the there key parameters. The distribution of lead time consists of a point mass at zero and a positive continuous probability density. Because the lead time will be zero for the interval incident cases and it will be greater than zero for the screen-detected cases. In this project, a persons lifetime T is considered as a random variable, and the number of screenings $K = \lceil (T - t_0)/\Delta \rceil$ is a function of T, hence it is also random. The derived probability formulae for lead time when lifetime is random can be found in Section 1.3.3. From Equations 1.28 and 1.29, when lifetime is a random variable, the distribution of lead time is a weighted average distribution of different lengths of lifetimes.

The 1000 posterior samples of the six unknown parameters obtained in the first project are used to estimate the lead time distribution. We use 1000 Bayesian posterior samples θ_i^* in the inference for the lead time distribution, where θ_i^* is one of the posterior samples generated using MCMC, and $i = 1, 2, \ldots, 1000$. Then the posterior predictive distribution of lead time is

$$f_L^{NLST}(l) \approx \frac{1}{n} \sum_{i=1}^{n} f_L^{NLST}(l|\theta_i^*), \tag{3.1}$$

where $f_L^{NLST}(l|\theta_i^*)$ is the mixture distribution defined by Equations 1.28 and 1.29.

3.2 Application

To obtain the projected lead time distribution for cohorts with different initial screening ages and different future schedules, we conducted simulations with different settings. For each gender, we assumed that there are four cohorts of initially asymptomatic individuals, with initial age t_0=55, 60, 65, and 70, respectively; Then within each cohort, we examined four different future screening intervals at Δ=12, 18, 24, and 30 months. We present simulation results of these 16 scenarios for both men and women.

Table 3.1 and Table 3.2 present the Bayesian predictive inference for the lead time in years and the probability of no-early-detection (P_0) and early-detection (1-P_0) for men and women, respectively. The probability that the lead time is zero (P_0) and its corresponding 95% C.I., the probability that the lead time is positive (1-P_0) and its corresponding standard deviation are reported as percentages. The mean lead time (EL) in years was estimated by $EL = 0 \times P(L = 0|D = 1, T \geq t_0) + \int_0^\infty z f_L(z|D = 1, T \geq t_0)\, dz = \int_0^\infty z f_L(z|D = 1, T \geq t_0)\, dz$. To measure the location and the spread for the distribution of lead time greater than zero, we estimated the median and divided it over the interquartile range (Med/IQR), where IQR is the first quartile subtracted from the third quartile.

For both genders, simulation results show a clear trend that the probability of no-early-detection will increase as the screening time interval increases within the same age group. For example, if a 60-year-old man begins to take CT screening exams annually (i.e., Δ=12 months) and assuming that he will develop lung cancer at some point in his lifetime, the chance that he will not be detected early by the regular screening exams is 11.65%. However, this probability of no-early-detection will increase to 36.35% if the exams are biennial (i.e., Δ=24 months). Across the

age groups, although the probability of no-early-detection does not seem to have significant differences, it tends to decrease as the initial screening age increases with the same screening interval. In addition, for all 16 scenarios of different initial ages and screening intervals, the probabilities of no-early-detection for men are larger than that for women. It seems that men have smaller chances to be detected early by CT screening exam than women do.

The probability density curves of the lead time for men and women are shown in Figure 3.1 and Figure 3.2, respectively. For initial screening age t_0= 55, 60, 65 and 70, four curves represent four different screening intervals (Δ=12, 18, 24 and 30 months). For men and women with the same initial screening age and screening interval, the mean lead time appears longer for women than for men. For both genders, the mean lead time becomes shorter as screening interval increases within initial age group. In other words, if the screening exams are more frequent, the lead time will be longer. This result matches with the trend for the probability of lead time is zero. That is, the increase in the mean lead time when screening interval decreases is due partly to the smaller point mass at zero of the lead time. However, the mean lead time seems stable across different initial age groups, which implies the length of lead time may only relate to gender and screening interval.

Table 3.1: A projection of the lead time distribution for men by initial screening age age and screening interval

Δ (months)	P_0 (95% C.I.)	$1 - P_0$ (s.d.)	EL (s.d.)	Med/IQR
Age at initial screen t_0=55				
12	11.80 (7.86, 16.98)	88.20 (2.34)	0.87 (0.69)	0.94
18	24.42 (18.07, 31.03)	75.58 (3.29)	0.68 (0.68)	0.94
24	37.06 (28.68, 45.34)	62.94 (4.34)	0.55 (0.66)	0.83
30	47.41 (37.92, 55.62)	52.59 (4.66)	0.45 (0.64)	0.83
Age at initial screen t_0=60				
12	11.65 (7.28, 17.76)	88.35 (2.61)	0.86 (0.68)	0.94
18	23.98 (17.14, 31.20)	76.02 (3.59)	0.68 (0.68)	0.83
24	36.35 (27.67, 45.06)	63.65 (4.54)	0.55 (0.66)	0.83
30	46.49 (36.73, 54.86)	53.51 (4.83)	0.46 (0.64)	0.83
Age at initial screen t_0=65				
12	11.58 (6.70, 19.16)	88.42 (3.12)	0.85 (0.68)	0.94
18	23.53 (16.19, 32.29)	76.47 (4.12)	0.67 (0.67)	0.83
24	35.51 (26.17, 45.14)	64.49 (4.92)	0.55 (0.65)	0.83
30	45.31 (34.90, 54.44)	54.69 (5.11)	0.46 (0.63)	0.94
Age at initial screen t_0=70				
12	11.66 (6.19, 21.83)	88.34 (3.97)	0.82 (0.67)	1.06
18	23.14 (15.02, 34.52)	76.86 (4.97)	0.66 (0.66)	0.83
24	34.55 (24.56, 45.82)	65.45 (5.53)	0.54 (0.65)	0.94
30	43.86 (32.82, 54.44)	56.14 (5.57)	0.46 (0.62)	0.81

Table 3.2: A projection of the lead time distribution for women by initial screening age age and screening interval

Δ (months)	P_0 (95% C.I.)	$1 - P_0$ (s.d.)	EL (s.d.)	Med/IQR
Age at initial screen t_0=55				
12	6.89 (4.06, 10.73)	93.11 (1.73)	1.06 (0.73)	1.17
18	16.88 (11.28, 23.55)	83.12 (3.09)	0.85 (0.74)	0.95
24	28.85 (20.06, 38.55)	71.15 (4.63)	0.69 (0.73)	0.94
30	39.84 (28.97, 50.17)	60.16 (5.42)	0.57 (0.71)	0.94
Age at initial screen t_0=60				
12	6.76 (3.91, 10.68)	93.24 (1.76)	1.05 (0.72)	1.17
18	16.53 (10.92, 23.18)	83.47 (3.13)	0.85 (0.73)	0.95
24	28.26 (19.41, 37.85)	71.74 (4.65)	0.69 (0.73)	0.94
30	39.06 (27.92, 49.33)	60.94 (5.44)	0.58 (0.71)	0.85
Age at initial screen t_0=65				
12	6.65 (3.72, 10.56)	93.35 (1.87)	1.03 (0.72)	1.05
18	16.13 (10.41, 22.92)	83.87 (3.23)	0.84 (0.73)	0.94
24	27.54 (18.52, 37.16)	72.46 (4.71)	0.69 (0.72)	0.94
30	38.03 (26.98, 48.14)	61.97 (5.47)	0.58 (0.71)	0.85
Age at initial screen t_0=70				
12	6.58 (3.49, 11.20)	93.42 (2.10)	1.01 (0.72)	1.06
18	15.74 (9.95, 22.72)	84.26 (3.45)	0.82 (0.72)	0.94
24	26.71 (17.58, 36.35)	73.29 (4.82)	0.68 (0.71)	0.85
30	36.79 (25.86, 46.84)	63.21 (5.53)	0.58 (0.70)	0.75

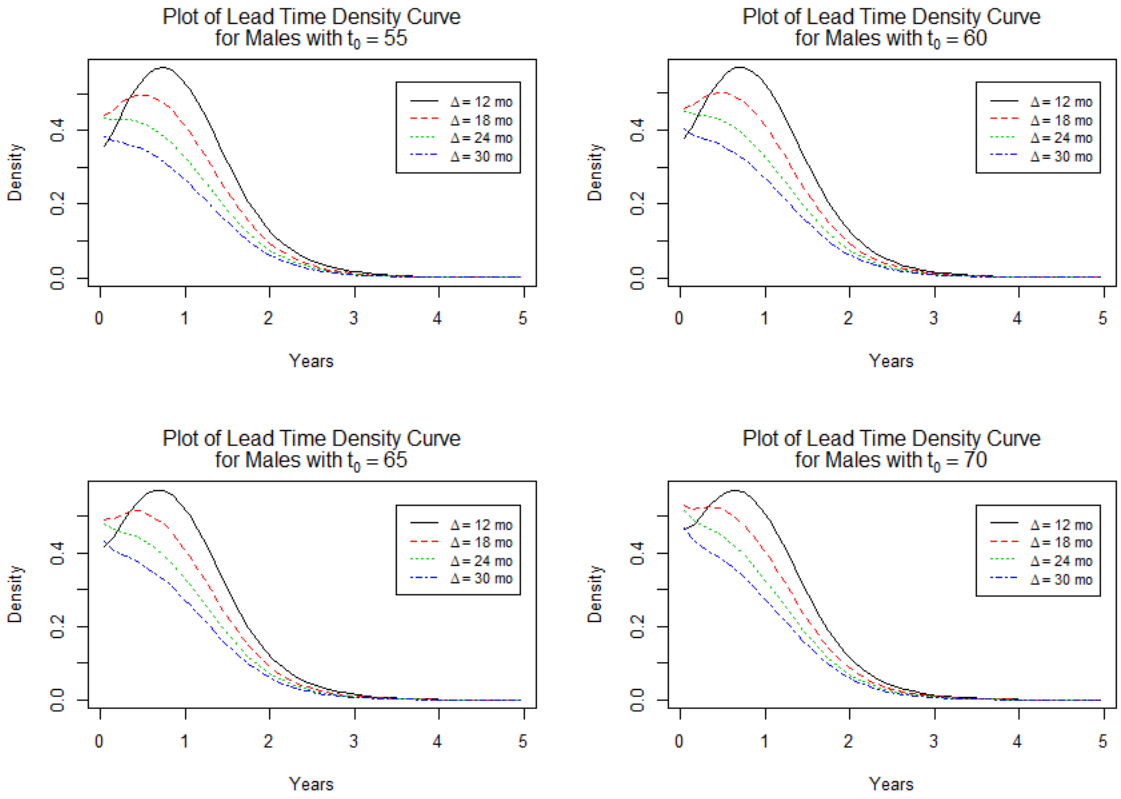

Figure 3.1: The sub-PDF curves of the lead time for males: Four curves representing different screening intervals are plotted for $t_0 = 55$ (upper left panel), $t_0 = 60$ (upper right panel), $t_0 = 65$ (bottom left panel) and $t_0 = 70$ (bottom right panel), respectively. The area under the curve is $1 - P_0$.

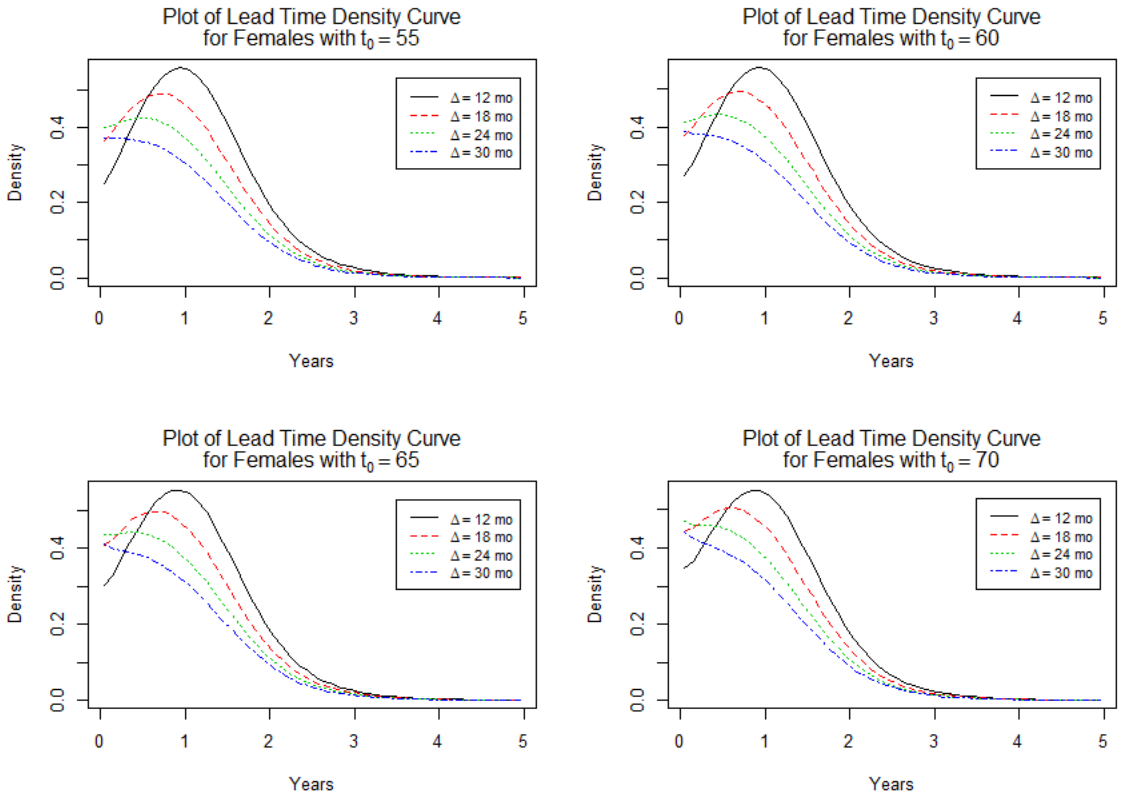

Figure 3.2: The sub-PDF curves of the lead time for females: Four curves representing different screening intervals are plotted for $t_0 = 55$ (upper left panel), $t_0 = 60$ (upper right panel), $t_0 = 65$ (bottom left panel) and $t_0 = 70$ (bottom right panel), respectively. The area under the curve is $1 - P_0$.

3.3 Discussion

In this project, we estimated the lead time distribution when lifetime T is a random variable for lung cancer screening using CT. When lifetime is treated as a random variable, the lead time distribution is a weighted average of lead time under different lifetime lengths. We explored the relation of lead time by gender for several initial screening ages and different screening intervals.

We compared our results to the simulation results using the MLP data in Wu et al. (2011). The participants of MLP were all male heavy smokers, and they took a screening test every four months. Each test includes a chest X-ray and a three-day pooled sputum cytology sampling. The authors made inference on lead time distribution while the initial screening age t_0 was set at 45 years and human lifetime was fixed to 75 year. The simulation results showed the probability of no-early-detection was 32.74% for men with an annual screening interval (i.e. Δ=12 months). Although we did not present the simulation results for initial age t_0=45, the probability of no-early-detection is 11.81% for men with an annual screening interval and initial screening age t_0=55 years, which is much smaller than 32.74%. Therefore, it seems that lung cancer screening using CT will result in more early-detected cases than a chest X-ray and a three-day pooled sputum cytology sampling do. We also compared the lead time estimates to the lead time distribution estimated using the JHLP data (Jang et al., 2013). In the JHLP study, a projection of the lead time distribution was obtained for male smokers received screening exam using X-ray. To estimate the lead time distribution, the human lifetime was considered as a random variable in the JHLP study, as well as in this NLST study. For initial screening age t_0=60, the probability of no-early-detection is 46.54% if the screening exams are given annually, and it was 59.97% if the screening exams are offered biennially. For the NLST CT arm, this probability is 36.36% and 46.49% for annual and biennial exams,

respectively. The dramatic decrease in probability of no-early-detection in screening using CT may demonstrate that lung cancer screening using CT is more effective than using X-ray and X-ray combined with pooled sputum cytology sampling.

The lifetime distribution was calculated based on the actuarial life table 2013 published in 2016 from the United States Social Security Administration (SSA) (SSA, 2016). Some may argue that we should use the lifetime reports from the same year when the NLST study underwent to better represent the study participants. In fact, the life table does not change much over the years, so we just use the recent published data to obtain the lifetime distribution. Furthermore, some may also say we need to use the lifetime distribution for lung cancer population instead of using the one for the general population. However, the cancer population is small and there is no such source for us to obtain the life table, so we just use the whole population data to estimate the lifetime time distribution.

CHAPTER 4

ESTIMATION OF THE LEAD TIME DISTRIBUTION FOR INDIVIDUALS WITH SCREENING HISTORY

In previous studies, it is assumed that a person has no screening history before entering the study. However, participants aged 55 and over may already have at least one prior lung cancer screening exam in the past and look healthy. In the third project, we extend the models of lead time distribution developed in Wu et al. (2007) and Wu et al. (2012).

4.1 Method

Lead time distribution for individuals with screening history can be derived as following. We define D as a binary random variable with $D = 1$ indicating a person develops (clinical) cancer before death and $D = 0$ indicating the person is cancer free before death. The time variable t represents a person's age, and T represents a person's lifetime. If a person already had K_1 screening exams, we define this event as

$$H_{K_1} = \left\{ \begin{array}{l} \text{An individual had screening exams at age } t_0 < t_1 < \cdots < t_{K_1-1}, \\ \text{no cancer was detected,} \\ \text{and the person is asymptomatic at his or her current age} \end{array} \right\}.$$

4.1.1 Lead time distribution for individuals with screening history when T is fixed

Suppose an individual has received K_1 screening exams, and he or she will continue with K screenings. To derive this lead time distribution, we need to calculate two parts of the mixture distribution as in Equations 1.21 and 1.22. That is, the conditional probability $P(L = 0|D = 1, H_{K_1}, T = t_{K_1+K})$ and conditional probability density $f_L(z|D = 1, H_{K_1}, T = t_{K_1+K})$. When the lifetime $T = t_{K_1+K}$ is a fixed value, the distribution of lead time is

$$P(L = 0|D = 1, H_{K_1}, T = t_{K_1+K}) = \frac{P(L = 0, D = 1, H_{K_1}|T = t_{K_1+K})}{P(D = 1, H_{K_1}|T = t_{K_1+K})}, \tag{4.1}$$

$$f_L(z|D = 1, H_{K_1}, T = t_{K_1+K}) = \frac{f_L(z, D = 1, H_{K_1}|T = t_{K_1+K})}{P(D = 1, H_{K_1}|T = t_{K_1+K})}. \tag{4.2}$$

The following probability will be calculated separately, $P(D = 1, H_{K_1}|T = t_{K_1+K})$, $P(L = 0, D = 1, H_{K_1}|T = t_{K_1+K})$ and $f_L(z, D = 1, H_{K_1}|T = t_{K_1+K})$. $P(D = 1, H_{K_1}|T = t_{K_1+K})$ is the probability that a person develops clinical cancer after K_1 screening exams given lifetime $T = t_{K_1+K}$. $P(L = 0, D = 1, H_{K_1}|T = t_{K_1+K})$ is the probability that an individual being an interval case after K_1 exams given lifetime $T = t_{K_1+K}$. They can be calculated by

$$\begin{aligned}
&P(D = 1, H_{K_1}|T = t_{K_1+K}) \\
&= \sum_{i=0}^{K_1-1} (1 - \beta_i) \cdots (1 - \beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)[Q(t_{K_1} - x) - Q(T - x)]\, dx \\
&\quad + \int_{t_{K_1-1}}^{t_{K_1}} w(x)[Q(t_{K_1} - x) - Q(T - x)]\, dx + \int_{t_{K_1}}^{T} w(x)[1 - Q(T - x)]\, dx.
\end{aligned} \tag{4.3}$$

$$P(L = 0, D = 1, H_{K_1}|T = t_{K_1+K}) = I_{K_1+K,K_1+1} + I_{K_1+K,K_1+2} + \cdots + I_{K_1+K,K_1+K}, \tag{4.4}$$

where

$$I_{K_1+K,\ j} = \sum_{i=0}^{j-1}(1-\beta_i)\cdots(1-\beta_{j-1})\int_{t_{i-1}}^{t_i} w(x)[Q(t_{j-1}-x)-Q(t_j-x)]\,dx + \int_{t_{j-1}}^{t_j} w(x)[1-Q(t_j-x)]\,dx,\ \text{for } j = K_1+1,\cdots,K_1+K. \tag{4.5}$$

$$\begin{aligned} &f_L(z, D=1, H_{K_1}|T=t_{K_1+K}) \\ &\quad= \sum_{i=K_1}^{j-1}\beta_i\left\{\sum_{r=0}^{i-1}(1-\beta_r)\cdots(1-\beta_{i-1})\int_{t_{r-1}}^{t_r} w(x)q(t_i+z-x)\,dx \right. \\ &\quad\left. + \int_{t_{i-1}}^{t_i} w(x)q(t_i+z-x)\,dx\right\}, \quad \text{if } T-t_j < z \le T-t_{j-1}, \\ &\quad \text{for } j = K_1+1,\ldots,K_1+K. \end{aligned} \tag{4.6}$$

We have proved that this mixed probability distribution is valid since

$$P(L=0|D=1,H_{K_1},T=t_{K_1+K}) + \int_0^{T-t_{K_1}} f_L(z|D=1,H_{K_1},T=t_{K_1+K})\,dz \equiv 1. \tag{4.7}$$

The detailed derivation and proof can be found in Appendix.

4.1.2 Lead time distribution for individuals with screening history when T is a random variable

When lifetime T is a random variable, the lead time distribution when T is greater than the current age t_{K_1} can be obtained by

$$P(L=0|D=1,H_{K_1},T\ge t_{K_1}) = \int_{t_{K_1}}^{\infty} P(L=0|D=1,H_{K_1},T=t) f_T(t|T\ge t_{K_1})\,dt, \tag{4.8}$$

$$f_L(z|D=1,H_{K_1},T\geq t_{K_1}) \\ =\int_{t_{K_1}+z}^{\infty} f_L(z|D=1,H_{K_1},T=t)f_T(t|T\geq t_{K_1})\,dt, z\in(0,\infty), \tag{4.9}$$

where $P(L=0|D=1,H_{K_1},T\geq t_{K_1})$ and $f_L(z|D=1,H_{K_1},T\geq t_{K_1})$ were given in Equations 4.1 and 4.2, respectively. Again, the conditional lifetime distribution density $f_T(t|T\geq t_{K_1})$ can be estimated using the actuarial life table, as shown in Equation 1.31.

4.2 Simulation Study

Screening for breast, lung, colon and cervical cancers are recommended by the USPSTF (2016), each of these cancers has different screening methods, and the sensitivities for these screening methods are not the same. In addition, the speed of cancer grows and spreads may also vary for different disease, which means the length of sojourn time in preclinical state is different. To explore the characteristics of newly developed lead time distribution for different cancer diseases and different screening schedules, simulations were done under each of the combinations of following settings:

1. Three different initial screening ages: $t_0 = 56$, 60 and 64 years.
2. Two different screening sensitivities: $\beta = 0.7$ and 0.9.
3. Three different mean sojourn time: MST = 2, 5 and 10 years.

For a given initial screening age t_0, we examined the lead time distribution for four current ages t_{K_1} with four years interval. That is, we conducted simulations by setting $t_{K_1} = 60$, 64, 68 and 72 when $t_0 = 56$, $t_{K_1} = 64$, 68, 72 and 76 when $t_0 = 60$, and $t_{K_1} = 68$, 72, 76 and 80 when $t_0 = 64$. For each combination of initial screening ages, sensitivities, mean sojourn times and current ages, we considered four screening schedules in the past and in the future with intervals $(\Delta_1, \Delta_2) = (1,1)$, $(2,1)$, $(1,2)$ and $(2,2)$. For example, $(\Delta_1, \Delta_2) = (1,2)$ means that the individual received annual

screening exams in the past and will take screening exams biennially in the future. We did not consider gender effect in this simulation study and only used lifetime table for males.

In the simulation, the parametric models of the transition probability and the sojourn time are the same with aforementioned Equations 2.3 and 2.4. μ and σ^2 were decided based on the mode of the log-normal distribution, here we let the mode be 70, as most of the cancer cases are diagnosed around age 70 years. For different mean sojourn time, we chose different values of λ. The values of the unknown parameters in the simulation were chosen as shown in Table 4.1.

Table 4.1: Values of unknown parameters in simulation study

	Parameter	Settings	Value
Sensitivity	N/A	Sensitivity is a fixed value	0.7 or 0.9
Transition Probability	μ	The mode of the log-normal	4.4
	σ^2	distribution is set to be 70	0.16
Sojourn Time	λ	MST=2	0.1963
		MST=5	0.0314
		MST=10	0.0079
	α	α is a fixed value	2

Simulation results for MST = 2, 5 and 10 are shown in Tables 4.2-4.4, respectively. In each table, we report the probability of lead time is zero P_0 as percentages. When the lead time is positive, we also report the mean lead time EL and its standard deviation, the median of lead time and the mode of lead time in years. As I mentioned in previous sections, P_0 stands for the probability that the person is not early-detected by the screening exams. Longer lead time means the person benefits more from the screening program since the treatments and interventions can be given earlier.

Intuitively, the length of lead time highly depends on the length of sojourn time. Usually, the longer sojourn time will lead to a longer lead time. We can

also see this from our simulation results. For example, the probability of no-early-detection P_0 is 20.23% and the mean lead time is 1.11 for an individual who started biennial screening exam at age 56 and will begin screening annually from current age $t_{K_1} = 64$, given that MST= 2 and $\beta = 0.7$. The probability P_0 decreases to 6.87% and the mean lead time increases to 3.29 when MST= 5. The probability P_0 decreases to only 3.82% and the mean lead time increases to 5.92 when MST= 10.

For different sensitivities, it is clear that larger screening sensitivity will result in higher probability of early-detection $(1-P_0)$. The probability of no-early-detection P_0 is smaller and the mean lead time is also longer for $\beta = 0.9$ compared to $\beta = 0.7$.

By examining three different initial screening ages t_0, we want to see if the initial screening age affects the lead time distribution given a person looks healthy at current age, or if the length of screening history has any influence on the lead time distribution given the same current age. However, we found that the lead time distribution tends to be the same for different t_0 if the current age t_{K_1} is fixed. To illustrate, simply look at the results of $t_{K_1} = 68$ and $t_{K_1} = 72$. Because we ran simulations of $t_{K_1} = 68$ and $t_{K_1} = 72$ for all three initial screening ages, and we can compare the results of these two current ages separately across three initial screening age groups. In Tables 4.2-4.4, the results of (t_0, t_{K_1}) =(56, 68), (60, 68) and (64, 68) are almost the same. It is also true when $t_{K_1} = 72$. This indicates that the screening initial age does not seem to affect the lead time distribution too much as long as the person still looks healthy at current age. Figure 4.1 shows the density plots of the lead time is positive for $t_{K_1} = 68$ and $t_{K_1} = 72$. In the figure, each curve actually represents the density of three different initial screening ages ($t_0 = 56$, 60 and 64), because the curves overlap each other and we can only observe four curves in each panel.

The probability of no-early-detection P_0 is slightly increasing with a participant's current age given that all other factors are the same, which means the younger

participants will benefit more from the screening program. This increasing is more obvious in the simulation when MST is larger. For example, in Table 4.3, the probability P_0 is 17.21% and the mean lead time is 2.80 for an individual who started annual screening exam at age 56, and will begin screening biennially from current age $t_{K_1} = 60$, given the screening sensitivity $\beta = 0.7$. The probability P_0 slightly goes up to 20.19% and the mean lead time becomes 2.36 when the individual's age is 72. Figure 4.2 gives percentages of P_0 and P_1 $(1 - P_0)$, we can see P_0 increases as t_{K_1} increases for all screening schedules. Since the results are the same for different t_0, we put results of all t_{K_1} together in the bar plots regardless of t_0.

For a given combination of sensitivity, initial age and current age, we can compare the results of different screening schedules. For the cases with the same Δ_1 but different Δ_2, the lead time distribution tends to be very similar. For example, we can compare the results of $(\Delta_1, \Delta_2) = (1, 1)$, $(1, 2)$ and $(2, 1)$. It is easy to see that the results for $(\Delta_1, \Delta_2) = (1, 1)$ are significantly different from the results for $(\Delta_1, \Delta_2) = (1, 2)$, but the results for $(\Delta_1, \Delta_2) = (1, 1)$ and $(\Delta_1, \Delta_2) = (2, 1)$ are very close. We can also see this from the PDF curves of lead time when MST = 2, 5 and 10, as shown in Figures 4.3-4.5, respectively. As the results are similar, we only present curves of $t_0 = 56$ and $\beta = 0.7$ for different t_{K_1} and MST. The PDF curves for lead time with same future screening interval Δ_2 almost overlap each other for given initial screening age, current age, sensitivity and mean sojourn time.

However, we can still find a trend that larger Δ_1 will result in smaller P_0 and longer mean lead time if Δ_2 remains the same, and it is more obvious when MST is longer. In Table 4.4, the probability P_0 is 9.19% and the mean lead time is 5.58 for an individual whose initial screening age $t_0 = 56$ and current age $t_{K_1} = 60$ with screening schedules $(\Delta_1, \Delta_2) = (1, 2)$ given the screening sensitivity $\beta = 0.7$. The probability P_0 decreases to 8.64% and the mean lead time increases to 5.63 if the individual's past screening interval $\Delta_1 = 2$. We can see the trend more clearly in Figure 4.6.

Table 4.2: A projection of the lead time distribution for individuals with screening history by current age and screening intervals, with MST=2

(Δ_1, Δ_2)	$\beta = 0.7$				$\beta = 0.9$			
(years)	P_0	EL (s.d.)	Median	Mode	P_0	EL (s.d.)	Median	Mode
initial screening age $t_0 = 56$, current age $t_{K_1} = 60$								
(1,1)	20.10	1.14 (1.05)	1.25	0.85	10.32	1.33 (1.04)	1.35	0.95
(2,1)	20.02	1.13 (1.04)	1.25	0.75	10.05	1.32 (1.04)	1.35	0.95
(1,2)	41.26	0.76 (0.97)	1.15	0.35	28.00	0.95 (1.01)	1.15	0.45
(2,2)	40.69	0.76 (0.96)	1.15	0.35	27.08	0.96 (1.00)	1.15	0.45
initial screening age $t_0 = 56$, current age $t_{K_1} = 64$								
(1,1)	20.34	1.12 (1.04)	1.25	0.75	10.45	1.31 (1.03)	1.35	0.95
(2,1)	20.23	1.11 (1.03)	1.25	0.65	10.10	1.30 (1.03)	1.35	0.95
(1,2)	41.34	0.75 (0.96)	1.15	0.15	28.09	0.94 (1.00)	1.15	0.45
(2,2)	40.63	0.75 (0.96)	1.05	0.15	26.96	0.94 (0.99)	1.15	0.45
initial screening age $t_0 = 56$, current age $t_{K_1} = 68$								
(1,1)	20.69	1.09 (1.03)	1.25	0.65	10.63	1.29 (1.03)	1.25	0.85
(2,1)	20.53	1.08 (1.02)	1.25	0.65	10.19	1.27 (1.02)	1.25	0.85
(1,2)	41.46	0.74 (0.95)	1.05	0.15	28.21	0.92 (0.99)	1.15	0.45
(2,2)	40.58	0.74 (0.94)	1.05	0.15	26.81	0.93 (0.98)	1.05	0.15
initial screening age $t_0 = 56$, current age $t_{K_1} = 72$								
(1,1)	21.15	1.06 (1.01)	1.15	0.65	10.88	1.25 (1.01)	1.25	0.85
(2,1)	20.93	1.05 (1.01)	1.15	0.65	10.32	1.24 (1.01)	1.25	0.65
(1,2)	41.58	0.72 (0.93)	1.05	0.15	28.36	0.90 (0.97)	1.05	0.15
(2,2)	40.50	0.72 (0.93)	1.05	0.15	26.61	0.91 (0.97)	1.05	0.15
initial screening age $t_0 = 60$, current age $t_{K_1} = 64$								
(1,1)	20.34	1.12 (1.04)	1.25	0.75	10.45	1.31 (1.03)	1.35	0.95
(2,1)	20.23	1.11 (1.03)	1.25	0.65	10.10	1.30 (1.03)	1.35	0.95
(1,2)	41.34	0.75 (0.96)	1.15	0.15	28.09	0.94 (1.00)	1.15	0.45
(2,2)	40.63	0.75 (0.96)	1.05	0.15	26.96	0.94 (0.99)	1.15	0.45
initial screening age $t_0 = 60$, current age $t_{K_1} = 68$								
(1,1)	20.69	1.09 (1.03)	1.25	0.65	10.63	1.29 (1.02)	1.25	0.85
(2,1)	20.53	1.08 (1.02)	1.25	0.65	10.19	1.27 (1.02)	1.25	0.85
(1,2)	41.46	0.74 (0.95)	1.05	0.15	28.21	0.92 (0.99)	1.15	0.45
(2,2)	40.58	0.74 (0.94)	1.05	0.15	26.81	0.93 (0.98)	1.05	0.15
initial screening age $t_0 = 60$, current age $t_{K_1} = 72$								
(1,1)	21.15	1.06 (1.01)	1.15	0.65	10.88	1.25 (1.01)	1.25	0.85
(2,1)	20.93	1.05 (1.01)	1.15	0.65	10.32	1.24 (1.01)	1.25	0.65
(1,2)	41.58	0.72 (0.93)	1.05	0.15	28.36	0.90 (0.97)	1.05	0.15
(2,2)	40.50	0.72 (0.93)	1.05	0.15	26.61	0.91 (0.97)	1.05	0.15
initial screening age $t_0 = 60$, current age $t_{K_1} = 76$								
(1,1)	21.77	1.01 (0.99)	1.15	0.45	11.22	1.20 (1.00)	1.25	0.65
(2,1)	21.46	1.00 (0.98)	1.15	0.15	10.49	1.18 (0.99)	1.15	0.65
(1,2)	41.77	0.69 (0.91)	1.05	0.15	28.55	0.87 (0.95)	1.05	0.15
(2,2)	40.43	0.70 (0.91)	0.95	0.05	26.38	0.88 (0.94)	1.05	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 68$								
(1,1)	20.69	1.09 (1.03)	1.25	0.65	10.63	1.29 (1.02)	1.25	0.85
(2,1)	20.53	1.08 (1.02)	1.25	0.65	10.19	1.27 (1.02)	1.25	0.85
(1,2)	41.46	0.74 (0.95)	1.05	0.15	28.21	0.92 (0.99)	1.15	0.45
(2,2)	40.58	0.74 (0.94)	1.05	0.15	26.81	0.93 (0.98)	1.05	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 72$								
(1,1)	21.15	1.06 (1.01)	1.15	0.65	10.88	1.25 (1.01)	1.25	0.85
(2,1)	20.93	1.05 (1.01)	1.15	0.65	10.32	1.24 (1.01)	1.25	0.65
(1,2)	41.58	0.72 (0.93)	1.05	0.15	28.36	0.90 (0.97)	1.05	0.15
(2,2)	40.50	0.72 (0.93)	1.05	0.15	26.61	0.91 (0.97)	1.05	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 76$								
(1,1)	21.77	1.01 (0.99)	1.15	0.45	11.22	1.20 (1.00)	1.25	0.65
(2,1)	21.46	1.00 (0.98)	1.15	0.15	10.49	1.18 (0.99)	1.15	0.65
(1,2)	41.77	0.69 (0.91)	1.05	0.15	28.55	0.87 (0.95)	1.05	0.15
(2,2)	40.43	0.70 (0.91)	0.95	0.05	26.38	0.88 (0.94)	1.05	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 80$								
(1,1)	22.60	0.95 (0.96)	1.05	0.15	11.68	1.13 (0.97)	1.15	0.45
(2,1)	22.15	0.94 (0.95)	1.05	0.15	10.71	1.12 (0.96)	1.15	0.15
(1,2)	41.88	0.66 (0.88)	0.95	0.05	28.68	0.82 (0.92)	1.05	0.15
(2,2)	40.21	0.66 (0.88)	0.95	0.05	25.97	0.84 (0.92)	0.95	0.05

Table 4.3: A projection of the lead time distribution for individuals with screening history by current age and screening intervals, with MST=5

(Δ_1, Δ_2)	$\beta = 0.7$				$\beta = 0.9$			
(years)	P_0	EL (s.d.)	Median	Mode	P_0	EL (s.d.)	Median	Mode
initial screening age $t_0 = 56$, current age $t_{K_1} = 60$								
(1,1)	6.47	3.44 (2.50)	3.35	2.45	2.77	3.72 (2.48)	3.45	2.75
(2,1)	6.44	3.42 (2.50)	3.35	2.45	2.55	3.71 (2.47)	3.45	2.75
(1,2)	17.21	2.80 (2.49)	2.95	1.85	8.81	3.21 (2.48)	3.15	2.25
(2,2)	16.73	2.80 (2.49)	2.95	1.85	8.09	3.23 (2.47)	3.15	2.25
initial screening age $t_0 = 56$, current age $t_{K_1} = 64$								
(1,1)	6.95	3.31 (2.46)	3.25	2.15	3.00	3.59 (2.44)	3.35	2.45
(2,1)	6.87	3.29 (2.46)	3.15	2.15	2.72	3.58 (2.44)	3.35	2.45
(1,2)	17.93	2.69 (2.44)	2.85	1.45	9.31	3.09 (2.44)	3.05	1.95
(2,2)	17.28	2.70 (2.44)	2.85	1.45	8.38	3.12 (2.43)	3.05	1.95
initial screening age $t_0 = 56$, current age $t_{K_1} = 68$								
(1,1)	7.64	3.14 (2.41)	3.05	1.95	3.34	3.41 (2.39)	3.15	1.95
(2,1)	7.51	3.12 (2.41)	3.05	1.95	2.95	3.41 (2.39)	3.15	1.95
(1,2)	18.91	2.54 (2.38)	2.75	1.45	10.01	2.93 (2.38)	2.85	1.95
(2,2)	18.05	2.56 (2.38)	2.75	1.15	8.79	2.97 (2.38)	2.85	1.95
initial screening age $t_0 = 56$, current age $t_{K_1} = 72$								
(1,1)	8.58	2.93 (2.33)	2.85	1.65	3.80	3.18 (2.32)	2.95	1.95
(2,1)	8.37	2.91 (2.33)	2.85	1.65	3.28	3.19 (2.32)	2.95	1.95
(1,2)	20.19	2.36 (2.29)	2.55	0.15	10.93	2.73 (2.30)	2.65	1.45
(2,2)	19.06	2.39 (2.29)	2.55	0.15	9.32	2.78 (2.30)	2.75	0.15
initial screening age $t_0 = 60$, current age $t_{K_1} = 64$								
(1,1)	6.96	3.31 (2.46)	3.25	2.15	3.00	3.59 (2.44)	3.35	2.45
(2,1)	6.90	3.28 (2.46)	3.15	2.15	2.72	3.58 (2.44)	3.35	2.45
(1,2)	17.93	2.69 (2.44)	2.85	1.45	9.31	3.09 (2.44)	3.05	1.95
(2,2)	17.29	2.69 (2.44)	2.85	1.45	8.38	3.12 (2.43)	3.05	1.95
initial screening age $t_0 = 60$, current age $t_{K_1} = 68$								
(1,1)	7.64	3.14 (2.41)	3.05	1.95	3.34	3.41 (2.39)	3.15	1.95
(2,1)	7.51	3.12 (2.41)	3.05	1.95	2.95	3.41 (2.39)	3.15	1.95
(1,2)	18.91	2.54 (2.38)	2.75	1.45	10.01	2.93 (2.38)	2.85	1.95
(2,2)	18.05	2.56 (2.38)	2.75	1.15	8.79	2.97 (2.38)	2.85	1.95
initial screening age $t_0 = 60$, current age $t_{K_1} = 72$								
(1,1)	8.58	2.93 (2.33)	2.85	1.65	3.80	3.18 (2.32)	2.95	1.95
(2,1)	8.37	2.91 (2.33)	2.85	1.65	3.28	3.19 (2.32)	2.95	1.95
(1,2)	20.19	2.36 (2.29)	2.55	0.15	10.93	2.73 (2.30)	2.65	1.45
(2,2)	19.06	2.39 (2.29)	2.55	0.15	9.32	2.78 (2.30)	2.75	0.15
initial screening age $t_0 = 60$, current age $t_{K_1} = 76$								
(1,1)	9.84	2.65 (2.23)	2.55	0.15	4.42	2.90 (2.22)	2.65	1.45
(2,1)	9.50	2.64 (2.23)	2.55	0.15	3.71	2.92 (2.23)	2.65	0.15
(1,2)	21.85	2.14 (2.17)	2.35	0.15	12.13	2.47 (2.19)	2.45	0.15
(2,2)	20.35	2.17 (2.18)	2.35	0.05	10.01	2.55 (2.20)	2.45	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 68$								
(1,1)	7.65	3.14 (2.41)	3.05	1.95	3.34	3.41 (2.39)	3.15	1.95
(2,1)	7.54	3.11 (2.41)	3.05	1.95	2.95	3.41 (2.39)	3.15	1.95
(1,2)	18.91	2.54 (2.38)	2.75	1.45	10.01	2.93 (2.38)	2.85	1.95
(2,2)	18.06	2.55 (2.38)	2.75	1.15	8.78	2.97 (2.38)	2.85	1.95
initial screening age $t_0 = 64$, current age $t_{K_1} = 72$								
(1,1)	8.58	2.93 (2.33)	2.85	1.65	3.80	3.18 (2.32)	2.95	1.95
(2,1)	8.37	2.91 (2.33)	2.85	1.65	3.28	3.19 (2.32)	2.95	1.95
(1,2)	20.19	2.36 (2.29)	2.55	0.15	10.93	2.73 (2.30)	2.65	1.45
(2,2)	19.06	2.39 (2.29)	2.55	0.15	9.32	2.78 (2.30)	2.75	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 76$								
(1,1)	9.84	2.65 (2.23)	2.55	0.15	4.42	2.90 (2.22)	2.65	1.45
(2,1)	9.50	2.64 (2.23)	2.55	0.15	3.71	2.92 (2.23)	2.65	0.15
(1,2)	21.85	2.14 (2.17)	2.35	0.15	12.13	2.47 (2.19)	2.45	0.15
(2,2)	20.35	2.17 (2.18)	2.35	0.05	10.01	2.55 (2.20)	2.45	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 80$								
(1,1)	11.58	2.32 (2.08)	2.25	0.15	5.29	2.55 (2.08)	2.35	0.15
(2,1)	11.06	2.32 (2.09)	2.25	0.05	4.31	2.58 (2.09)	2.35	0.05
(1,2)	23.94	1.86 (2.01)	2.05	0.05	13.67	2.17 (2.04)	2.15	0.15
(2,2)	21.98	1.91 (2.02)	2.05	0.05	10.87	2.26 (2.05)	2.15	0.05

Table 4.4: A projection of the lead time distribution for individuals with screening history by current age and screening intervals, with MST=10

(Δ_1, Δ_2)	$\beta = 0.7$				$\beta = 0.9$			
(years)	P_0	EL (s.d.)	Median	Mode	P_0	EL (s.d.)	Median	Mode
initial screening age $t_0 = 56$, current age $t_{K_1} = 60$								
(1,1)	3.43	6.32 (4.41)	5.95	3.95	1.47	6.60 (4.39)	6.05	4.25
(2,1)	3.38	6.32 (4.44)	5.95	3.95	1.24	6.66 (4.41)	6.15	4.25
(1,2)	9.19	5.58 (4.42)	5.45	2.85	4.48	6.07 (4.39)	5.75	3.85
(2,2)	8.64	5.63 (4.44)	5.45	2.85	3.78	6.17 (4.41)	5.75	3.85
initial screening age $t_0 = 56$, current age $t_{K_1} = 64$								
(1,1)	4.00	5.87 (4.23)	5.45	3.65	1.73	6.13 (4.21)	5.65	3.75
(2,1)	3.82	5.92 (4.26)	5.55	3.65	1.44	6.22 (4.24)	5.65	3.75
(1,2)	10.31	5.15 (4.23)	5.05	2.65	5.14	5.61 (4.21)	5.25	2.65
(2,2)	9.52	5.25 (4.26)	5.15	2.65	4.24	5.75 (4.23)	5.35	3.25
initial screening age $t_0 = 56$, current age $t_{K_1} = 68$								
(1,1)	4.82	5.31 (4.01)	4.95	2.95	2.11	5.57 (3.99)	5.05	2.95
(2,1)	4.55	5.38 (4.04)	5.05	2.95	1.72	5.67 (4.02)	5.15	3.15
(1,2)	11.79	4.64 (3.98)	4.55	0.15	6.04	5.07 (3.97)	4.75	1.95
(2,2)	10.74	4.76 (4.03)	4.65	0.15	4.84	5.23 (4.01)	4.85	2.65
initial screening age $t_0 = 56$, current age $t_{K_1} = 72$								
(1,1)	5.92	4.68 (3.72)	4.35	0.15	2.63	4.92 (3.70)	4.45	2.15
(2,1)	5.54	4.76 (3.77)	4.45	0.15	2.09	5.04 (3.75)	4.55	0.15
(1,2)	13.70	4.05 (3.68)	4.05	0.15	7.23	4.45 (3.68)	4.15	0.15
(2,2)	12.31	4.20 (3.73)	4.15	0.15	5.63	4.63 (3.73)	4.25	0.15
initial screening age $t_0 = 60$, current age $t_{K_1} = 64$								
(1,1)	4.00	5.86 (4.23)	5.45	3.65	1.73	6.13 (4.21)	5.65	3.75
(2,1)	3.93	5.87 (4.27)	5.45	3.65	1.44	6.21 (4.24)	5.65	3.75
(1,2)	10.27	5.15 (4.23)	5.05	2.65	5.14	5.61 (4.21)	5.25	2.65
(2,2)	9.55	5.21 (4.26)	5.05	2.65	4.21	5.74 (4.23)	5.35	3.25
initial screening age $t_0 = 60$, current age $t_{K_1} = 68$								
(1,1)	4.82	5.31 (4.01)	4.95	2.95	2.11	5.57 (3.99)	5.05	2.95
(2,1)	4.56	5.38 (4.05)	4.95	2.95	1.72	5.67 (4.02)	5.15	3.15
(1,2)	11.79	4.64 (3.98)	4.55	0.15	6.04	5.07 (3.97)	4.75	1.95
(2,2)	10.74	4.76 (4.03)	4.65	0.15	4.84	5.23 (4.01)	4.85	2.65
initial screening age $t_0 = 60$, current age $t_{K_1} = 72$								
(1,1)	5.92	4.68 (3.72)	4.35	0.15	2.63	4.92 (3.70)	4.45	2.15
(2,1)	5.54	4.76 (3.77)	4.45	0.15	2.09	5.04 (3.75)	4.55	0.15
(1,2)	13.70	4.05 (3.68)	4.05	0.15	7.23	4.45 (3.68)	4.15	0.15
(2,2)	12.31	4.20 (3.73)	4.15	0.15	5.63	4.63 (3.73)	4.25	0.15
initial screening age $t_0 = 60$, current age $t_{K_1} = 76$								
(1,1)	7.38	3.99 (3.37)	3.65	0.15	3.33	4.21 (3.35)	3.75	0.15
(2,1)	6.86	4.08 (3.43)	3.75	0.05	2.59	4.34 (3.41)	3.85	0.15
(1,2)	16.11	3.42 (3.31)	3.45	0.15	8.76	3.78 (3.32)	3.55	0.15
(2,2)	14.29	3.57 (3.38)	3.55	0.05	6.62	3.98 (3.38)	3.65	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 68$								
(1,1)	4.82	5.31 (4.01)	4.95	2.95	2.11	5.57 (3.99)	5.05	2.95
(2,1)	4.70	5.32 (4.05)	4.95	0.15	1.71	5.66 (4.03)	5.15	3.15
(1,2)	11.73	4.64 (3.98)	4.55	0.15	6.04	5.07 (3.97)	4.75	1.95
(2,2)	10.76	4.72 (4.02)	4.55	0.15	4.80	5.22 (4.01)	4.85	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 72$								
(1,1)	5.92	4.68 (3.72)	4.35	0.15	2.63	4.92 (3.70)	4.45	2.15
(2,1)	5.55	4.76 (3.77)	4.45	0.15	2.09	5.04 (3.75)	4.55	0.15
(1,2)	13.70	4.05 (3.68)	4.05	0.15	7.23	4.45 (3.68)	4.15	0.15
(2,2)	12.31	4.19 (3.73)	4.15	0.15	5.63	4.63 (3.73)	4.25	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 76$								
(1,1)	7.38	3.99 (3.37)	3.65	0.15	3.33	4.21 (3.35)	3.75	0.15
(2,1)	6.86	4.08 (3.43)	3.75	0.05	2.59	4.34 (3.41)	3.85	0.15
(1,2)	16.11	3.42 (3.31)	3.45	0.15	8.76	3.78 (3.32)	3.55	0.15
(2,2)	14.30	3.57 (3.38)	3.55	0.05	6.62	3.98 (3.38)	3.65	0.15
initial screening age $t_0 = 64$, current age $t_{K_1} = 80$								
(1,1)	9.38	3.26 (2.97)	3.05	0.05	4.29	3.47 (2.96)	3.05	0.15
(2,1)	8.67	3.35 (3.03)	3.05	0.05	3.28	3.60 (3.02)	3.15	0.05
(1,2)	19.09	2.76 (2.89)	2.85	0.05	10.72	3.08 (2.91)	2.85	0.15
(2,2)	16.76	2.92 (2.97)	2.85	0.05	7.86	3.28 (2.99)	2.95	0.05

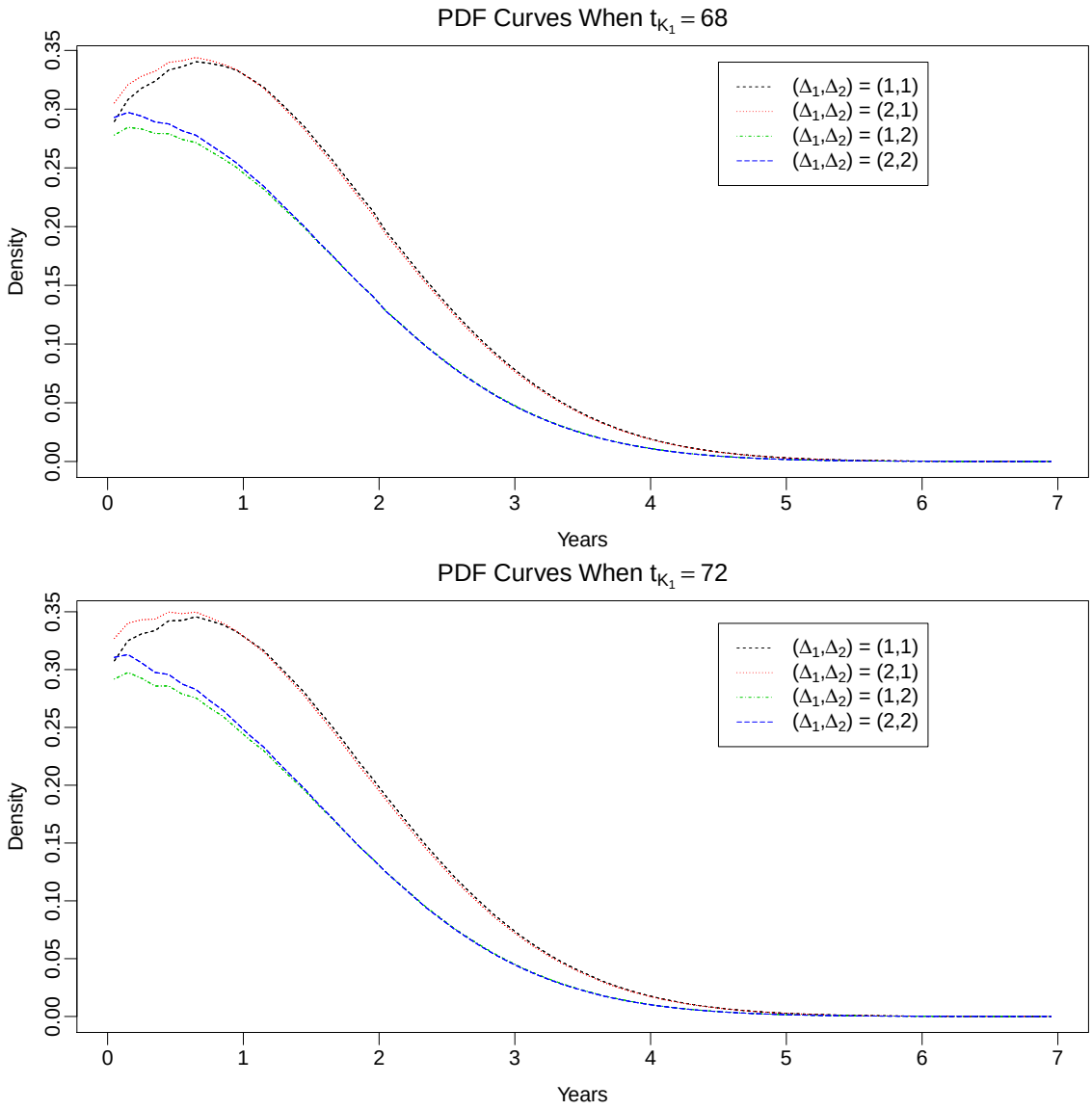

Figure 4.1: The PDF curves of the lead time for $t_{K_1} = 68$ and $t_{K_1} = 72$ with different t_0: 12 curves representing different screening schedules and different initial screening age t_0 are plotted for $t_{K_1} = 68$ (upper panel) and $t_{K_1} = 72$ (bottom panel), respectively. Curves with the same t_0 overlap each other, and only one curve for each t_0 shows. $\beta = 0.7$, MST $= 2$.

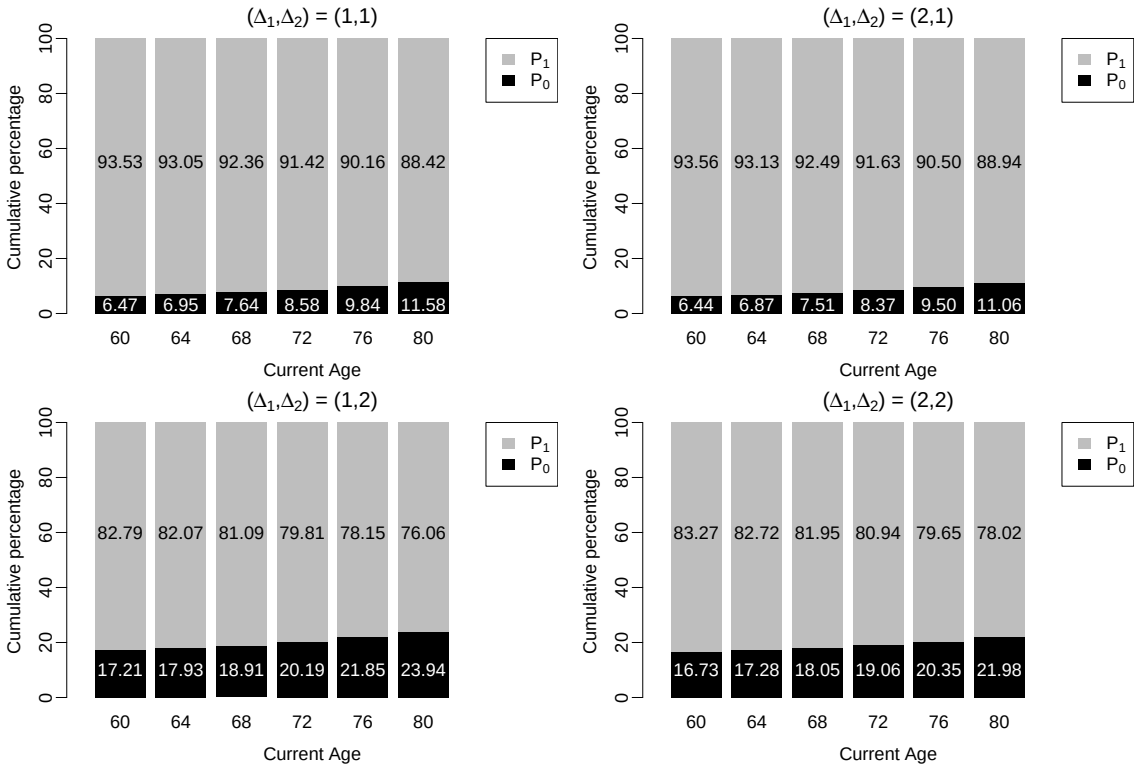

Figure 4.2: The bar plots of percentage changes for P_0 and P_1 with different t_{K_1}: Six bars representing different current ages are plotted for each of the four screening schedules, $(\Delta_1, \Delta_2) = (1, 1)$ (upper left panel), $(\Delta_1, \Delta_2) = (2, 1)$ (upper right panel), $(\Delta_1, \Delta_2) = (1, 2)$ (bottom left panel) and $(\Delta_1, \Delta_2) = (2, 2)$ (bottom right panel). $\beta = 0.7$, MST $= 5$, any t_0.

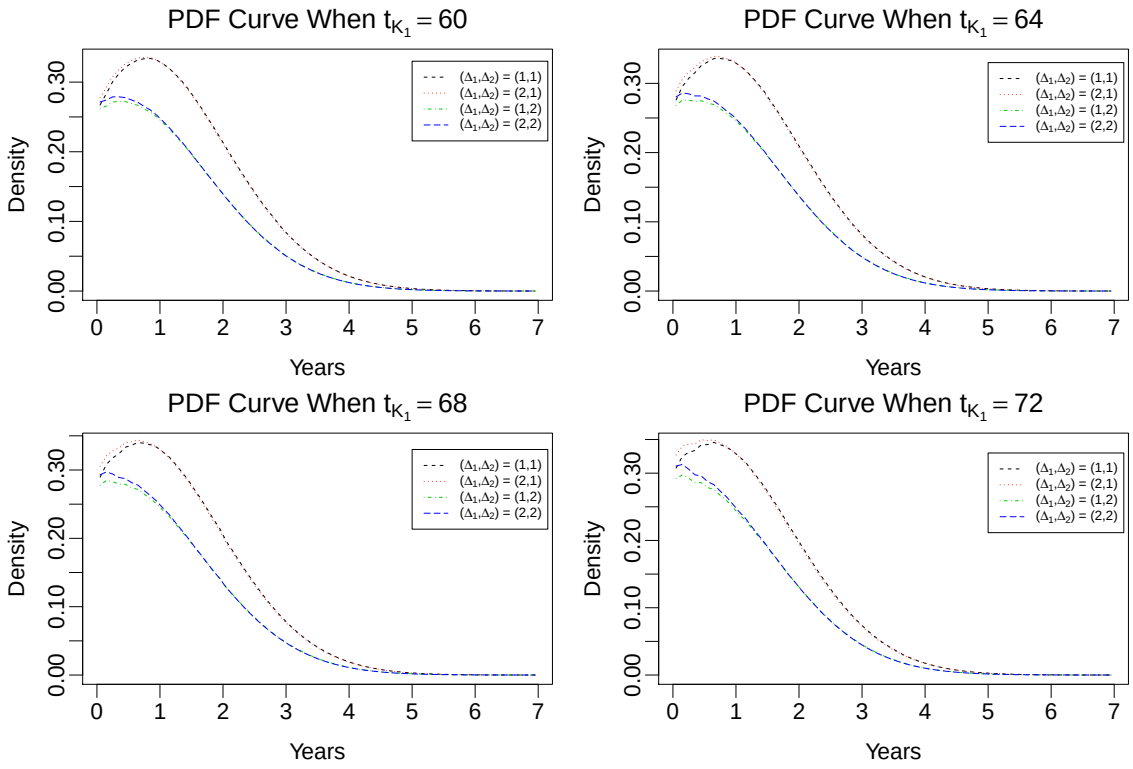

Figure 4.3: The sub-PDF curves of the lead time for $t_0 = 56$, $\beta = 0.7$, MST $= 2$

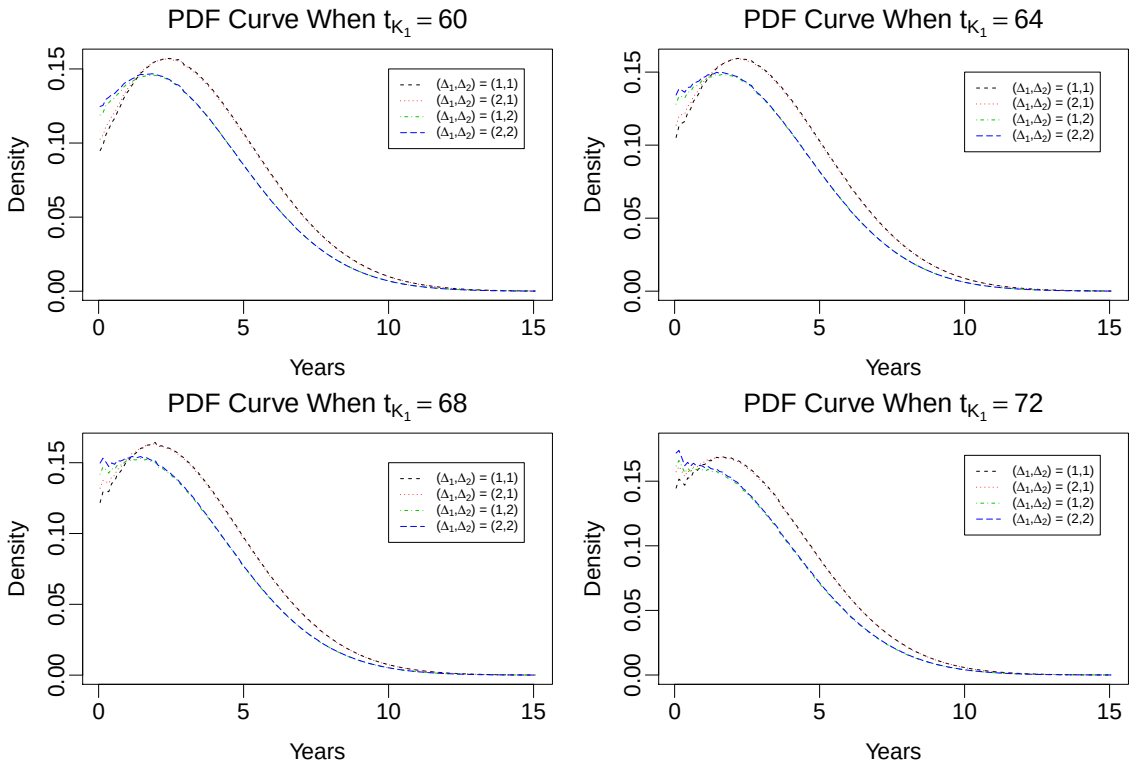

Figure 4.4: The sub-PDF curves of the lead time for $t_0 = 56$, $\beta = 0.7$, MST $= 5$

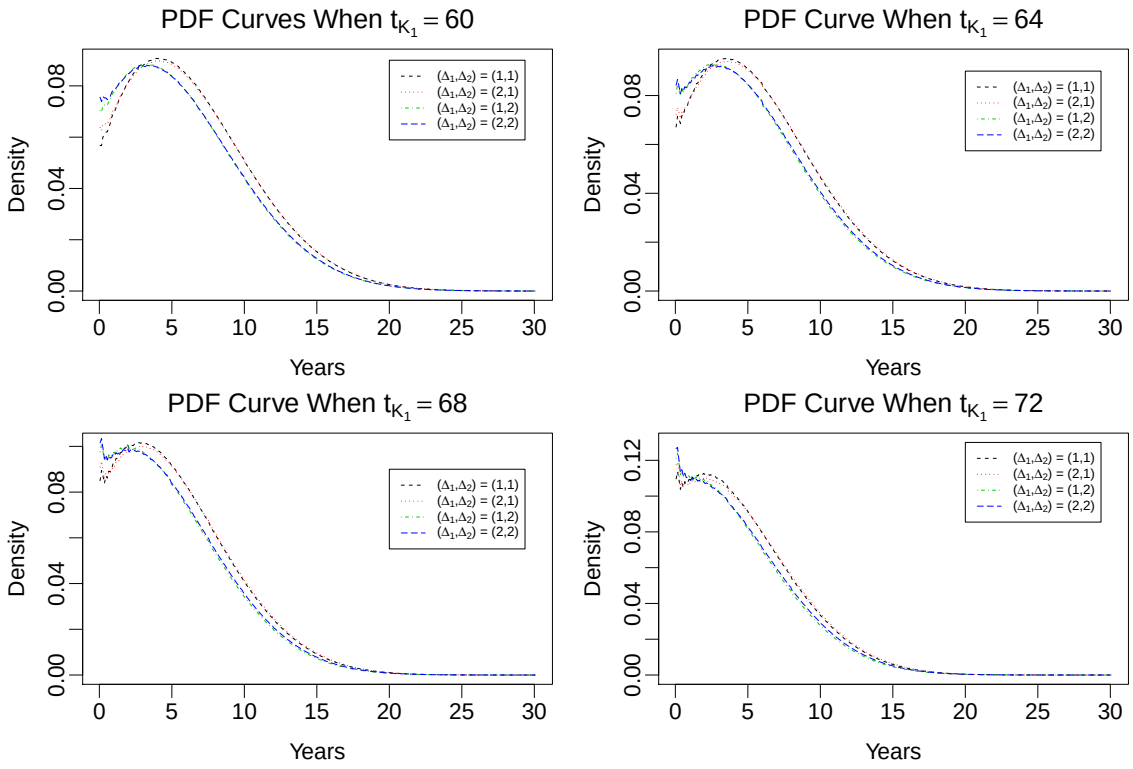

Figure 4.5: The sub-PDF curves of the lead time for $t_0 = 56$, $\beta = 0.7$, MST $= 10$

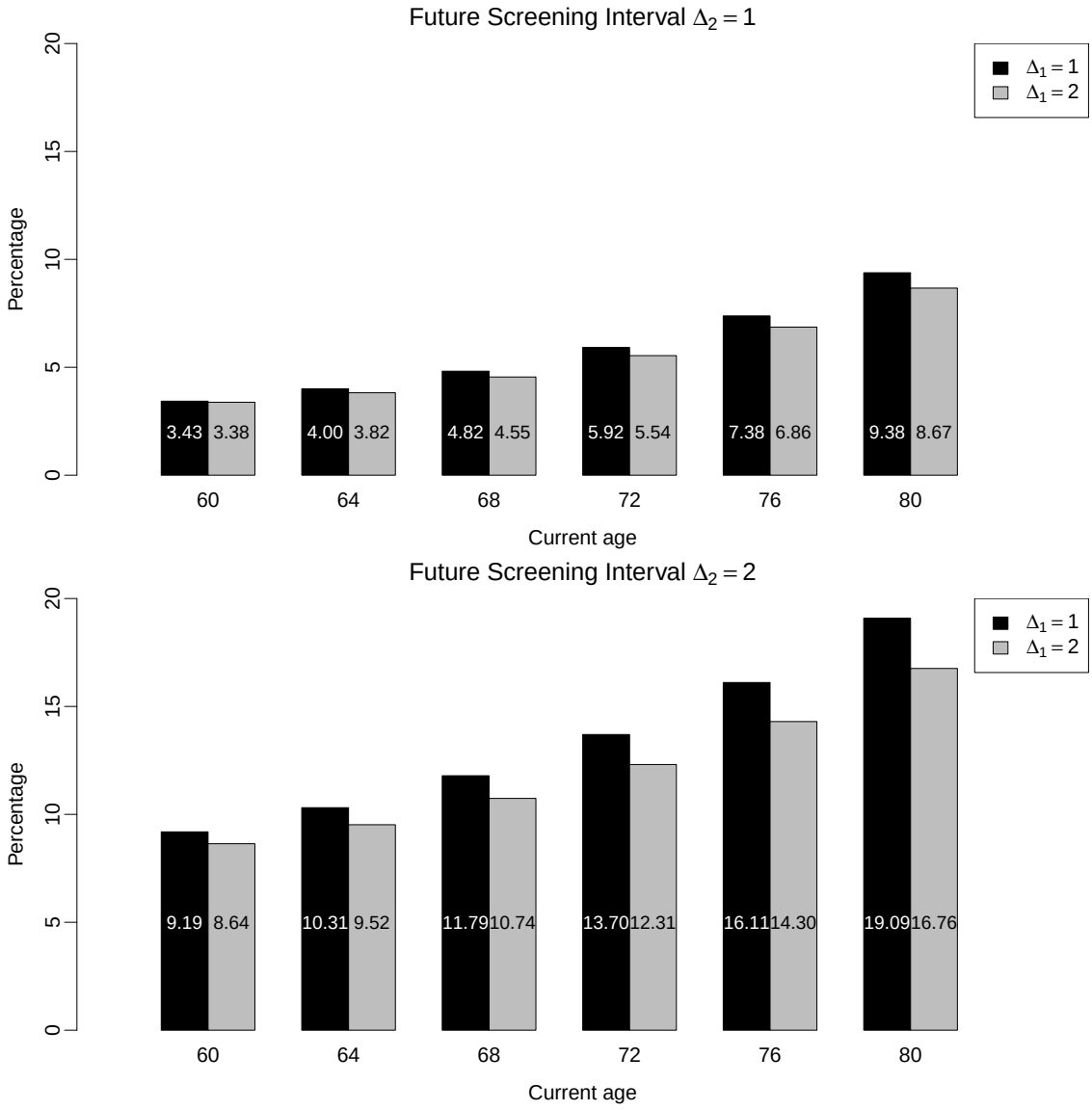

Figure 4.6: The bar plots of percentage changes for P_0 with different Δ_1 and the same Δ_2: Bars grouped by six different current ages are plotted for two future schedules, $\Delta_2 = 1$ (upper panel) and $\Delta_2 = 2$ (bottom panel). $\beta = 0.7$, MST $= 10$, any t_0.

4.3 Application

We then applied 1000 Bayesian posterior samples generated from the NLST data (Project 1) to the extended lead time distribution for individuals with screening history. In Project 2, we obtained the lead time distribution for 16 scenarios, and we followed the same strategy to obtain the projected lead time distribution for cohorts with different current ages and different past and future schedules.

For each gender, we assumed that there are four cohorts of initially asymptomatic individuals, with current age $t_{K_1} = 60$, 64, 68 and 72, respectively. Then within each cohort, we examined four different screening schedules $(\Delta_1, \Delta_2) = (1, 1)$, $(2, 1)$, $(1, 2)$ and $(2, 2)$, which are the same with the simulation study. We assumed the initial screening age $t_0 = 56$ for all scenarios, since the simulation study showed the initial screening age has no significant influence on the lead time distribution. We present our results of these 16 scenarios for both men and women.

Table 4.5 and Table 4.6 present the Bayesian predictive inference for the probability of no-early-detection (P_0), the probability of early-detection ($1 - P_0$) and the lead time in years for men and women, respectively. The probability that the lead time is zero (P_0) and its corresponding 95% C.I., the probability that the lead time is positive ($1-P_0$) and its corresponding standard deviation are reported as percentages. We also report median over the interquartile range (Med/IQR).

Coinciding with our expectations, for both genders, we can see the results of $(\Delta_1, \Delta_2) = (1, 1)$ and $(2, 1)$ are very similar, the same for results of $(\Delta_1, \Delta_2) = (1, 2)$ and $(2, 2)$. The future screening schedule plays an more important role than the past screening schedule regarding the lead time distribution. To compare with the lead time results for individuals with no history, we also add the projected lead time for individual with no history with corresponding future screening intervals. For example, in Table 4.5, the probability P_0 is 11.83% and the mean lead time is 0.87 years for

screening schedules $(\Delta_1, \Delta_2) = (1, 1)$, and P_0 is 11.67% and the mean lead time is 0.86 for $(\Delta_1, \Delta_2) = (2, 1)$ given the person's current age $t_{K_1} = 60$. The probability P_0 is 11.65% and the mean lead time is 0.86 years if the person's future screening interval is 1 year given the person took the first screening test at age 60. We also present the sub-PDF curves of lead time is positive for men and women in Figure 4.7 and Figure 4.8, respectively. In each figure, four panels represent four different current ages ($t_{K_1} = 60$, 65, 70 and 75). In each panel, six curves are the lead time density for six different screening intervals ($(\Delta_1, \Delta_2) = (1, 1)$, (1,2), (-,1), (2,1), (2,2) and (-,2)). The sub-PDF curves of lead time for the same future screening interval Δ_2 almost overlap each other given the same current age.

Like the simulation study, we also find a trend that larger Δ_1 will result in smaller P_0 if Δ_2 is the same. In Table 4.5, the probability P_0 is 11.83% for current age $t_{K_1} = 60$ with screening schedules $(\Delta_1, \Delta_2) = (1, 1)$, and it decreases to 11.67% if the individual's past screening interval $\Delta_1 = 2$.

For both genders, it is obvious that the probability P_0 increases and the mean lead time decreases as the future screening interval Δ_2 increases within the same age group. Across the age groups, the probability P_0 and mean lead time does not seem to have significant differences. To illustrate, let us see the lead time density curves of men and women in Figure 4.9 and Figure 4.10, respectively. Four panels represent four different screening schedules, and four curves represent four current ages in each panel. In each panel, the curves do not differentiate too much except the left tail.

In addition, the projected lead time is significantly affected by gender. Comparing to women, men have larger P_0 and shorter mean lead time given the same age and the same screening schedule. It seems that men have smaller chances to be detected early by CT screening exam than women do.

Table 4.5: A projection of the lead time distribution for men with screening history by current age and screening intervals with initial screening age $t_0 = 56$

(Δ_1, Δ_2) (years)	P_0 (95% C.I.)	$1 - P_0$ (s.d.)	EL (s.d.)	Med/IQR
current age $t_{K_1} = 60$				
(1,1)	11.83 (7.37, 17.93)	88.17 (2.66)	0.87 (0.69)	1.06
(2,1)	11.67 (7.28, 17.79)	88.33 (2.62)	0.86 (0.68)	0.94
(-,1)	11.65 (7.28, 17.76)	88.35 (2.61)	0.86 (0.68)	0.94
(1,2)	36.91 (28.21, 45.48)	63.09 (4.49)	0.54 (0.66)	0.83
(2,2)	36.30 (27.59, 44.99)	63.70 (4.53)	0.54 (0.66)	0.83
(-,2)	36.35 (27.67, 45.06)	63.65 (4.54)	0.55 (0.66)	0.83
current age $t_{K_1} = 64$				
(1,1)	11.86 (6.93, 19.12)	88.14 (3.04)	0.86 (0.68)	0.94
(2,1)	11.62 (6.82, 18.83)	88.38 (3.00)	0.85 (0.68)	0.94
(-,1)	11.58 (6.82, 18.77)	88.42 (2.99)	0.85 (0.68)	0.94
(1,2)	36.60 (27.69, 45.84)	63.40 (4.72)	0.55 (0.66)	0.83
(2,2)	35.68 (26.67, 45.19)	64.32 (4.78)	0.54 (0.65)	0.83
(-,2)	35.68 (26.42, 45.11)	64.32 (4.83)	0.55 (0.66)	0.83
current age $t_{K_1} = 68$				
(1,1)	12.00 (6.47, 20.89)	88.00 (3.62)	0.85 (0.68)	0.94
(2,1)	11.66 (6.42, 20.48)	88.34 (3.59)	0.84 (0.68)	0.94
(-,1)	11.61 (6.39, 20.48)	88.39 (3.58)	0.83 (0.67)	0.94
(1,2)	36.31 (27.09, 46.40)	63.69 (5.08)	0.54 (0.65)	0.83
(2,2)	35.01 (25.60, 45.55)	64.99 (5.18)	0.54 (0.65)	0.83
(-,2)	34.96 (25.23, 45.45)	65.04 (5.26)	0.54 (0.65)	0.94
current age $t_{K_1} = 72$				
(1,1)	12.29 (6.07, 23.67)	87.71 (4.43)	0.83 (0.68)	0.94
(2,1)	11.84 (5.96, 23.35)	88.16 (4.43)	0.81 (0.67)	0.94
(-,1)	11.77 (5.96, 23.33)	88.23 (4.43)	0.81 (0.67)	1.06
(1,2)	36.04 (26.18, 47.83)	63.96 (5.60)	0.54 (0.65)	0.83
(2,2)	34.34 (24.37, 46.75)	65.66 (5.78)	0.54 (0.64)	0.94
(-,2)	34.22 (23.88, 46.78)	65.78 (5.89)	0.54 (0.64)	0.81

Table 4.6: A projection of the lead time distribution for women with screening history by current age and screening intervals with initial screening age $t_0 = 56$

(Δ_1, Δ_2) (years)	P_0 (95% C.I.)	$1 - P_0$ (s.d.)	EL (s.d.)	Med/IQR
current age $t_{K_1} = 60$				
(1,1)	6.87 (3.94, 10.93)	93.13 (1.81)	1.06 (0.72)	1.17
(2,1)	6.78 (3.91, 10.73)	93.22 (1.77)	1.05 (0.72)	1.17
(-,1)	6.76 (3.91, 10.68)	93.24 (1.76)	1.05 (0.72)	1.17
(1,2)	28.69 (19.83, 38.26)	71.31 (4.64)	0.69 (0.73)	0.94
(2,2)	28.15 (19.39, 37.77)	71.85 (4.63)	0.69 (0.73)	0.94
(-,2)	28.26 (19.41, 37.85)	71.74 (4.65)	0.69 (0.73)	0.94
current age $t_{K_1} = 64$				
(1,1)	6.84 (3.82, 10.98)	93.16 (1.90)	1.05 (0.72)	1.17
(2,1)	6.69 (3.74, 10.68)	93.31 (1.85)	1.04 (0.72)	1.17
(-,1)	6.67 (3.74, 10.60)	93.33 (1.84)	1.04 (0.72)	1.05
(1,2)	28.44 (19.38, 37.99)	71.56 (4.65)	0.69 (0.73)	0.94
(2,2)	27.63 (18.76, 37.18)	72.37 (4.65)	0.69 (0.72)	0.94
(-,2)	27.69 (18.71, 37.29)	72.31 (4.69)	0.69 (0.72)	0.94
current age $t_{K_1} = 68$				
(1,1)	6.85 (3.70, 11.24)	93.15 (2.06)	1.04 (0.72)	1.17
(2,1)	6.64 (3.59, 10.99)	93.36 (2.00)	1.03 (0.72)	1.05
(-,1)	6.60 (3.59, 10.96)	93.40 (1.99)	1.02 (0.72)	1.17
(1,2)	28.20 (19.36, 37.59)	71.80 (4.68)	0.69 (0.72)	0.94
(2,2)	27.06 (18.24, 36.58)	72.94 (4.69)	0.69 (0.72)	0.94
(-,2)	27.06 (18.00, 36.61)	72.94 (4.77)	0.68 (0.72)	0.94
current age $t_{K_1} = 72$				
(1,1)	6.92 (3.57, 12.24)	93.08 (2.31)	1.03 (0.72)	1.17
(2,1)	6.63 (3.39, 11.78)	93.37 (2.26)	1.00 (0.72)	1.06
(-,1)	6.59 (3.39, 11.63)	93.41 (2.25)	1.00 (0.71)	1.06
(1,2)	27.97 (19.11, 37.41)	72.03 (4.76)	0.69 (0.72)	0.94
(2,2)	26.45 (17.74, 35.99)	73.55 (4.78)	0.68 (0.71)	0.94
(-,2)	26.39 (17.20, 36.03)	73.61 (4.90)	0.68 (0.71)	0.75

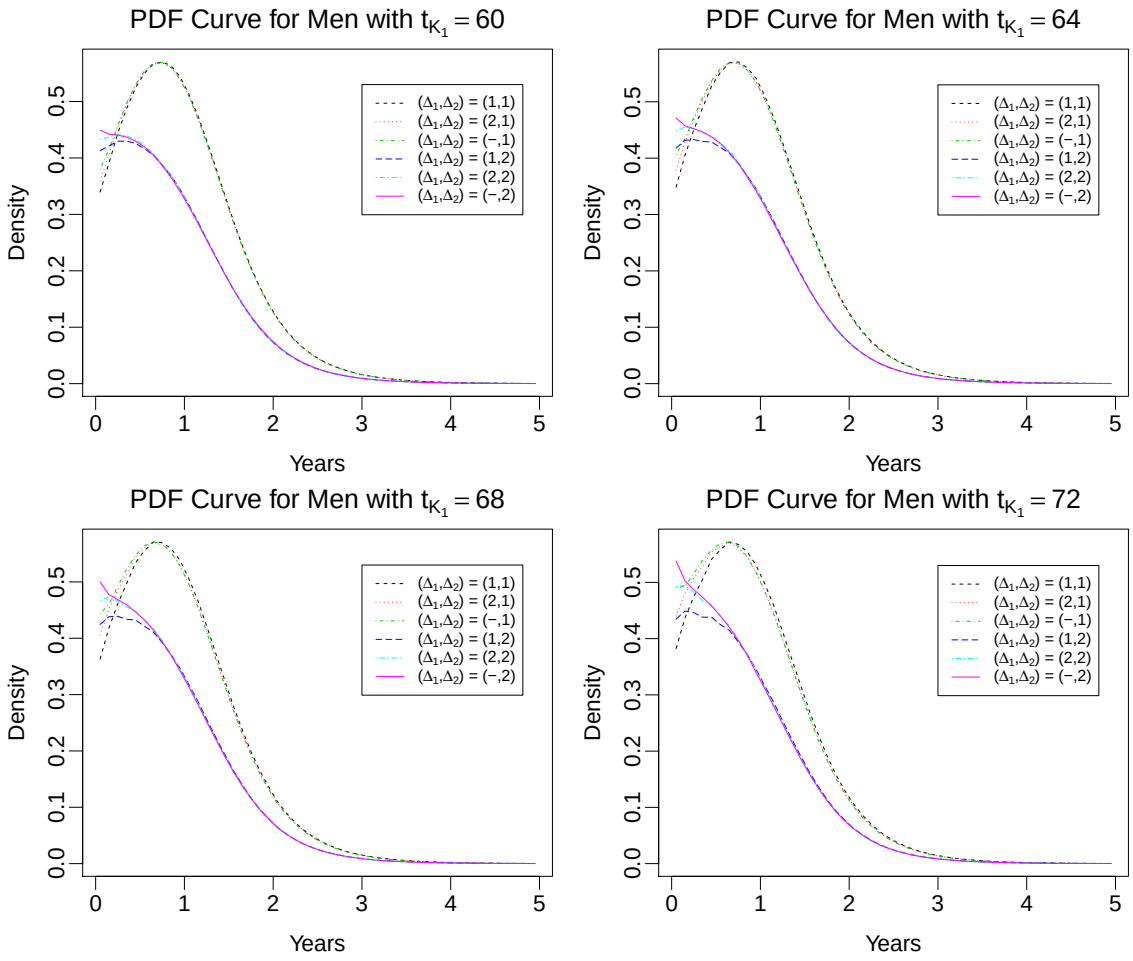

Figure 4.7: The sub-PDF curves of the lead time for men with screening history by screening intervals when $t_0 = 56$: Six curves representing different screening schedules are provided for each of the four current ages, $t_{K_1} = 60$ (upper left panel), $t_{K_1} = 64$ (upper right panel), $t_{K_1} = 68$ (bottom left panel) and $t_{K_1} = 72$ (bottom right panel).

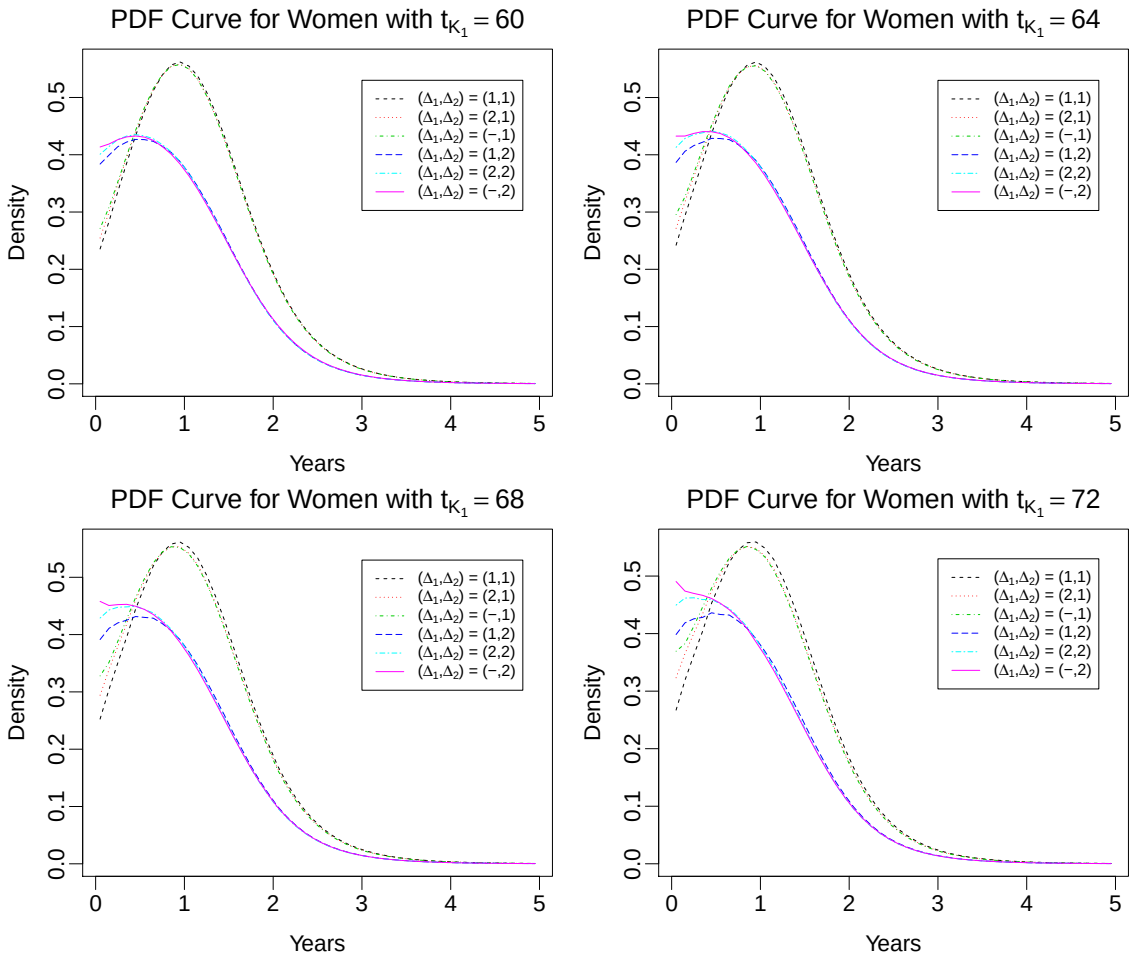

Figure 4.8: The sub-PDF curves of the lead time for women with screening history by screening intervals when $t_0 = 56$: Six curves representing different screening schedules are provided for each of the four current ages, $t_{K_1} = 60$ (upper left panel), $t_{K_1} = 64$ (upper right panel), $t_{K_1} = 68$ (bottom left panel) and $t_{K_1} = 72$ (bottom right panel).

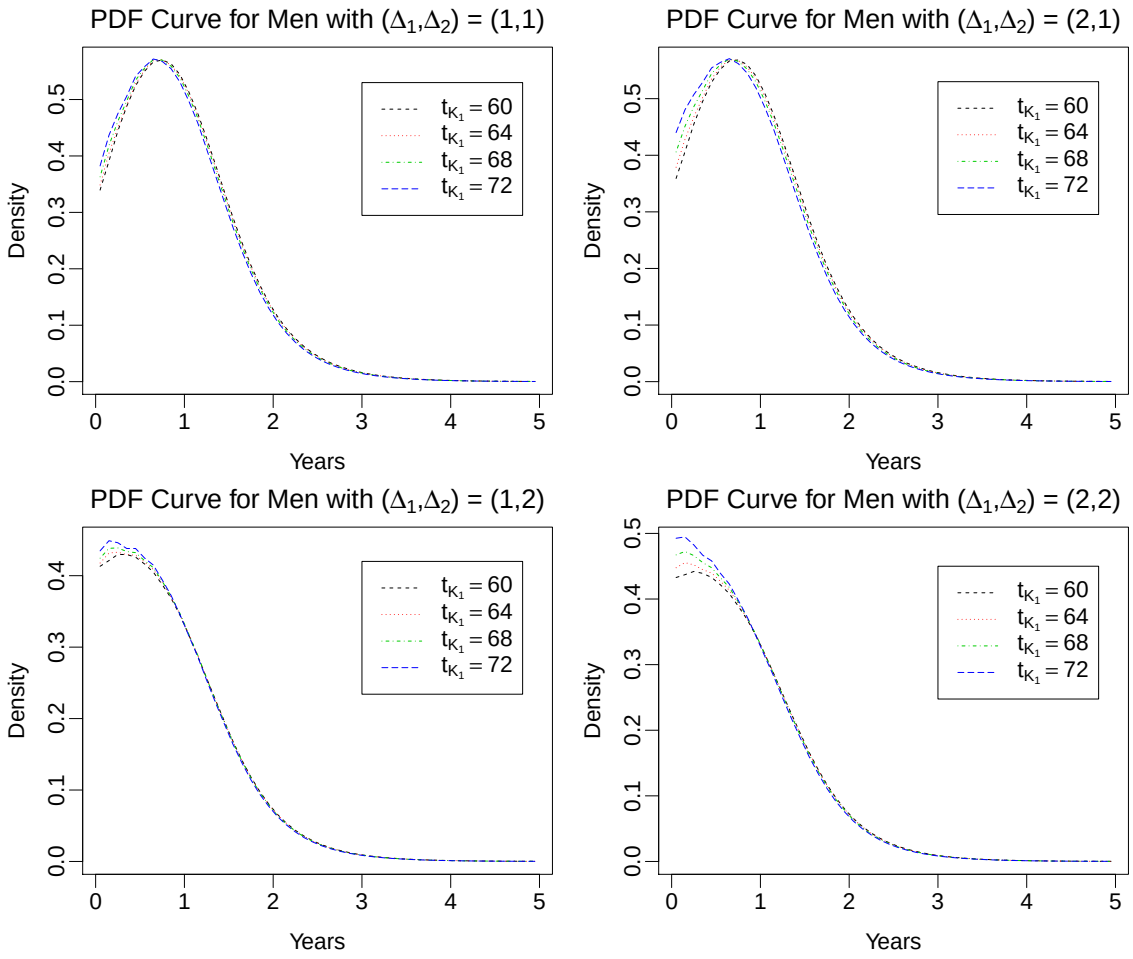

Figure 4.9: The sub-PDF curves of the lead time for men with screening history by current age when $t_0 = 56$: Four curves representing different current ages are provided for each of the four screening schedules, $(\Delta_1, \Delta_2) = (1, 1)$ (upper left panel), $(\Delta_1, \Delta_2) = (2, 1)$ (upper right panel), $(\Delta_1, \Delta_2) = (1, 2)$ (bottom left panel) and $(\Delta_1, \Delta_2) = (2, 2)$ (bottom right panel).

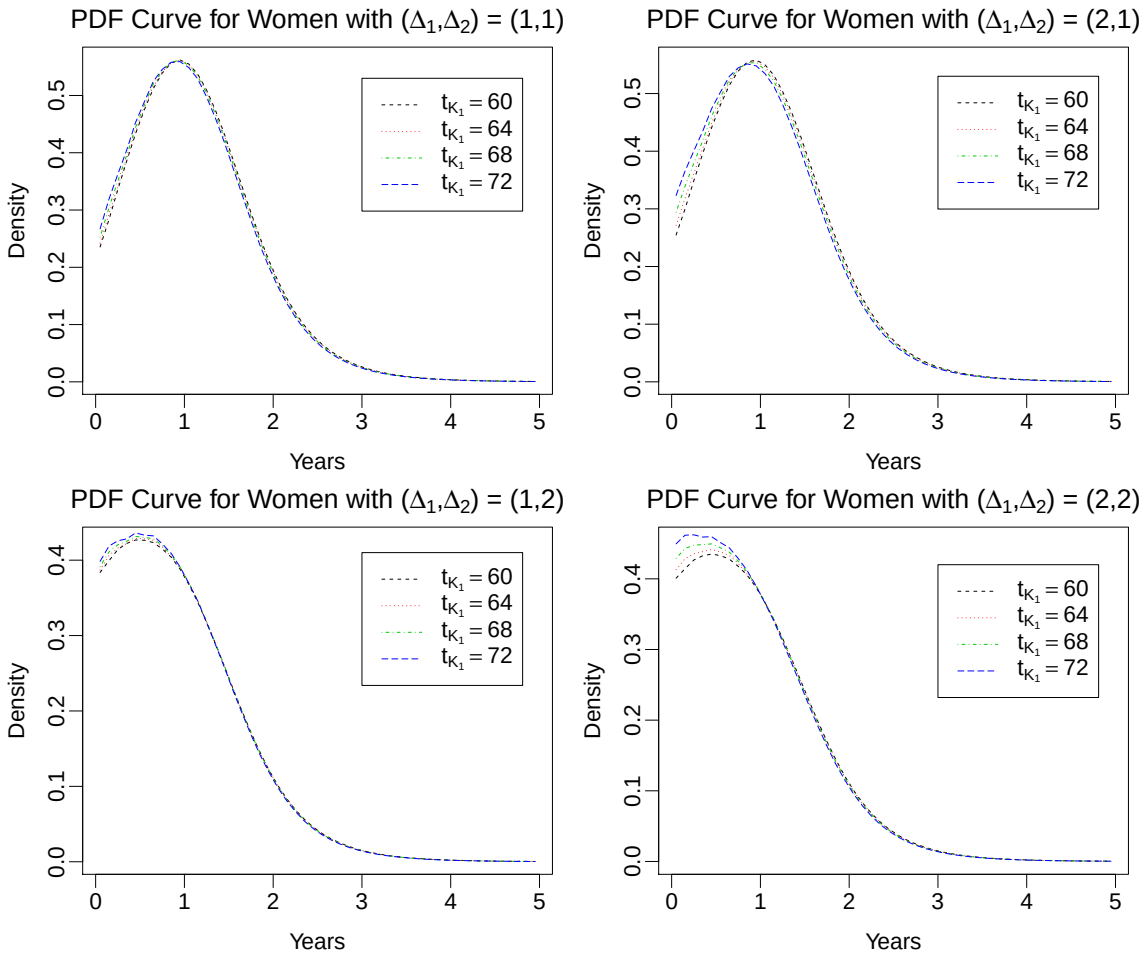

Figure 4.10: The sub-PDF curves of the lead time for women with screening history by current age when $t_0 = 56$: Four curves representing different current ages are provided for each of the four screening schedules, $(\Delta_1, \Delta_2) = (1, 1)$ (upper left panel), $(\Delta_1, \Delta_2) = (2, 1)$ (upper right panel), $(\Delta_1, \Delta_2) = (1, 2)$ (bottom left panel) and $(\Delta_1, \Delta_2) = (2, 2)$ (bottom right panel).

4.4 Discussion

In this project, we derived the lead time distribution for individuals with screening history. Simulation study was done to investigate the effect of a person's screening history on the lead time distribution. We also estimated the projected lead time distribution for different ages and screening schedules considering screening history using the NLST CT arm data.

In this simulation study, we found a small trend that larger past screening interval Δ_1 will result in larger chance of early-detection and longer mean lead time if the individual's future screening interval is decided. In the NLST application, we can also find this trend. But for a given current age, the length of screening history does not really affect the lead time distribution too much. This indicates the current age is more important to the lead time distribution than the initial screening age, since the person still looks healthy at current age.

However, the influence of a person's screening history on the lead time distribution is not as much as we thought. We may note that, the person who has lead time is under the assumption that he or she will develop clinical cancer during the lifetime. This fact may explain why the lead time distribution is more related to the current age. In addition, our model is relatively simple, the sensitivity is not age-dependent or depends on the sojourn time. The screening history may play an important role on lead time distribution when other model applied.

CHAPTER 5

FUTURE WORK

In this dissertation, we estimated the three key parameters of lung cancer screening using the NLST data, then used these parameters to make inference of the lead time distribution for individuals without or with screening history. In the future we will develop an R package to implement the three key parameters and lead time distribution estimation and make it available for other researchers.

As the inference of lead time and other terms (e.g. over-diagnosis, long term outcomes and etc.) all depend on the three key parameters, it is important to accurately estimate them. Another possible future plan would be to adjust the likelihood function and change parametric models of the three key parameters.

In addition, we can also explore the effect of a person's screening history on lead time distribution when the sensitivity depends on the sojourn time, or under other model assumptions.

REFERENCES

Aberle, D. R., DeMello, S., Berg, C. D., Black, W. C., Brewer, B., Church, T. R., Clingan, K. L., Duan, F., Fagerstrom, R. M., Gareen, I. F., Gatsonis, C. A., Gierada, D. S., Jain, A., Jones, G. C., Mahon, I., Marcus, P. M., Rathmell, J. M., and Sicks, J. (2013). Results of the Two Incidence Screenings in the National Lung Screening Trial. *New England Journal of Medicine*, 369(10):920–931.

Chen, Y., Erwin, D., and Wu, D. (2014). Over-diagnosis in Lung Cancer Screening using the MSKC-LCSP Data. *Journal of Biometrics and Biostatistics*, 05(04).

Jang, H., Kim, S., and Wu, D. (2013). Bayesian Lead Time Estimation for the Johns Hopkins Lung Project Data. *Journal of Epidemiology and Global Health*, 3(3):157 – 163.

Kafadar, K. and Prorok, P. C. (1994). A Data-Analytic Approach for Estimating Lead Time and Screening Benefit Based on Survival Curves in Randomized Cancer Screening Trials. *Statistics in Medicine*, 13(5-7):569–586.

Kafadar, K. and Prorok, P. C. (1996). Computer Simulation of Randomized Cancer Screening Trials to Compare Methods of Estimating Lead Time and Benefit Time. *Computational Statistics and Data Analysis*, 23(2):263 – 291.

Kafadar, K. and Prorok, P. C. (2003). Alternative Definitions of Comparable Case Groups and Estimates of Lead Time and Benefit Time in Randomized Cancer Screening Trials. *Statistics in Medicine*, 22(1):83–111.

Kendrick, S. K., Rai, S. N., and Wu, D. (2015). Simulation Study for the Sensitivity and Mean Sojourn Time Specific Lead Time in Cancer Screening When Human Lifetime is a Competing Risk. *Journal of Biometrics and Biostatistics*, 6(4).

Kim, S. and Erwin, D. (2012). Efficacy of Dual Lung Cancer Screening by Chest X-Ray and Sputum Cytology Using Johns Hopkins Lung Project Data. *Journal of Biometrics and Biostatistics*, 03(03).

Kim, S. and Wu, D. (2016). Estimation of Sensitivity Depending on Sojourn Time and Time Spent in Preclinical State. *Statistical Methods in Medical Research*, 25(2):728–740.

Liu, R., Gaskins, J., Mitra, R., and Wu, D. (2017). A Review of Estimation of Key Parameters and Lead Time in Cancer Screening. *Revista Colombiana de Estadstica*, 40(2):263–278.

Liu, R., Levitt, B., Riley, T., and Wu, D. (2015). Bayesian Estimation of the Three Key Parameters in CT for the National Lung Screening Trial Data. *Journal of Biometrics and Biostatistics*, 6(5).

Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E., and Adjei, A. A. (2008). Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clinic Proceedings*, 83(5):584 – 594.

NCI (2015). SEER Fast Stats Results. `http://seer.cancer.gov/statfacts/html/lungb.html` [Accessed: 2015-08-30].

Pinsky, P. F., Gierada, D. S., Black, W., Munden, R., Nath, H., Aberle, D., and Kazerooni, E. (2015). Performance of Lung-RADS in the National Lung Screening Trial. *Annals of Internal Medicine*, 162(7):485.

Prorok, P. C. (1976). The Theory of Periodic Screening I: Lead Time and Proportion Detected. *Advances in Applied Probability*, 8(1):127–143.

Prorok, P. C. (1982). Bounded Recurrence Times and Lead Time in the Design of a Repetitive Screening Program. *Journal of Applied Probability*, 19(1):10–19.

Shen, Y., Wu, D., and Zelen, M. (2001). Testing the Independence of Two Diagnostic Tests. *Biometrics*, 57(4):1009–1017.

Shen, Y. and Zelen, M. (1999). Parametric Estimation Procedures for Screening Programmes: Stable and Nonstable Disease Models for Multimodality Case Finding. *Biometrika*, 86(3):503–515.

SSA (2016). The Actuarial Life Table 2013. `https://www.ssa.gov/oact/STATS/table4c6.html` [Accessed: 2016-12-1].

Straatman, H., Peer, P. G. M., and Verbeek, A. L. M. (1997). Estimating Lead Time and Sensitivity in a Screening Program without Estimating the Incidence in the Screened Group. *Biometrics*, 53(1):217–229.

USPSTF (2016). The United States Preventive Services Task Force. `https://www.uspreventiveservicestaskforce.org/BrowseRec/Index` [Accessed: 2016-10-10].

Wu, D., Cariño, R. L., and Wu, X. (2008). When Sensitivity is a Function of Age and Time Spent in the Preclinical State in Periodic Cancer Screening. *Journal of Modern Applied Statistical Methods*, 7(1):297–303.

Wu, D., Erwin, D., and Rosner, G. L. (2009a). A Projection of Benefits Due to Fecal Occult Blood Test for Colorectal Cancer. *Cancer Epidemiology*, 33(3):212–215.

Wu, D., Erwin, D., and Rosner, G. L. (2009b). Estimating Key Parameters in FOBT Screening for Colorectal Cancer. *Cancer Causes and Control*, 20(1):41–46.

Wu, D., Erwin, D., and Rosner, G. L. (2011). Sojourn Time and Lead Time Projection in Lung Cancer Screening. *Lung Cancer*, 72(3):322– 326.

Wu, D., Kafadar, K., Rosner, G. L., and Broemeling, L. D. (2012). The Lead Time Distribution When Lifetime is Subject to Competing Risks in Cancer Screening. *The International Journal of Biostatistics*, 8(1).

Wu, D., Rosner, G. L., and Broemeling, L. (2005a). MLE and Bayesian Inference of Age-Dependent Sensitivity and Transition Probability in Periodic Screening. *Biometrics*, 61(4):1056–1063.

Wu, D., Rosner, G. L., and Broemeling, L. D. (2007). Bayesian Inference for the Lead Time in Periodic Cancer Screening. *Biometrics*, 63(3):873–880.

Wu, D., Wu, X., Banicescu, I., and Cariño, R. L. (2005b). Simulation Procedure in Periodic Cancer Screening Trials. *Journal of Modern Applied Statistical Methods*, 4(2):522–527.

Zelen, M. and Feinleib, M. (1969). On the Theory of Screening for Chronic Diseases. *Biometrika*, 56(3):601–614.

APPENDIX

The derivation of the conditional probabilities

$$
\begin{aligned}
&(1)\ P(D=1, H_{K_1}|T=t_{K_1+K}) \\
&= P(\text{entered } S_p \text{ in } (0,t_0), \text{ and not detected by first } K_1 \text{ exams}) \\
&\quad + \ P(\text{entered } S_p \text{ in } (t_0,t_1), \text{ and not detected by first } K_1-1 \text{ exams}) \\
&\quad + \cdots + P(\text{entered } S_p \text{ in } (t_{K_1-1}, t_{K_1})) + P(\text{entered } S_p \text{ after } t_{K_1}) \\
&= \int_{t_{K_1}}^{T}\int_{0}^{t_0} w(x)q(t-x)\,dx\,dt \cdot (1-\beta_0)\cdots(1-\beta_{K_1-1}) \\
&\quad + \int_{t_{K_1}}^{T}\int_{t_0}^{t_1} w(x)q(t-x)\,dx\,dt \cdot (1-\beta_1)\cdots(1-\beta_{K_1-1}) \\
&\quad + \cdots + \int_{t_{K_1}}^{T}\int_{t_{K_1-1}}^{t_{K_1}} w(x)q(t-x)\,dx\,dt + \int_{t_{K_1}}^{T}\int_{t_{K_1}}^{t} w(x)q(t-x)\,dx\,dt \\
&= (1-\beta_0)\cdots(1-\beta_{K_1-1})\int_{0}^{t_0} w(x)[Q(t_{K_1}-x)-Q(T-x)]\,dx \\
&\quad + (1-\beta_1)\cdots(1-\beta_{K_1-1})\int_{t_0}^{t_1} w(x)[Q(t_{K_1}-x)-Q(T-x)]\,dx \\
&\quad + \cdots + \int_{t_{K_1-1}}^{t_{K_1}} w(x)[Q(t_{K_1}-x)-Q(T-x)]\,dx + \int_{t_{K_1}}^{T} w(x)[1-Q(T-x)]\,dx \\
&= \sum_{i=0}^{K_1-1}(1-\beta_i)\cdots(1-\beta_{K_1-1})\int_{t_{i-1}}^{t_i} w(x)[Q(t_{K_1}-x)-Q(T-x)]\,dx \\
&\quad + \int_{t_{K_1-1}}^{t_{K_1}} w(x)[Q(t_{K_1}-x)-Q(T-x)]\,dx + \int_{t_{K_1}}^{T} w(x)[1-Q(T-x)]\,dx.
\end{aligned}
$$

(2) To calculate $P(L=0, D=1, H_{K_1}|T=t_{K_1+K})$, if a person's lead time is 0, it means this is an interval case, the disease shows symptoms between two screening exams.

Then we sum the probabilities of being interval cases after age t_{K_1}.

$$P(L=0, D=1, H_{K_1} | T = t_{K_1+K}) = I_{K_1+K,K_1+1} + I_{K_1+K,K_1+2} + \cdots + I_{K_1+K,K_1+K}. \quad \text{(A.1)}$$

The probability of being an interval case in (t_{K_1}, t_{K_1+1}) is

$$\begin{aligned}
I_{K_1+K,K_1+1} = \; & P(\text{entered } S_p \text{ in } (0,t_0), \text{ and be an interval case in } (t_{K_1}, t_{K_1+1})) \\
& + \; P(\text{entered } S_p \text{ in } (t_0,t_1), \text{ and be an interval case in } (t_{K_1}, t_{K_1+1})) \\
& + \cdots + P(\text{entered } S_p \text{ in } (t_{K_1-1},t_{K_1}), \text{ and be an interval case in } (t_{K_1}, t_{K_1+1})) \\
& + P(\text{entered } S_p \text{ after } t_{K_1}, \text{ and be an interval case in } (t_{K_1}, t_{K_1+1})) \\
= & \int_{t_{K_1}}^{t_{K_1+1}} \int_0^{t_0} w(x) q(t-x)\, dx\, dt \cdot (1-\beta_0) \cdots (1-\beta_{K_1}) \\
& + \int_{t_{K_1}}^{t_{K_1+1}} \int_{t_0}^{t_1} w(x) q(t-x)\, dx\, dt \cdot (1-\beta_1) \cdots (1-\beta_{K_1}) \\
& + \cdots + \int_{t_{K_1}}^{t_{K_1+1}} \int_{t_{K_1-1}}^{t_{K_1}} w(x) q(t-x)\, dx\, dt \cdot (1-\beta_{K_1}) + \int_{t_{K_1}}^{t_{K_1+1}} \int_{t_{K_1}}^{t} w(x) q(t-x)\, dx\, dt \\
= & (1-\beta_0) \cdots (1-\beta_{K_1}) \int_0^{t_0} w(x) [Q(t_{K_1}-x) - Q(t_{K_1+1}-x)]\, dx \\
& + (1-\beta_1) \cdots (1-\beta_{K_1}) \int_{t_0}^{t_1} w(x) [Q(t_{K_1}-x) - Q(t_{K_1+1}-x)]\, dx \\
& + \cdots + (1-\beta_{K_1}) \int_{t_{K_1-1}}^{t_{K_1}} w(x) [Q(t_{K_1}-x) - Q(t_{K_1+1}-x)]\, dx \\
& + \int_{t_{K_1}}^{t_{K_1+1}} w(x) [1 - Q(t_{K_1+1}-x)]\, dx \\
= & \sum_{i=0}^{K_1} (1-\beta_i) \cdots (1-\beta_{K_1}) \int_{t_{i-1}}^{t_i} w(x) [Q(t_{K_1}-x) - Q(t_{K_1+1}-x)]\, dx \\
& + \int_{t_{K_1}}^{t_{K_1+1}} w(x) [1 - Q(t_{K_1+1}-x)]\, dx.
\end{aligned}$$

Then we obtain

$$\begin{aligned}
I_{K_1+K,\, j} = & \sum_{i=0}^{j-1} (1-\beta_i) \cdots (1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x) [Q(t_{j-1}-x) - Q(t_j-x)]\, dx \\
& + \int_{t_{j-1}}^{t_j} w(x) [1 - Q(t_j - x)]\, dx, \text{ for all } j = K_1+1, \cdots, K_1+K.
\end{aligned} \quad \text{(A.2)}$$

(3) For $f_L(z, D = 1, H_{K_1}|T = t_{K_1+K})$, we consider it piecewisely. When $T - t_j < z \leq T - t_{j-1}$, the detection must occur at or before t_{j-1}, with $j = K_1 + 1, \cdots, K_1 + K$. When $T - t_{K_1+1} < z \leq T - t_{K_1}$, we have

$$
\begin{aligned}
f_L(z, D = 1, H_{K_1}|T = t_{K_1+K}) = \ & P(\text{entered } S_p \text{ in } (0, t_0), \text{ and detected at } t_{K_1}) \\
& + \ P(\text{entered } S_p \text{ in } (t_0, t_1), \text{ and detected at } t_{K_1}) \\
& + \cdots + \ P(\text{entered } S_p \text{ in } (t_{K_1-1}, t_{K_1}), \text{ and detected at } t_{K_1}) \\
= & \int_0^{t_0} w(x) q(t_{K_1} + z - x)\, dx \cdot (1 - \beta_0) \cdots (1 - \beta_{K_1-1}) \cdot \beta_{K_1} \\
& + \int_{t_0}^{t_1} w(x) q(t_{K_1} + z - x)\, dx \cdot (1 - \beta_1) \cdots (1 - \beta_{K_1-1}) \cdot \beta_{K_1} \\
& + \cdots + \int_{t_{K_1-1}}^{t_{K_1}} w(x) q(t_{K_1} + z - x)\, dx \cdot \beta_{K_1} \\
= & \ \beta_{K_1} \left\{ \sum_{i=0}^{K_1-1} (1 - \beta_i) \cdots (1 - \beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) q(t_{K_1} + z - x)\, dx \right. \\
& \left. + \int_{t_{K_1-1}}^{t_{K_1}} w(x) q(t_{K_1} + z - x)\, dx \right\}.
\end{aligned}
$$

In general, when $T - t_j < z \leq T - t_{j-1}$,

$$
\begin{aligned}
& f_L(z, D=1, H_{K_1} | T = t_{K_1+K}) \\
&= P(\text{detected at } t_{K_1}) + \ P(\text{detected at } t_{K_1+1}) + \cdots + \ P(\text{detected at } t_{j-1}) \\
&= \beta_{K_1} \left\{ \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) q(t_{K_1} + z - x)\, dx \right. \\
&\quad \left. + \int_{t_{K_1-1}}^{t_{K_1}} w(x) q(t_{K_1} + z - x)\, dx \right\} \\
&\quad + \beta_{K_1+1} \left\{ \sum_{i=0}^{K_1} (1-\beta_i) \cdots (1-\beta_{K_1}) \int_{t_{i-1}}^{t_i} w(x) q(t_{K_1+1} + z - x)\, dx \right. \\
&\quad \left. + \int_{t_{K_1}}^{t_{K_1+1}} w(x) q(t_{K_1+1} + z - x)\, dx \right\} \\
&\quad + \beta_{j-1} \left\{ \sum_{i=0}^{j-2} (1-\beta_i) \cdots (1-\beta_{j-2}) \int_{t_{i-1}}^{t_i} w(x) q(t_{j-1} + z - x)\, dx \right. \\
&\quad \left. + \int_{t_{j-2}}^{t_{j-1}} w(x) q(t_{j-2} + z - x)\, dx \right\} \\
&= \sum_{i=K_1}^{j-1} \beta_i \left\{ \sum_{r=0}^{i-1} (1-\beta_r) \cdots (1-\beta_{i-1}) \int_{t_{r-1}}^{t_r} w(x) q(t_i + z - x)\, dx \right. \\
&\quad \left. + \int_{t_{i-1}}^{t_i} w(x) q(t_i + z - x)\, dx \right\}.
\end{aligned}
\tag{A.3}
$$

The validity of the lead time distribution

To verify this is a valid probability distribution, we will prove that

$$
P(L=0 | D=1, H_{K_1}, T = t_{K_1+K}) + \int_0^{T-t_{K_1}} f_L(z | D=1, H_{K_1}, T = t_{K_1+K})\, dz \equiv 1, \tag{A.4}
$$

that is equivalent to prove

$$
\begin{aligned}
& P(L=0, D=1, H_{K_1} | T = t_{K_1+K}) + \int_0^{T-t_{K_1}} f_L(z, D=1, H_{K_1} | T = t_{K_1+K})\, dz \\
& \qquad \equiv P(D=1, H_{K_1} | T = t_{K_1+K}).
\end{aligned}
$$

Proof:

$$
\begin{aligned}
&P(L=0, D=1, H_{K_1}|T=t_{K_1+K}) \\
&= I_{K_1+K,K_1+1} + I_{K_1+K,K_1+2} + \cdots + I_{K_1+K,K_1+K} = \sum_{j=K_1+1}^{K_1+K} I_{K_1+K,\ j} \\
&= \sum_{j=K_1+1}^{K_1+K} \left\{ \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)[Q(t_{j-1}-x) - Q(t_j-x)]\,dx \right. \\
&\quad \left. + \int_{t_{j-1}}^{t_j} w(x)[1-Q(t_j-x)]\,dx \right\} \\
&= \sum_{j=K_1+1}^{K_1+K} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)[Q(t_{j-1}-x) - Q(t_j-x)]\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K} \int_{t_{j-1}}^{t_j} w(x)[1-Q(t_j-x)]\,dx \\
&= \sum_{j=K_1+1}^{K_1+K} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{j-1}-x)\,dx \\
&\quad - \sum_{j=K_1+1}^{K_1+K} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K} \int_{t_{j-1}}^{t_j} w(x)\,dx - \sum_{j=K_1+1}^{K_1+K} \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
&= \sum_{j=K_1+1}^{K_1+K} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{j-1}-x)\,dx \\
&\quad - \sum_{j=K_1+1}^{K_1+K} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx + \int_{t_{K_1}}^{T} w(x)\,dx \\
&\quad - \sum_{j=K_1+1}^{K_1+K} \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
&= \boldsymbol{A - B + C - D}.
\end{aligned}
$$

For the integration of the pdf,

$$
\begin{aligned}
&\int_0^{T-t_{K_1}} f_L(z, D=1, H_{K_1} | T = t_{K_1+K})\, dz \\
&= \sum_{j=K_1+1}^{K_1+K} \int_{T-t_j}^{T-t_{j-1}} f_L(z, D=1, H_{K_1} | T = t_{K_1+K})\, dz \\
&= \sum_{j=K_1+1}^{K_1+K} \int_{T-t_j}^{T-t_{j-1}} \sum_{i=K_1}^{j-1} \beta_i \left\{ \sum_{r=0}^{i-1} (1-\beta_r) \cdots (1-\beta_{i-1}) \int_{t_{r-1}}^{t_r} w(x) q(t_i + z - x)\, dx \right. \\
&\quad \left. + \int_{t_{i-1}}^{t_i} w(x) q(t_i + z - x)\, dx \right\} dz \\
&= \sum_{j=K_1+1}^{K_1+K} \sum_{i=K_1}^{j-1} \beta_i \left\{ \sum_{r=0}^{i-1} (1-\beta_r) \cdots (1-\beta_{i-1}) \int_{t_{r-1}}^{t_r} w(x) \int_{T-t_j}^{T-t_{j-1}} q(t_i + z - x)\, dz\, dx \right. \\
&\quad \left. + \int_{t_{i-1}}^{t_i} w(x) \int_{T-t_j}^{T-t_{j-1}} q(t_i + z - x)\, dz\, dx \right\} \quad \text{(swap } dx \text{ and } dz\text{)} \\
&= \sum_{j=K_1+1}^{K_1+K} \sum_{i=K_1}^{j-1} \beta_i \\
&\quad \left\{ \sum_{r=0}^{i-1} (1-\beta_r) \cdots (1-\beta_{i-1}) \int_{t_{r-1}}^{t_r} w(x) [Q(T - t_j + t_i - x) - Q(T - t_{j-1} + t_i - x)]\, dx \right. \\
&\quad \left. + \int_{t_{i-1}}^{t_i} w(x) [Q(T - t_j + t_i - x) - Q(T - t_{j-1} + t_i - x)]\, dx \right\} \\
&= \sum_{i=K_1}^{K_1+K-1} \sum_{j=i+1}^{K_1+K} \beta_i \\
&\quad \left\{ \sum_{r=0}^{i-1} (1-\beta_r) \cdots (1-\beta_{i-1}) \int_{t_{r-1}}^{t_r} w(x) [Q(T - t_j + t_i - x) - Q(T - t_{j-1} + t_i - x)]\, dx \right. \\
&\quad \left. + \int_{t_{i-1}}^{t_i} w(x) [Q(T - t_j + t_i - x) - Q(T - t_{j-1} + t_i - x)]\, dx \right\} \quad \text{(swap sum of } i \text{ and } j\text{)} \\
&= \sum_{i=K_1}^{K_1+K-1} \beta_i \sum_{r=0}^{i-1} (1-\beta_r) \cdots (1-\beta_{i-1}) \times \\
&\quad \sum_{j=i+1}^{K_1+K} \int_{t_{r-1}}^{t_r} w(x) [Q(T - t_j + t_i - x) - Q(T - t_{j-1} + t_i - x)]\, dx \\
&\quad + \sum_{i=K_1}^{K_1+K-1} \beta_i \sum_{j=i+1}^{K_1+K} \int_{t_{i-1}}^{t_i} w(x) [Q(T - t_j + t_i - x) - Q(T - t_{j-1} + t_i - x)]\, dx
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=K_1}^{K_1+K-1} \beta_i \sum_{r=0}^{i-1} (1-\beta_r)\cdots(1-\beta_{i-1}) \times \left\{ \sum_{j=i+1}^{K_1+K} \int_{t_{r-1}}^{t_r} w(x) Q(T - t_j + t_i - x)\, dx \right. \\
&\quad \left. - \sum_{j=i+1}^{K_1+K} \int_{t_{r-1}}^{t_r} w(x) Q(T - t_{j-1} + t_i - x)\, dx \right\} \\
&\quad + \sum_{i=K_1}^{K_1+K-1} \beta_i \times \\
&\quad \left\{ \sum_{j=i+1}^{K_1+K} \int_{t_{i-1}}^{t_i} w(x) Q(T - t_j + t_i - x)\, dx - \sum_{j=i+1}^{K_1+K} \int_{t_{i-1}}^{t_i} w(x) Q(T - t_{j-1} + t_i - x)]\, dx \right\} \\
&= \sum_{i=K_1}^{K_1+K-1} \beta_i \sum_{r=0}^{i-1} (1-\beta_r)\cdots(1-\beta_{i-1}) \times \left\{ \sum_{j=i+1}^{K_1+K} \int_{t_{r-1}}^{t_r} w(x) Q(T - t_j + t_i - x)\, dx \right. \\
&\quad \left. - \sum_{l=i}^{K_1+K-1} \int_{t_{r-1}}^{t_r} w(x) Q(T - t_l + t_i - x)\, dx \right\} \quad \text{(change index } j-1 \to l\text{)} \\
&\quad + \sum_{i=K_1}^{K_1+K-1} \beta_i \times \left\{ \sum_{j=i+1}^{K_1+K} \int_{t_{i-1}}^{t_i} w(x) Q(T - t_j + t_i - x)\, dx \right. \\
&\quad \left. - \sum_{l=i}^{K_1+K-1} \int_{t_{i-1}}^{t_i} w(x) Q(T - t_l + t_i - x)\, dx \right\} \quad \text{(change index } j-1 \to l\text{)} \\
&= \sum_{i=K_1}^{K_1+K-1} \beta_i \sum_{r=0}^{i-1} (1-\beta_r)\cdots(1-\beta_{i-1}) \left\{ \int_{t_{r-1}}^{t_r} w(x) Q(t_i - x)\, dx - \int_{t_{r-1}}^{t_r} w(x) Q(T-x)\, dx \right\} \\
&\quad + \sum_{i=K_1}^{K_1+K-1} \beta_i \left\{ \int_{t_{i-1}}^{t_i} w(x) Q(t_i - x)\, dx - \int_{t_{i-1}}^{t_i} w(x) Q(T-x)\, dx \right\} \\
&= \sum_{j=K_1}^{K_1+K-1} \beta_j \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \left\{ \int_{t_{i-1}}^{t_i} w(x) Q(t_j - x)\, dx \right. \\
&\quad \left. - \int_{t_{i-1}}^{t_i} w(x) Q(T-x)\, dx \right\} \quad \text{(only } i \text{ and } r \text{ left, change index } i \to j,\ r \to i\text{)} \\
&\quad + \sum_{i=K_1}^{K_1+K-1} \beta_i \left\{ \int_{t_{i-1}}^{t_i} w(x) Q(t_i - x)\, dx - \int_{t_{i-1}}^{t_i} w(x) Q(T-x)\, dx \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j=K_1}^{K_1+K-1} \beta_j \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j - x)\,dx \\
&\quad - \sum_{j=K_1}^{K_1+K-1} \beta_j \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T - x)\,dx \\
&\quad + \sum_{i=K_1}^{K_1+K-1} \beta_i \int_{t_{i-1}}^{t_i} w(x)Q(t_i - x)\,dx - \sum_{i=K_1}^{K_1+K-1} \beta_i \int_{t_{i-1}}^{t_i} w(x)Q(T - x)\,dx \\
&= \boldsymbol{E - F + G - H}.
\end{aligned}
$$

Compare to $P(L=0, D=1, H_{K_1}|T=t_{K_1+K}) = \boldsymbol{A}-\boldsymbol{B}+\boldsymbol{C}-\boldsymbol{D}$, we have

$$
\begin{aligned}
\boldsymbol{E}-\boldsymbol{B} &= \sum_{j=K_1}^{K_1+K-1} \beta_j \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&\quad - \sum_{j=K_1+1}^{K_1+K} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&= \beta_{K_1} \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} \beta_j \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&\quad - \sum_{j=K_1+1}^{K_1+K-1} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&\quad - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \beta_{K_1} \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} (\beta_j-1) \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&\quad - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \beta_{K_1} \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad - \sum_{j=K_1+1}^{K_1+K-1} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1})(1-\beta_j) \int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&\quad - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\boldsymbol{A}-\boldsymbol{B}+\boldsymbol{E} &= \sum_{j=K_1+1}^{K_1+K}\sum_{i=0}^{j-1}(1-\beta_i)\cdots(1-\beta_{j-1})\int_{t_{i-1}}^{t_i} w(x)Q(t_{j-1}-x)\,dx \\
&+ \beta_{K_1}\sum_{i=0}^{K_1-1}(1-\beta_i)\cdots(1-\beta_{K_1-1})\int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&- \sum_{j=K_1+1}^{K_1+K-1}\sum_{i=0}^{j-1}(1-\beta_i)\cdots(1-\beta_{j-1})(1-\beta_j)\int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&- \sum_{i=0}^{K_1+K-1}(1-\beta_i)\cdots(1-\beta_{K_1+K-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{l=K_1}^{K_1+K-1}\sum_{i=0}^{l}(1-\beta_i)\cdots(1-\beta_l)\int_{t_{i-1}}^{t_i} w(x)Q(t_l-x)\,dx \quad (\text{change index } j-1\to l) \\
&+ \beta_{K_1}\sum_{i=0}^{K_1-1}(1-\beta_i)\cdots(1-\beta_{K_1-1})\int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&- \sum_{j=K_1+1}^{K_1+K-1}\sum_{i=0}^{j-1}(1-\beta_i)\cdots(1-\beta_{j-1})(1-\beta_j)\int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&- \sum_{i=0}^{K_1+K-1}(1-\beta_i)\cdots(1-\beta_{K_1+K-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{i=0}^{K_1}(1-\beta_i)\cdots(1-\beta_{K_1})\int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&+ \sum_{l=K_1+1}^{K_1+K-1}\sum_{i=0}^{l}(1-\beta_i)\cdots(1-\beta_l)\int_{t_{i-1}}^{t_i} w(x)Q(t_l-x)\,dx \\
&+ \beta_{K_1}\sum_{i=0}^{K_1-1}(1-\beta_i)\cdots(1-\beta_{K_1-1})\int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&- \sum_{j=K_1+1}^{K_1+K-1}\sum_{i=0}^{j-1}(1-\beta_i)\cdots(1-\beta_{j-1})(1-\beta_j)\int_{t_{i-1}}^{t_i} w(x)Q(t_j-x)\,dx \\
&- \sum_{i=0}^{K_1+K-1}(1-\beta_i)\cdots(1-\beta_{K_1+K-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=0}^{K_1} (1-\beta_i)\cdots(1-\beta_{K_1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \ \beta_{K_1} \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \quad \text{(change index back } l \to j\text{)} \\
&\quad - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= (1-\beta_{K_1}) \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \ (1-\beta_{K_1}) \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \ \beta_{K_1} \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
&\quad - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= (1-\beta_{K_1}) \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
&\quad - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\boldsymbol{A-B+E+G} = {} & (1-\beta_{K_1}) \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx \\
& + \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
& + \sum_{j=K_1+1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
& - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
& + \sum_{i=K_1}^{K_1+K-1} \beta_i \int_{t_{i-1}}^{t_i} w(x)Q(t_i-x)\,dx \\
= {} & (1-\beta_{K_1}) \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx \\
& + \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
& + \sum_{j=K_1+1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
& - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
& + \sum_{j=K_1+1}^{K_1+K-1} \beta_j \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx + \beta_{K_1} \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx \,(\text{change index } i \to j) \\
= {} & \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx + \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
& + \sum_{j=K_1+1}^{K_1+K-1} \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
& - \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx.
\end{aligned}
$$

We also have

$$
\begin{aligned}
-\boldsymbol{F} &= -\sum_{j=K_1}^{K_1+K-1} \beta_j \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{j=K_1}^{K_1+K-1} (-\beta_j) \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{j=K_1}^{K_1+K-1} (1-\beta_j-1) \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{j=K_1}^{K_1+K-1} (1-\beta_j) \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{j=K_1}^{K_1+K-1} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{j=K_1}^{K_1+K-1} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1})(1-\beta_j) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{j=K_1}^{K_1+K-1} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{j=K_1}^{K_1+K-1} \left\{ \sum_{i=0}^{j} (1-\beta_i)\cdots(1-\beta_j) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx - (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(T-x)\,dx \right\} \\
&\quad - \sum_{j=K_1}^{K_1+K-1} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{j=K_1}^{K_1+K-1} \sum_{i=0}^{j} (1-\beta_i)\cdots(1-\beta_j) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{j=K_1}^{K_1+K-1} \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{j=K_1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(T-x)\,dx
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j=K_1}^{K_1+K-1} \sum_{i=0}^{j} (1-\beta_i)\cdots(1-\beta_j) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{l=K_1-1}^{K_1+K-2} \sum_{i=0}^{l} (1-\beta_i)\cdots(1-\beta_l) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \quad \text{(change index } j-1 \to l\text{)} \\
&\quad - \sum_{j=K_1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(T-x)\,dx \\
&= \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad + \sum_{j=K_1}^{K_1+K-2} \sum_{i=0}^{j} (1-\beta_i)\cdots(1-\beta_j) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{l=K_1}^{K_1+K-2} \sum_{i=0}^{l} (1-\beta_i)\cdots(1-\beta_l) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{j=K_1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(T-x)\,dx \\
&= \sum_{i=0}^{K_1+K-1} (1-\beta_i)\cdots(1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad - \sum_{j=K_1}^{K_1+K-1} (1-\beta_j) \int_{t_{j-1}}^{t_j} w(x)Q(T-x)\,dx.
\end{aligned}
$$

Then

$$
\begin{aligned}
-\boldsymbol{F}-\boldsymbol{H} &= \sum_{i=0}^{K_1+K-1}(1-\beta_i)\cdots(1-\beta_{K_1+K-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad -\sum_{i=0}^{K_1-1}(1-\beta_i)\cdots(1-\beta_{K_1-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad -\sum_{j=K_1}^{K_1+K-1}(1-\beta_j)\int_{t_{j-1}}^{t_j} w(x)Q(T-x)\,dx - \sum_{i=K_1}^{K_1+K-1}\beta_i\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&= \sum_{i=0}^{K_1+K-1}(1-\beta_i)\cdots(1-\beta_{K_1+K-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad -\sum_{i=0}^{K_1-1}(1-\beta_i)\cdots(1-\beta_{K_1-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad -\sum_{j=K_1}^{K_1+K-1}\int_{t_{j-1}}^{t_j} w(x)Q(T-x)\,dx \\
&= \sum_{i=0}^{K_1+K-1}(1-\beta_i)\cdots(1-\beta_{K_1+K-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad -\sum_{i=0}^{K_1-1}(1-\beta_i)\cdots(1-\beta_{K_1-1})\int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx \\
&\quad -\int_{t_{K_1-1}}^{t_{K_1+K-1}} w(x)Q(T-x)\,dx.
\end{aligned}
$$

From above, we have

$$
\begin{aligned}
&P(L=0, D=1, H_{K_1} | T = t_{K_1+K}) + \int_0^{T-t_{K_1}} f_L(z, D=1, H_{K_1} | T = t_{K_1+K})\, dz \\
&= \boldsymbol{C + (A - B + E + G - D) - F - H} \\
&= \int_{t_{K_1}}^{T} w(x)\, dx + \int_{t_{K_1-1}}^{t_{K_1}} w(x) Q(t_{K_1} - x)\, dx \\
&\quad + \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(t_{K_1} - x)\, dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} \int_{t_{j-1}}^{t_j} w(x) Q(t_j - x)\, dx \\
&\quad - \sum_{i=0}^{K_1+K-1} (1-\beta_i) \cdots (1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x) Q(T - x)\, dx \\
&\quad - \sum_{j=K_1+1}^{K_1+K} \int_{t_{j-1}}^{t_j} w(x) Q(t_j - x)\, dx \\
&\quad + \sum_{i=0}^{K_1+K-1} (1-\beta_i) \cdots (1-\beta_{K_1+K-1}) \int_{t_{i-1}}^{t_i} w(x) Q(T - x)\, dx \\
&\quad - \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(T - x)\, dx - \int_{t_{K_1-1}}^{t_{K_1+K-1}} w(x) Q(T - x)\, dx \\
&= \int_{t_{K_1}}^{T} w(x)\, dx + \int_{t_{K_1-1}}^{t_{K_1}} w(x) Q(t_{K_1} - x)\, dx \\
&\quad + \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(t_{K_1} - x)\, dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} \int_{t_{j-1}}^{t_j} w(x) Q(t_j - x)\, dx - \sum_{j=K_1+1}^{K_1+K} \int_{t_{j-1}}^{t_j} w(x) Q(t_j - x)\, dx \\
&\quad - \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(T - x)\, dx - \int_{t_{K_1-1}}^{t_{K_1+K-1}} w(x) Q(T - x)\, dx
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{t_{K_1}}^{T} w(x)\,dx + \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{j=K_1+1}^{K_1+K-1} \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx - \sum_{j=K_1+1}^{K_1+K-1} \int_{t_{j-1}}^{t_j} w(x)Q(t_j-x)\,dx \\
&\quad - \int_{t_{K_1+K-1}}^{t_{K_1+K}} w(x)Q(t_{K_1+K}-x)\,dx \\
&\quad - \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx - \int_{t_{K_1-1}}^{t_{K_1+K-1}} w(x)Q(T-x)\,dx \\
&= \int_{t_{K_1}}^{T} w(x)\,dx + \int_{t_{K_1-1}}^{t_{K_1}} w(x)Q(t_{K_1}-x)\,dx \\
&\quad + \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_{K_1}-x)\,dx - \int_{t_{K_1-1}}^{T} w(x)Q(T-x)\,dx \\
&\quad - \sum_{i=0}^{K_1-1} (1-\beta_i)\cdots(1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)Q(T-x)\,dx.
\end{aligned}
$$

On the other side of the equation,

$$
\begin{aligned}
&P(D=1, H_{K_1} | T = t_{K_1+K}) \\
=& \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x)[Q(t_{K_1}-x) - Q(T-x)]\, dx \\
&+ \int_{t_{K_1-1}}^{t_{K_1}} w(x)[Q(t_{K_1}-x) - Q(T-x)]\, dx + \int_{t_{K_1}}^{T} w(x)[1 - Q(T-x)]\, dx \\
=& \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(t_{K_1}-x)\, dx \\
&- \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(T-x)\, dx \\
&+ \int_{t_{K_1-1}}^{t_{K_1}} w(x) Q(t_{K_1}-x)\, dx - \int_{t_{K_1-1}}^{t_{K_1}} w(x) Q(T-x)\, dx \\
&+ \int_{t_{K_1}}^{T} w(x)\, dx - \int_{t_{K_1}}^{T} w(x) Q(T-x)\, dx \\
=& \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(t_{K_1}-x)\, dx \\
&- \sum_{i=0}^{K_1-1} (1-\beta_i) \cdots (1-\beta_{K_1-1}) \int_{t_{i-1}}^{t_i} w(x) Q(T-x)\, dx \\
&+ \int_{t_{K_1-1}}^{t_{K_1}} w(x) Q(t_{K_1}-x)\, dx + \int_{t_{K_1}}^{T} w(x)\, dx - \int_{t_{K_1-1}}^{T} w(x) Q(T-x)\, dx \\
=& P(L=0, D=1, H_{K_1} | T = t_{K_1+K}) + \int_0^{T-t_{K_1}} f_L(z, D=1, H_{K_1} | T = t_{K_1+K})\, dz.
\end{aligned}
$$

This finishes the proof.

CURRICULUM VITA

NAME: Ruiqi Liu

ADDRESS: Department of Bioinformatics and Biostatistics
485 E. Gray Street
University of Louisville, KY 40202

DOB: Yueyang, Hunan Province, the People's Republic of China, Auguest 29, 1988

EDUCATION: Ph.D., Biostatistics, 2013-2017
University of Louisville
M.S., Biostatistics 2011-2013
University of Louisville
B.S., Pharmacy, 2006-2010
China Jiliang University

PUBLICATIONS: Liu, R., Gaskins, J.T., Mitra, R., Wu, D. (2017) A Review of Estimation of Key Parameters and Lead Time in Cancer Screening. Accepted. *Colombian Journal of Statistics*.

Liu, R., Wu, D., Zhang, X., Kim, S. (2016) Compound identification using penalized linear regression in metabolomics. *Journal of Modern Applied Statistical Methods*, 15(1): 373-38

Wu, D., Liu, R., Levitt B., Riley T. (2016) Evaluating long-term outcomes using computed tomography in lung cancer screening. *Journal of Biometrics & Biostatistics*, 7:313.

Liu, R., Levitt, B., Riley, T., Wu, D. (2015) Bayesian estimation of the three key parameters in CT for the National Lung Screening Trial Data. *Journal of Biometrics & Biostatistics*, 6:263.

HONORS/AWARDS: Graduate Student Council Travel Award, University of Louisville, *June 2017*

Graduate Student Council Travel Award, University of Louisville, *August 2016*

Graduate Assistantship, Department of Bioinformatics and Biostatistics, University of Louisville, *August 2015 - July 2017*

University of Louisville Graduate Fellowship, *August 2013 - July 2015*