

Bayesian Estimation of the Three Key Parameters in CT for the National Lung Screening Trial Data

Ruiqi Liu¹, Beth Levitt², Tom Riley² and Dongfeng Wu^{1*}

¹Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY, USA

²Information Management Services Inc, USA

Abstract

In this study cancer screening likelihood method was used to analyze the CT scan group in the National Lung Screening Trial (NLST) data. Three key parameters: screening sensitivity, transition probability density from disease free to preclinical state, and sojourn time in the preclinical state, were estimated using Bayesian approach and Markov Chain Monte Carlo simulations. The sensitivity for lung cancer screening using CT scan is high; it does not depend on a patient's age, and is slightly higher in females than in males. The transition probability from the disease-free to the preclinical state has a peak around age 70 for both genders, which agrees with the fact that the highest lung cancer incidence rate appears between age 65 and 74. The posterior mean sojourn time is around 1.5 years for all groups, and that explains why screening only have a short time interval to catch lung cancer. Accurate estimation of the three key parameters is critical for other estimations such as lead time and over-diagnosis, because these quantities are functions of the three key parameters.

Keywords: Sensitivity; Transition probability density; Sojourn time; NLST

Introduction

Lung cancer, the leading cause of cancer death for both men and women, occurs in the lungs and claims more lives each year than do breast, colon, and prostate and ovarian cancers combined. There are two major types of lung cancer which have been identified: about 15% of lung cancers are small cell lung cancer, and the most common type is non-small cell lung cancer. The age-specific lung cancer incidence rate rises with advancing age and reaches its peak between 65 and 74 [1].

Smoking is the major risk factor for development of lung cancer. The general prognosis of lung cancer is poor because symptoms tend not to show up until it is at an advanced stage. Five-year survival is 54.8% for stage I lung cancer, but only 4.2% in advanced, inoperable lung cancer [2]. Since the survival rate for advanced lung cancer is low, early detection and treatment hopefully will lead to a better prognosis. Cancer screening for individuals at high risk has the potential to dramatically improve lung cancer survival rates by finding the disease at an earlier, more treatable stage. In August of 2011, the National Cancer Institute released results from its National Lung Screening Trial (NLST), a randomized clinical trial that screened at-risk smokers with either low-dose helical computed tomography (CT) or single-view chest radiography (X-ray). The final results showed a 20% reduction in lung cancer mortality in the CT arm relative to the X-ray arm.

In NLST, approximately 54,000 male and female heavy smokers (with 30 or more pack-years of cigarette smoking history, and at most 15 years since quitting if former smokers) aged 55-75 years were enrolled between August 2002 and April 2004. Male and female participants were randomly assigned to two study arms in equal proportions: low-dose helical computed tomography (CT), or single-view chest radiography (X-ray), resulting in 15621 male and 10831 female smokers in the CT arm, and 15500 male and 10726 female smokers in the X-ray arm. Participants received screening test annually for 3 years, with the first screening performed at study entry. 15537 male and 10769 female smokers in the CT arm and 15396 male and 10634 female smokers in the X-ray arm had first screening test. If any of the tests was positive, then the screen was considered positive and a definitive work-up exam, such as biopsy, was done. The data that were

used in this study were restricted to the overall, male and female group of CT data and only use age group 55-74, because there were too few people in the 75-year-old group that will cause large bias and variation in the parameter estimation. Each group of data included the number of participants in each screening exam, the number of detected and confirmed cancer cases in each screening exam, and the number of interval cases, stratified by initial age.

We assume the commonly followed disease progression model and the disease develops by progressing through three states[3], denoted by $S_0 \rightarrow S_p \rightarrow S_c$. The state S_0 refers to the disease-free state, where either a person does not have the disease, or the disease is in such an early stage that it cannot be detected by a screening exam. The state S_p is the preclinical disease state, in which an asymptomatic individual unknowingly has the disease that a screening exam can detect. The disease state, S_c , is a state at which the disease manifests itself with clinical symptoms. This was illustrated in Figure 1. The three key parameters in the probability model are: the sensitivity, the sojourn time and the transition probability. The sensitivity is the probability that the screening exam is positive given that the individual is in S_p . The sojourn time refers to the time interval between the beginning of the preclinical state, and the manifestation of clinical symptoms, i.e., $(S_c - S_p)$. The transition probability density from the disease free into the preclinical stage is the probability density function of making a transition from S_0 to S_p , it is in fact a sub-pdf.

We will focus on estimating the three key parameters in CT screening using the NLST data. The reason is 1). CT screening is the

***Corresponding author:** Dongfeng Wu, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA, Tel: 502-852-1888; E-mail: Dongfeng.Wu@Louisville.Edu

Received November 19, 2015; **Accepted** December 01, 2015; **Published** December 08, 2015

Citation: 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. J Biom Biostat 6: 263. doi:10.4172/2155-6180.1000263

Copyright: © 2015 Liu R, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

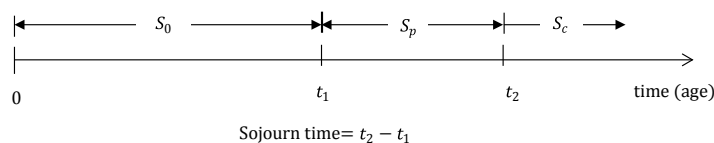


Figure 1: A graphical representation of the disease progressive model.

most current screening modality, commonly known with higher sensitivity. 2). There is little literature on the estimation of the three key parameters in CT scan. 3). Other interesting terms in screening, such as lead time (the diagnosis time advanced by screening), over diagnosis (those whose symptoms would not have appeared before death if untreated) are functions of the three key parameters. Therefore, accurate estimation of the three key parameters is essential and lays a foundation in the study of screening. Furthermore, accurate estimation of the screening sensitivity provides a way to evaluate the predictive performance of a screening modality. Knowing the sojourn time of a disease is necessary for guiding a screening procedure, as usually a case with a longer sojourn time will be easier to catch than the one with a shorter sojourn time. Finally, information about transition probability density can help us determine which age group of people is at higher risk for the disease, so people can take preventive steps before the symptoms show up [4].

Methodology

Let the time variable t represents the participants' 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 (pdf) 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 $t_0 < t_1 < \dots < t_{K-1}$, where $t_i = t_0 + i$ for annual screening exams in the NLST study. Define the i -th screening interval as the time interval between the i -th and the $(i+1)$ -th screening exams (t_{i-1}, t_i) , $i=1,2,\dots, K-1$. 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 i -th screening, S_{i,t_0} is the number of cases detected at the i -th screening exam, and I_{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 the NLST data, since the age of participants enrolled was between 55 to 74 at the study entry, the likelihood function for all groups is:

$$L = \prod_{t_0=55}^{74} \prod_{k=1}^3 D_{k,t_0}^{S_{k,t_0}} I_{k,t_0}^{I_{k,t_0}} \left(1 - D_{k,t_0} - I_{k,t_0}\right)^{n_{k,t_0} - S_{k,t_0} - I_{k,t_0}}, \quad (1)$$

where D_{k,t_0} is the probability that an individual will be diagnosed at the k -th scheduled exam given that he or she is in S_p , and I_{k,t_0} is the probability of being incident in the k -th screening interval. These two probabilities were originally derived in [5]:

$$D_{k,t_0} = \beta(t_{k-1}) \left\{ \sum_{i=0}^{k-2} [1 - \beta(t_i)] \dots [1 - \beta(t_{k-2})] \int_{t_{i-1}}^{t_i} w(x) Q(t_{k-1} - x) dx + \int_{t_{k-2}}^{t_{k-1}} w(x) Q(t_{k-1} - x) dx \right\}, \quad (2)$$

$$I_{k,t_0} = \sum_{i=0}^{k-1} [1 - \beta(t_i)] \dots [1 - \beta(t_{i-1})] \int_{t_{i-1}}^{t_i} w(x) [Q(t_{i+1} - x) - Q(t_i - x)] dx + \int_{t_{k-1}}^{t_k} w(x) [1 - Q(t_k - x)] dx. \quad (3)$$

The three key parameters were estimated from the NLST data using the following parametric models:

$$\beta(t) = \frac{1}{1 + \exp(-b_0 - b_1(t - m))}, \quad (4)$$

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

and

$$Q(x) = \exp(-\lambda x^\alpha), \quad \alpha > 0, \lambda > 0. \quad (5)$$

where t represents age and x is the sojourn time in the preclinical state S_p . We associate the sensitivity β with age t by a logistic link, m is the average age at entry in the whole study group, in this data, $m=61.4$ years. If $b_{-1} > 0$, then $\beta(t)$ will be a monotone increasing function of age t . The lognormal 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[2]. Since participants in the 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. [5,6]. We choose a different sojourn time distribution than Wu et al. [5], 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 2 parameters. However, Weibull is more flexible in that the n -th moments always exist.

Results

The six unknown parameters $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ were estimated based on the NLST data CT arm. We split the data 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 [5,7], 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 subchains, sampling the posterior for (b_0, b_1) , (μ, σ^2) , (λ, α) separately. 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 shows the MCMC trace for of overall group, the MCMC trace for male and females are similar to Figure 2 and we omit here. Figures 3-5 show the density plots for 3 groups, respectively. Bayesian output diagnosis showed that the chains had converged. The posterior estimates for parameters θ and the standard deviations are listed in Table 1.

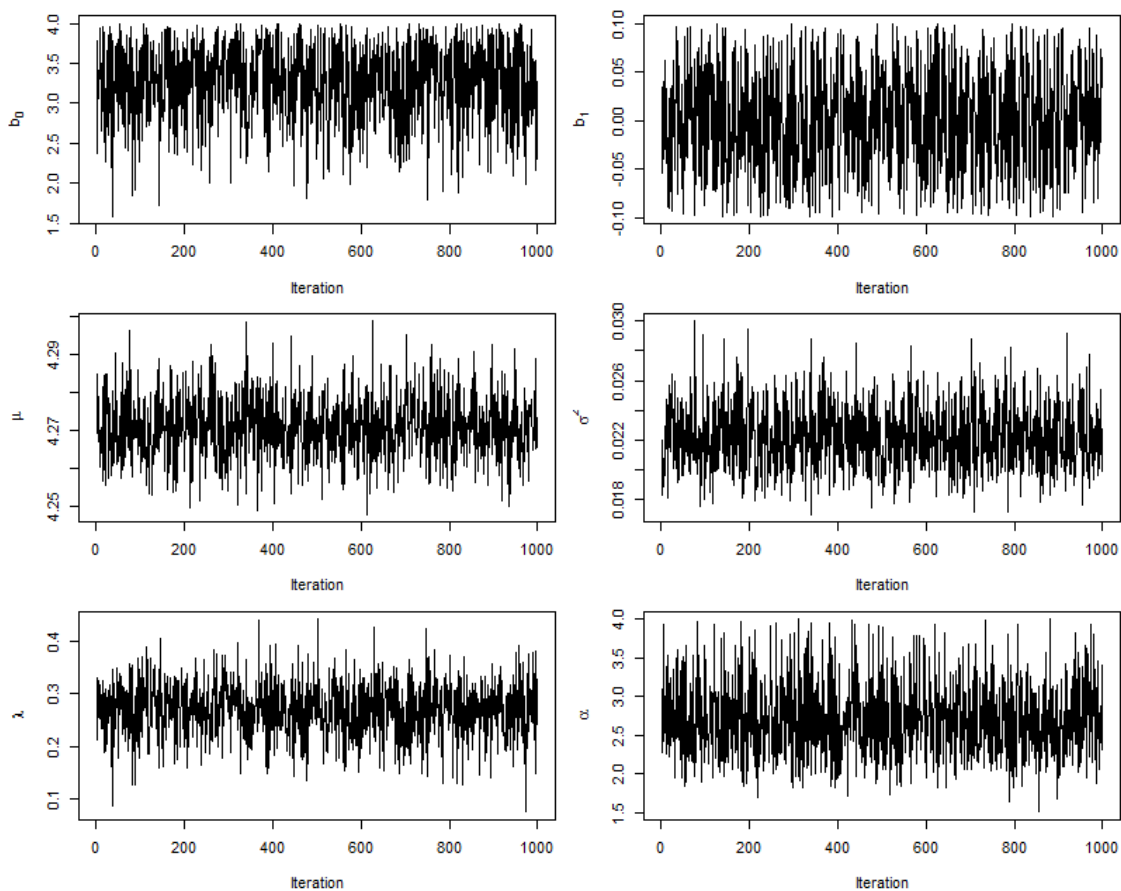


Figure 2: The MCMC trace plots of the parameters $\theta=(b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm overall group in NLST data.

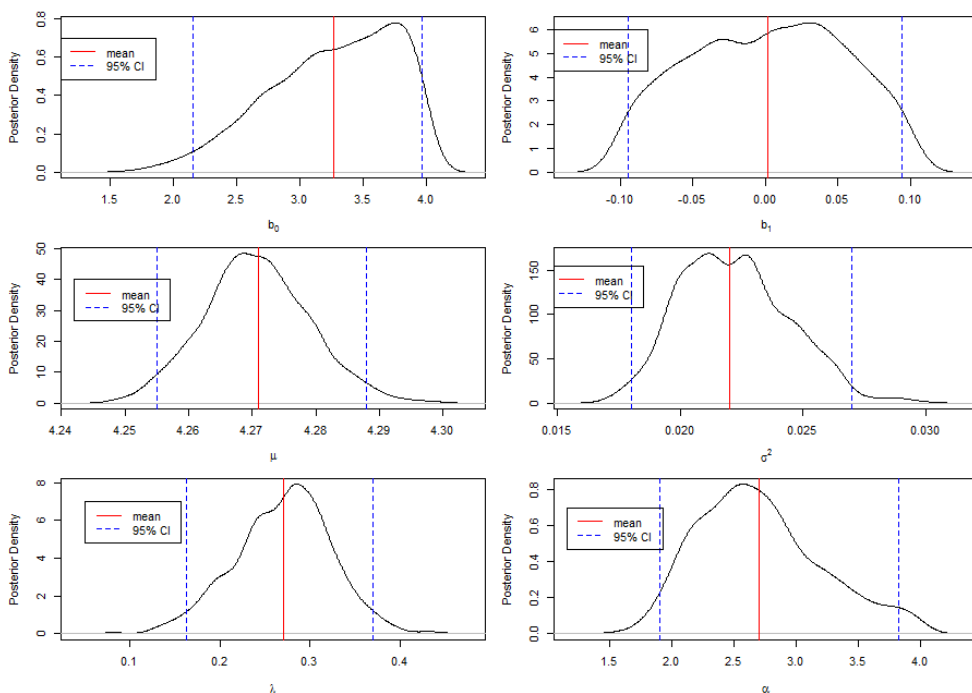


Figure 3: The posterior density plots of the parameters $\theta=(b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm overall group in NLST data.

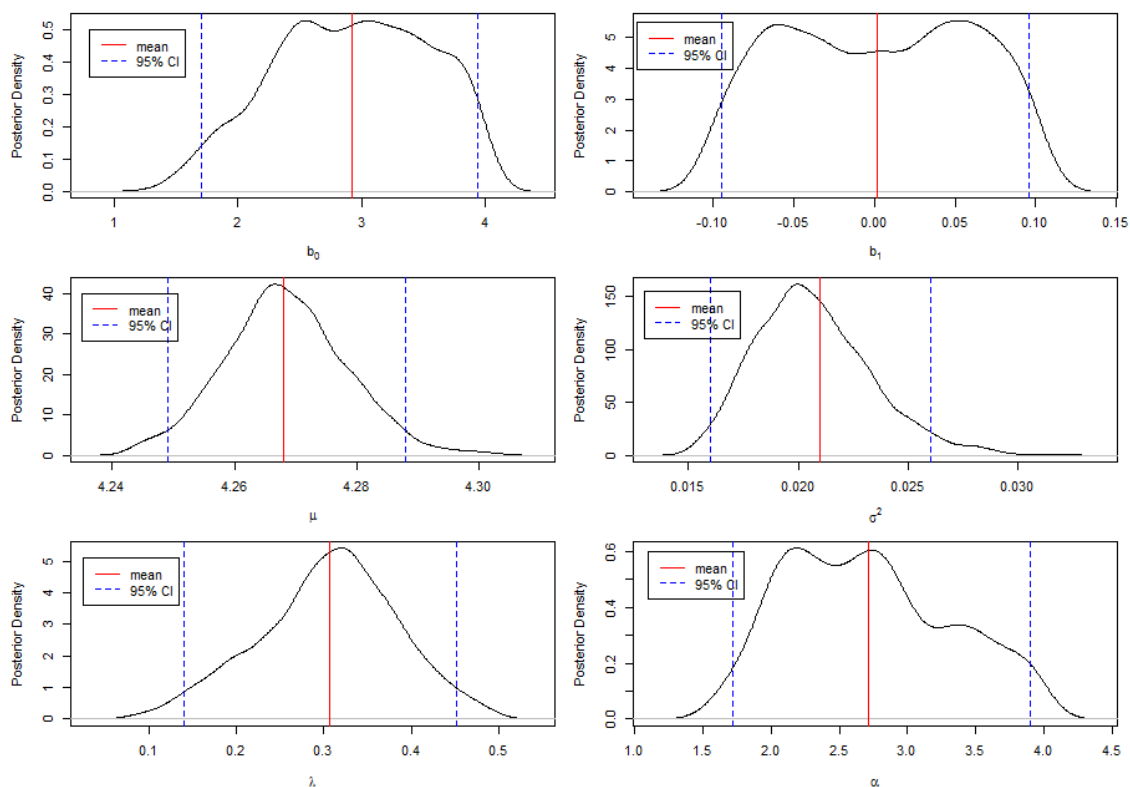


Figure 4: The posterior density plots of the parameters $\theta=(b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm male group in NLST data.

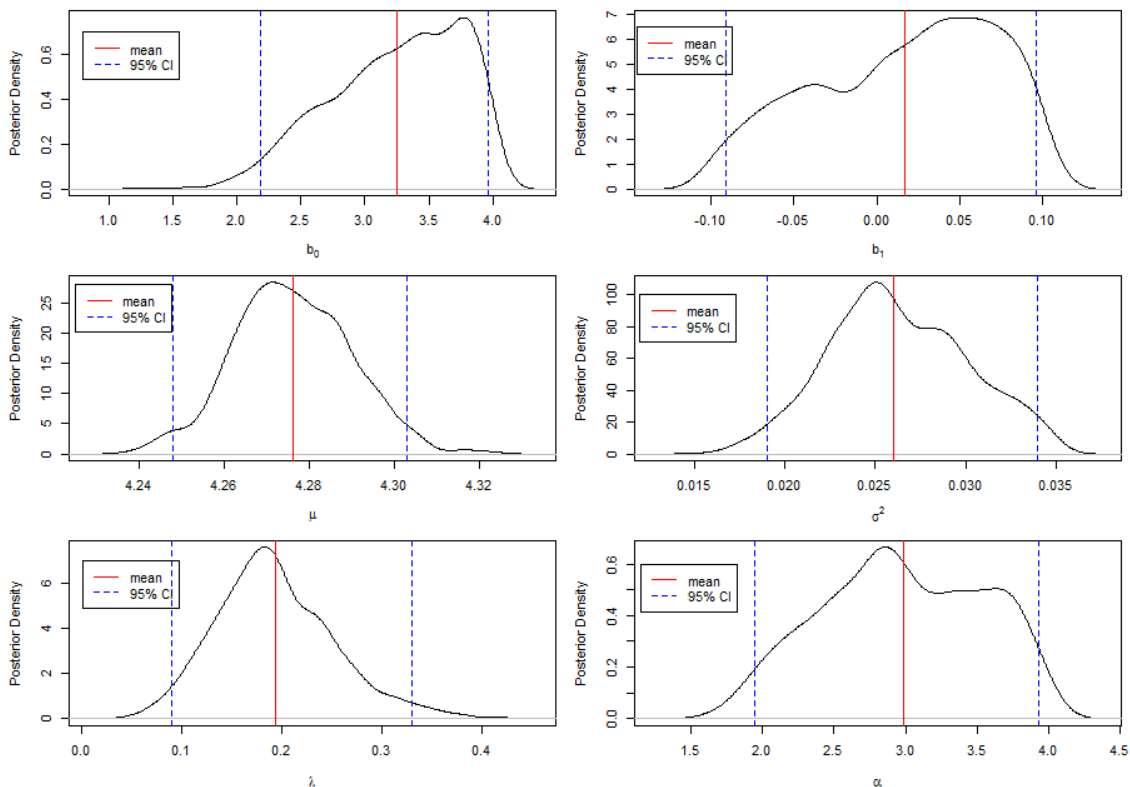


Figure 5: The posterior density plots of the parameters $\theta=(b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ using CT arm female group in NLST data.

	Mean	SD	2.5%	50%	97.5%
CT 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
CT 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
CT 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 1: Bayesian posterior estimates for the 6 parameters in NLST data CT arm.

The age-dependent Bayesian estimates of the sensitivity β and the transition density $w(t)$ for each group are listed in Table 2. Figures 6-8 show posterior quantiles of sensitivity and transition probability for each group.

From equation (4), 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 $P(b_1 > 0 | \text{Data}) = 0.532$; For males group, this posterior probability is $P(b_1 > 0 | \text{Data}) = 0.513$; 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 itself a sub-pdf from our model construction. The posterior density curve of the transition probability could be seen from Figures 6-8. 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 (HPD) 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.

Discussion

In this paper, the three key parameters, screening sensitivity, the transition probability density and the sojourn time distribution, were estimated using Bayesian approach. The NLST CT arm data have been used for the estimation.

For lung cancer, the estimated sensitivity was 56.8% for JHLP control group data, where only X-ray screenings were administered, from the study of Jang et al. [1]. Kim et al. [8] 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 male heavy smokers data, Wu et al. [6] estimated combined X-rays and sputum cytology sensitivity is 89.4%. Chen et al. [9] estimated the screening sensitivity of sputum cytology as a supplement to the chest X-ray using MSKC-LCSP data was 86.64%. Compared with these previous results, the sensitivity estimated in this study was 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. For the NLST data CT arm, Pinsky et al. [10] and Aberle et al. [11] estimated the sensitivity was 93.5% and 94.4%, respectively, which is also closed to our sensitivity estimation.

Age	Sensitivity β			Transition probability for $w(t)$		
	Median	Mean	SE	Median	Mean	SE
CT 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}
CT 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}
CT 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}

Table 2: Bayesian posterior estimates of β and $w(t)$ for each group.

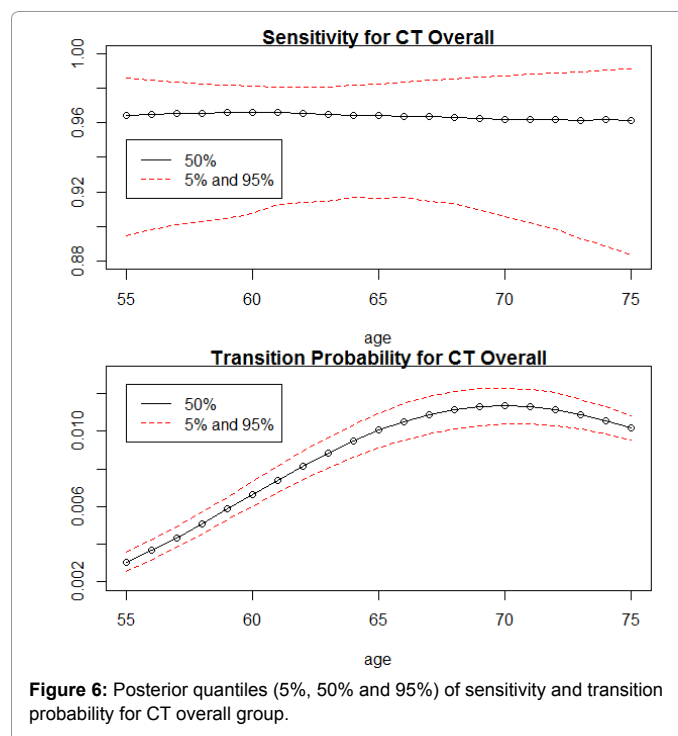
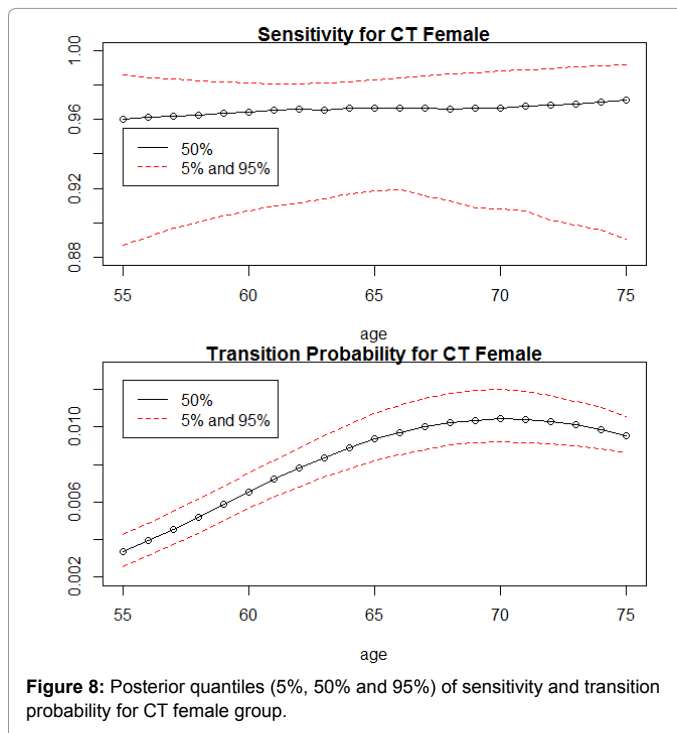
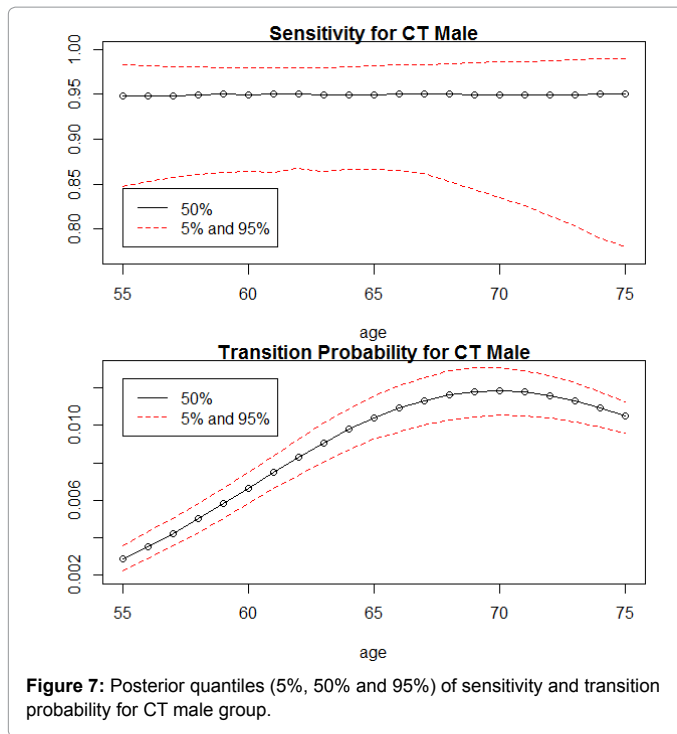


Figure 6: Posterior quantiles (5%, 50% and 95%) of sensitivity and transition probability for CT overall group.



The transition probability from disease-free to preclinical state has a peak around age 70 for both males and females. The transition probability also has a peak around age 70 from Chen's study [9]. The "SEER Fast Fact Stats" [2] show that the highest percent of new lung cancer cases is in 65-74 age group. Our results are consistent with that fact.

In the Mayo Lung Project study, the mean sojourn time was 2.2 years [6], the mean sojourn time for male heavy smokers in MSKC-LCSP data is about 3.35 years. The posterior mean sojourn in this study

is around 1.5 years for both gender groups in this study. Since these two studies were carried out about one or two decades ago, it maybe that today's heavy smokers have a shorter sojourn time. That is, the tumor grows faster than before to present clinical symptoms, and makes it harder to catch the disease during the preclinical state.

Conclusion

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, to lay 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.

Conflict of Interest Declaration

We authors declare here that we have no conflict of interest with other researchers.

Acknowledgements

We authors thank the National Cancer Institute's Cancer Data Access System for allowing us to use the NLST data. This work was conducted in part using the resources of the University of Louisville's research computing group and the Cardinal Research Cluster.

References

- Jang H, Kim S, Wu D (2013) Bayesian lead time estimation for the Johns Hopkins Lung Project data. *J Epidemiol Glob Health* 3: 157-163.
- SEER Fast Stats Results, NIH.
- Zelen M, Feinleib M (1969) On the Theory of Screening for Chronic Diseases. *Biometrika* 56: 601-614.
- Wu D, Erwin D, Rosner GL (2009) Estimating key parameters in FOBT screening for colorectal cancer. *Cancer Causes Control* 20: 41-46.
- Wu D, Rosner GL, Broemeling L (2005) MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. *Biometrics* 61: 1056-63.
- Wu D, Erwin D, Rosner GL (2011) Sojourn time and lead time projection in lung cancer screening. *Lung Cancer* 72: 322-326.
- Wu D, Erwin D, Kim S (2011) Projection of long-term outcomes using X-rays and pooled cytology in lung cancer screening. *Open Access Medical Statistics* 1: 13-9.
- Kim S, Erwin D, Wu D (2012) Efficacy of Dual Lung Cancer Screening by Chest X-Ray and Sputum Cytology Using Johns Hopkins Lung Project Data. *J Biomet Biostat* 3: 1-5.
- Chen Y, Erwin D, Wu D (2014) Over-diagnosis in Lung Cancer Screening using the MSKC-LCSP Data. *J Biomet Biostat* 5: 1-6.
- Pinsky PF, Gierada DS, Black W, Munden R, Nath H, et al. (2015) Performance of Lung-RADS in the National Lung Screening Trial: A retrospective assessment. *Ann Intern Med* 162: 485-491.
- Aberle DR, DeMello S, Berg CD, Black WC, Brewer B, et al. (2013) Results of the two incidence screenings in the National Lung Screening Trial. *N Engl J Med* 369: 920-931.