RESEARCH ARTICLE

Inference of Onset Age of Preclinical State and Sojourn Time for Breast Cancer

Authors

Dongfeng Wu[1]* and Seongho Kim[2]

Affiliations

[1]Department of Bioinformatics and Biostatistics,
University of Louisville, KY 20402
[2]Biostatistics and Bioinformatics Core, Karmanos Cancer Institute,
Department of Oncology, School of Medicine
Wayne State University. Detroit, Michigan 48201

Corresponding author:

Dongfeng Wu, Department of Bioinformatics and Biostatistics, University of Louisville, KY 40202, USA.
Email: dongfeng.wu@louisville.edu

Abstract

Aims: Accurate estimation of the three key parameters (sensitivity, time duration in disease-free state and sojourn time in preclinical state) in cancer screening are critical. Likelihood method with a new link function was applied to the Health Insurance Plan of Greater New York (HIP) breast cancer screening data, to estimate the onset age of preclinical state and the sojourn time in the preclinical state for breast cancer.

Materials and Methods: A new link function to model sensitivity as a function of time in the preclinical state and the sojourn time was adopted. Markov Chain Monte Carlo simulations were used to obtain posterior samples and make inference on the three key parameters. Maximum likelihood estimate was also used for comparison.

Results: The onset age of the preclinical state has a wide range for breast cancer; the peak onset age was 65.07 years (95% credible interval [C.I.], 55.76 to 73.02). The mean sojourn time was 2.00 years (95% C.I., 0.85 to 2.95). The 95 % C.I. for the sojourn time was 0.16 to 5.53 years. Sensitivity at onset of the preclinical state was 0.75 (95% C.I., 0.54 to 0.88); and sensitivity at the end of the preclinical state was 0.84 (95% C.I., 0.67 to 0.88).

Conclusion: The HIP study was the oldest breast cancer mass screening. The estimates reflect key parameters in those days with lower screening sensitivity. However, it is helpful to know other parameters in the planning for future breast cancer screening.

Keywords: Breast Cancer Screening, Sojourn Time, Transition Probability Density, Sensitivity, Markov Chain Monte Carlo.

1. Introduction

Breast cancer is the most diagnosed cancer among American women. In 2021, it is estimated that about 30% of newly diagnosed cancers in women will be breast cancers [1]. About 1 in 8 U.S. women (i.e., 12.5%) will develop invasive breast cancer in her lifetime. An estimated 281,550 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 49,290 new cases of non-invasive (in situ) breast cancer in 2021 [1]. The average 5-year survival rate for women with non-metastatic invasive breast cancer is 90% [2], however, the 5-year survival rate for women with metastatic breast cancer is only 28% [3]. Therefore, early detection through screening exam is important and could improve breast cancer survival when combined with efficient treatments. The goal of screening exam is to catch the disease in its preclinical state, when patients have better prognosis.

This is a brief review of the commonly used disease progressive model with three states: $S_0 \rightarrow S_p \rightarrow S_c$. S_0 is the disease-free state, when a person does not have the disease, or the disease is at an early stage that cannot be detected by any exam. S_p is the preclinical disease state, where an asymptomatic individual unknowingly has the disease that an exam can detect. S_c is the clinical state when clinical symptoms appear.

There are three key parameters in cancer screening: screening sensitivity, sojourn time distribution and transition density. Sensitivity is the probability of getting a positive test result when one is in the preclinical state. Sojourn time is defined as the length of time that one will stay in the preclinical state. Transition density measures the length of time that one stays in the disease-free state. Since all other terms, such as lead time, over-diagnosis, are functions of these three, it is important to get accurate estimation of them. And this is the goal of this study.

Many research has been done in modeling and estimating the three key parameters [4-10]. Wu et al. (2005) has developed a method to model screening sensitivity as a function of a woman's age [6], however, more evidence seems to point out that sensitivity may depend more on how long one has stayed in the preclinical state rather than one's age. And Wu et al. (2021) found a better way to link the sensitivity with the sojourn time and to find the MLE and the Bayesian posterior samples for the three key parameters [7]. In this study we will apply that method to the Health Insurance Plan of Greater New York (HIP) data to estimate the three key parameters in breast cancer.

2. Materials and Methods

The Health Insurance Plan of Greater New York (HIP) is the first randomized mass breast cancer screening trials in North America. It was initiated in December 1963, with about 62,000 asymptomatic women randomized to the study and the control group. The study group participated in 4 annual screening exams, and each exam consists of a mammogram and a clinical physical exam, while the control group did not have any screening except usual care. We will use the HIP study group screening data in the first four years.

We apply the same method in Wu et al. (2021) [7] to the HIP breast cancer screening data. We let $\beta(s|S)$ be the screening sensitivity, where s is the time one has stayed in the preclinical state and S is the total sojourn time in the S_p, and $0 \leq s \leq S$. We define $w(t)$ as the probability density function (PDF) of the time duration in the disease-free state S_0, one can also consider $w(t)dt$ as making a transition from S_0 to S_p in the time interval $(t, t + dt)$. We let $q(x)$ be the PDF of the sojourn time in S_p, and $Q(z) = \int_z^\infty q(x)dx$ is the survival function of the sojourn time.

For a cohort of asymptomatic women with age t_0 at the first exam, assume that there are K ordered screening exams at ages $t_0 < t_1 < \cdots < t_{K-1}$. We let $t_{-1} = 0$. At the t_{i-1}, let n_{i,t_0} be the total number of individuals examined at the ith exam, s_{i,t_0} is the number of diagnosed cases at the ith exam, and r_{i,t_0} is the number of clinical incident cases in $(t_{i-1}, t_i), i = 1, \ldots, K$. For the HIP study, $K = 4$, and the age of participants enrolled was between 40 to 64 at the study entry, the likelihood function for all age groups is:

$$L = \prod_{t_0=40}^{64} \prod_{k=1}^{4} D_{k,t_0}^{s_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} \left(1 - D_{k,t_0} - I_{k,t_0}\right)^{n_{k,t_0} - s_{k,t_0} - r_{k,t_o}} \tag{1}$$

where D_{k,t_0} is the probability that an individual is diagnosed at the kth exam given that she is in S_p and I_{k,t_0} is the probability of interval case in (t_{k-1}, t_k), as given in Wu et al. (2021) [7]:

$$D_{1,t_0} = \int_0^{t_0} w(x) \int_{t_0-x}^{\infty} q(t)\beta(t_0 - x|t)dtdx.$$

And for $k = 2, \ldots, K$,

$$D_{k,t_0} = \sum_{i=0}^{k-2} \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{\infty} q(t)\,\beta(t_{k-1} - x|t) \left\{ \prod_{j=i}^{k-2} \left[1 - \beta\left(t_j - x\middle|t\right)\right] \right\} dt\,dx$$
$$+ \int_{t_{k-2}}^{t_{k-1}} w(x) \int_{t_{k-1}-x}^{\infty} q(t)\,\beta(t_{k-1} - x|t)dt\,dx$$

$$I_{k,t_0} = \sum_{i=0}^{k-2} \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{t_k-x} q(t) \left\{ \prod_{j=i}^{k-1} \left[1 - \beta\left(t_j - x\middle|t\right)\right] \right\} dt\,dx$$
$$+ \int_{t_{k-1}}^{t_k} w(x)[1 - Q(t_k - x)]\,dx,$$

Parametric link functions are used here: $\beta(s|S) = [1 + \exp(-b_0 - b_1 x)]^{-1}$, where $x = \frac{s}{S} \in [0,1], b_0 \geq 0, b_1 \geq 0$. And

$$w(t|\mu, \sigma^2) = \frac{0.2}{\sqrt{2\pi}\sigma t} \exp\left\{ -\frac{(logt - \mu)^2}{2\sigma^2} \right\}$$

$q(x|\alpha, \lambda) = \alpha\lambda x^{\alpha-1} \exp(-\lambda x^\alpha)$, and $Q(x) = \exp(-\lambda x^\alpha)$, with $\lambda > 0$ *and* $\alpha > 0$.

There are 6 unknown parameters $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$ in the parametric link functions. We use two methods to estimate θ, the Maximum likelihood estimate (MLE) and the Bayesian posterior samples of θ.

3. Results

We applied our method to the HIP study group data. To find the MLE, we used the built-in function "nlminb" in R. To obtain the Bayesian posterior samples, we used Markov Chain Monte Carlo simulation and Gibbs sampler with non-informative priors for three sub-chains $(b_0, b_1), (\mu, \sigma^2)$, $(\lambda\ , \alpha)$. The sensitivity was estimated to be around 0.76 using the empirical method $\frac{\sum s_i}{\sum s_i + \sum r_i}$ (that is, using all screen-detected cases divided by the total cancer cases, including both screen-detected and interval cases), so uniform prior was chosen for (b_0, b_1), with a boundary of $b_0 \geq 0, b_1 \geq 0,$ and $b_0 + b_1 \leq 2,$ such that sensitivity at the onset of the S_p has a lower bound of $[1 + \exp(-b_0)]^{-1} = [1 + \exp(-0)]^{-1} = 0.5$, and sensitivity at the end of the S_p (or the onset of the S_c) has an upper bound of $[1 + \exp(-b_0 - b_1)]^{-1} = [1 + \exp(-2)]^{-1} = 0.88$. The prior distribution for (μ, σ^2) and $(\lambda\ , \alpha)$ are both non-informative bivariate Normal with (0,0) as the mean and a diagonal matrix as the variance with 10^{10} on the diagonal. The jumping density of the candidates were bivariate Normal centered at the current values. Twelve chains ran 6000 steps with over dispersed initial values. After a burn-in of 1000 steps, the Gelman-Rubin statistics were calculated and the chains showed convergence, and then we thin the chain every 100 steps and provide 50 posterior samples from each chain, the pooled posterior sample is 600. Figure 1 is the trace plot of the pooled posterior samples, which looks like random noise. Figure 2 is the estimated density curve for the six parameters. The MLE and the posterior mean, median, standard error (S.E.) and the corresponding 95% highest posterior density (HPD) credible interval (C.I.) for parameters θ are listed in Table 1.

Table 1 shows that the MLEs and the corresponding posterior medians are close to each other, except for b_0. This is compatible with the density of b_0 in Figure 2, which is flat in a large interval, while all other parameters have a density of unimodal. Based on the MLE and the posterior samples, we can obtain information on the sensitivity, the transition age into the preclinical state and the sojourn time in the preclinical state.

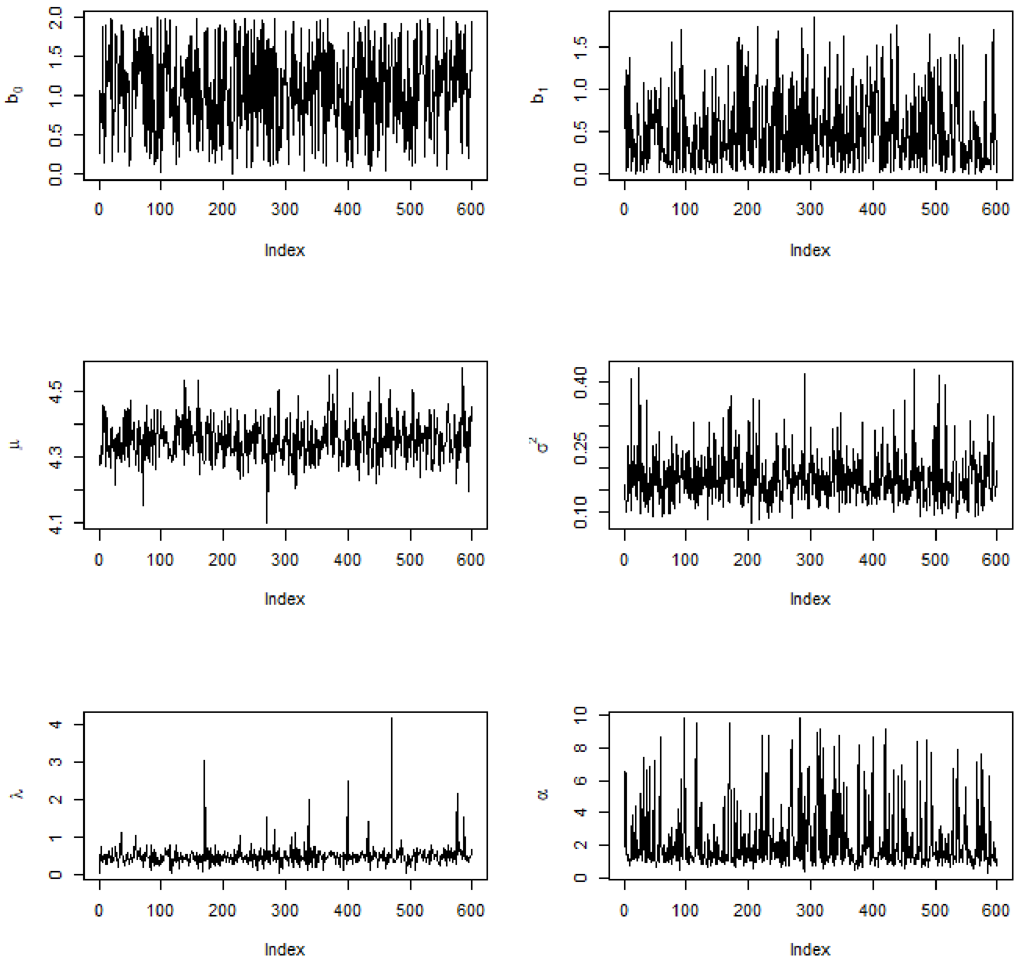

Figure 1: Trace Plot of the 600 pooled posterior samples.

Table 1: MLE and Bayesian posterior estimates using the HIP data.

Parameter	MLE	Bayesian posterior estimate			
		Mean	Median	S.E.	95% C.I.
b_0	0.000	1.075	1.081	0.553	(0.142, 1.976)
b_1	0.216	0.489	0.363	0.436	(0.001, 1.404)
μ	4.361	4.350	4.344	0.062	(4.228, 4.462)
σ^2	0.133	0.177	0.164	0.059	(0.087, 0.299)
λ	0.231	0.495	0.478	0.281	(0.115, 0.779)
α	1.388	2.240	1.395	1.999	(0.260, 6.962)

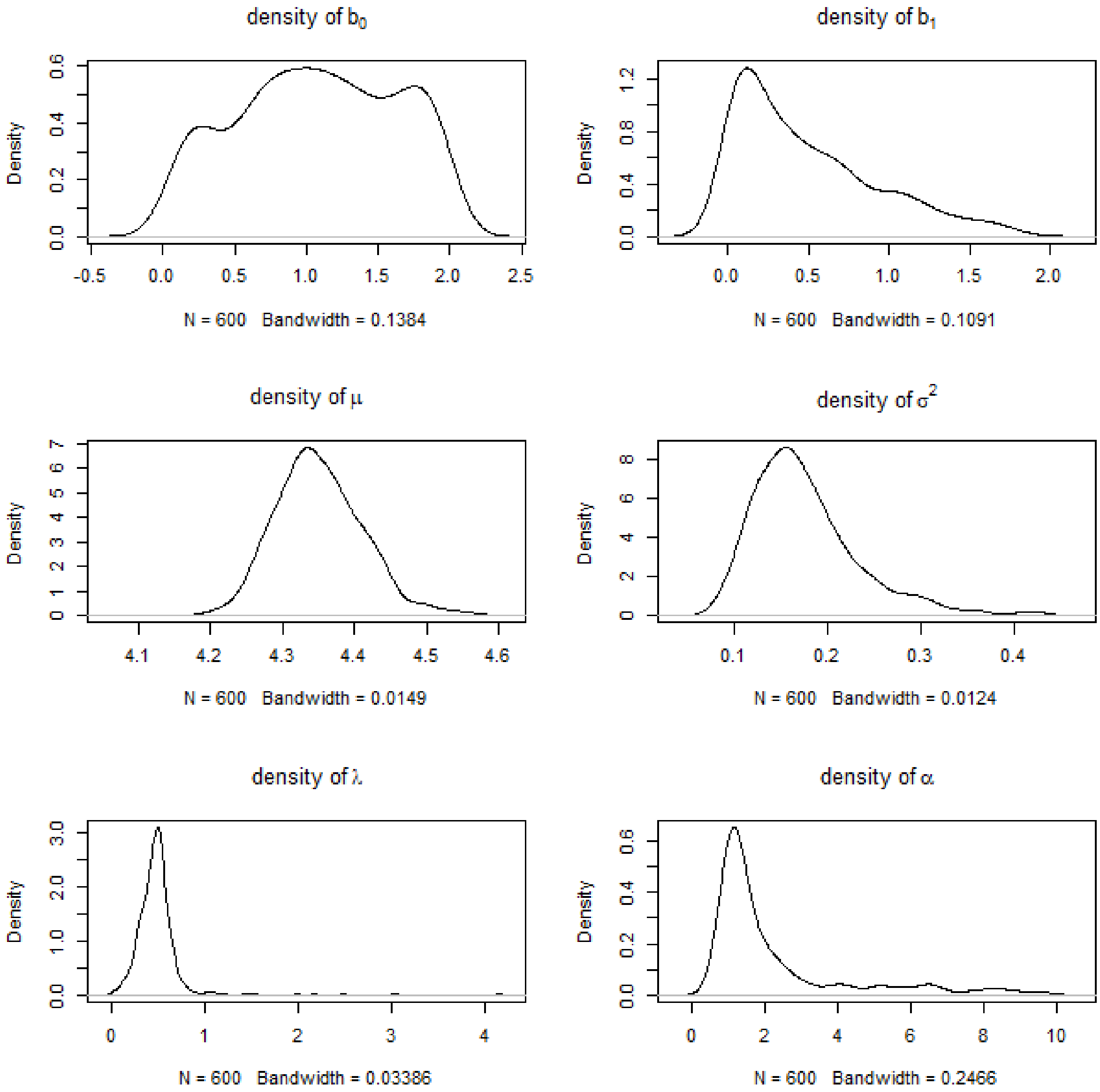

Figure 2: Estimated density of $\theta = (b_0, b_1, \mu, \sigma^2, \lambda, \alpha)$.

If we use the MLE $(b_0, b_1) = (0, 0.216)$, the sensitivity at the onset of S_p would be $\hat{\beta}_0 = \frac{1}{1+\exp(-\hat{b}_0)} = 0.5$, and at the end of S_p, $\hat{\beta}_1 = \frac{1}{1+\exp(-\hat{b}_0 - \hat{b}_1)} = 0.554$. For Bayesian inference, we can use all 600 posterior samples of (b_0, b_1), and each pair can provide an estimate of (β_0, β_1), so we get 600 pairs of the onset sensitivity and the end-of-state sensitivity. The mean, the median and the 95% highest posterior density (HPD) credible interval at the onset of S_p are 0.73, 0.75 and (0.54, 0.88) correspondingly; and the numbers become 0.82, 0.84, and (0.67, 0.88) at the end of S_p, see Table 2. The sensitivity as a function of $x = s/S$, (where s is the time one stayed in the preclinical state, and S is the sojourn time, $0 \leq s \leq S$) using the posterior mean of

(b_0, b_1) was plotted as the dotted line in Figure 3. The pointwise average of sensitivity and its 95% corresponding C.I. using all posterior samples are plotted in the same graph.

Table 2: Estimated sensitivity, transition age and sojourn time based on the MLE and posterior samples.

	MLE	Posterior samples		
		Mean	Median	95% C.I.
β_0	0.50	0.73	0.75	(0.54, 0.88)
β_1	0.55	0.82	0.84	(0.67, 0.88)
	Mean, median, mode	Mean	Median	Mode
w(t)	83.72, 78.34, 68.58	84.92 (73.75, 98.55)	77.64 (68.60, 86.69)	65.07 (55.76, 73.02)
q(x)	2.62, 2.21, 1.15	2.00 (0.85, 2.95)	1.34 (0.86, 1.90)	0.66 (0.00, 1.37)

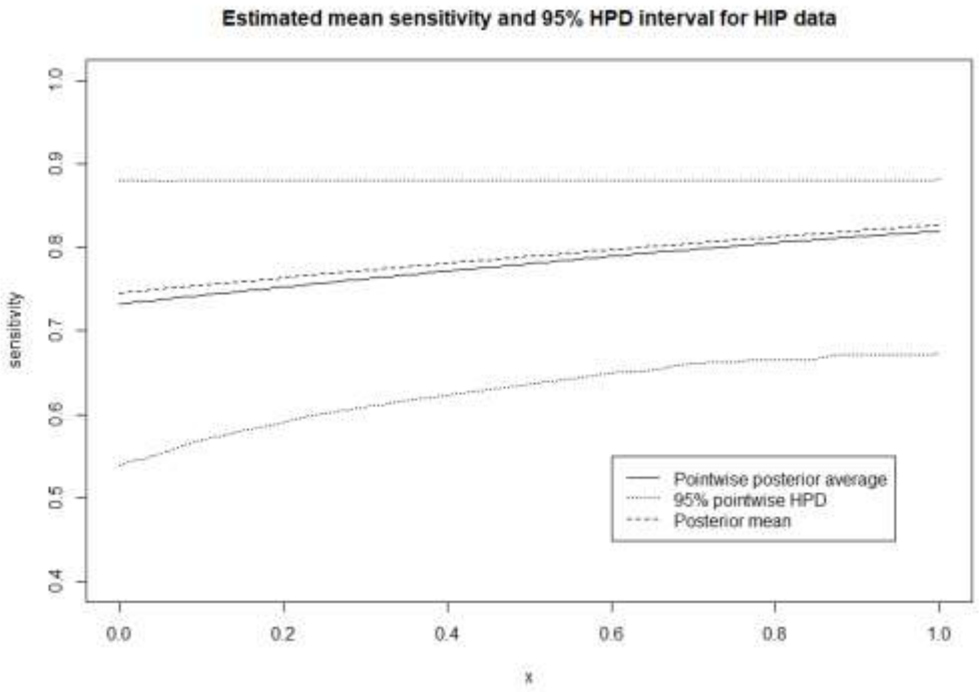

Figure 3: Estimated posterior mean sensitivity, pointwise sensitivity and 95% HPD C.I.

We can estimate the transition age from the disease-free to the preclinical state using the estimate of (μ, σ^2). For the lognormal PDF of w(t), the mean equals $e^{\mu+\frac{1}{2}\sigma^2}$, the mode is $e^{\mu-\sigma^2}$, the median is e^{μ}, and the standard deviation is $\sqrt{e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}}$. We plug in the MLE of $(\mu, \sigma^2) = (4.361, 0.133)$ to get the mean transition age of 83.72 years, median transition age is 78.34 years, with a mode (peak) transition age at 68.58 years and the standard deviation of 31.58 years. We can use all 600 posterior samples to estimate the time duration in the disease-free state: each pair (μ, σ^2) would generate a curve for *w(t)*, hence provide the corresponding mean, median and mode (in years), and we can get the average and the corresponding 95% C.I. (i.e., empirical HPD interval) in Table 2. The density curve using the posterior mean (dotted line) and pointwise posterior average (solid line) with the 95% pointwise credible band were plotted in Figure 4. We can estimate the HPD interval for the density *w(t)* as well. Since each pair of (μ, σ^2) would provide a density of *w(t)* and a corresponding 90% HPD interval for *w(t)*, we can take the average of the HPD intervals to get a better estimate. The 90% HPD interval for *w(t)* is (31.33, 137.61) years, and the 75% HPD interval for *w(t)* is (38.89, 109.88) years. If we use the MLE of (μ, σ^2), the 75% and the 90% HPD interval for *w(t)* are (43.84, 107.28) and (36.21, 129.88) correspondingly. The probability that the disease-free state lasts longer than 50 years old is about 85.46% among those who would develop breast cancer. The United States Preventive Services Task Force (USPSTF) recommends biennial screening mammography for women aged 50 to 74 years [4]. Therefore, we calculate the probability of making a transition to the preclinical in this age interval (50, 74), and it is 31.32%. This estimate may not be appropriate since human lifetime is assumed unlimited in the lognormal PDF of *w(t)*. However, if we use the average lifetime of US women, which is 80.5 years, then the conditional probability of making a transition from the disease-free to the preclinical state in the age interval (50, 74) would be approximately $P(50 \leq T_1 \leq 74 | T_1 \leq 80.5) = 57.96\%$, which is still not very large.

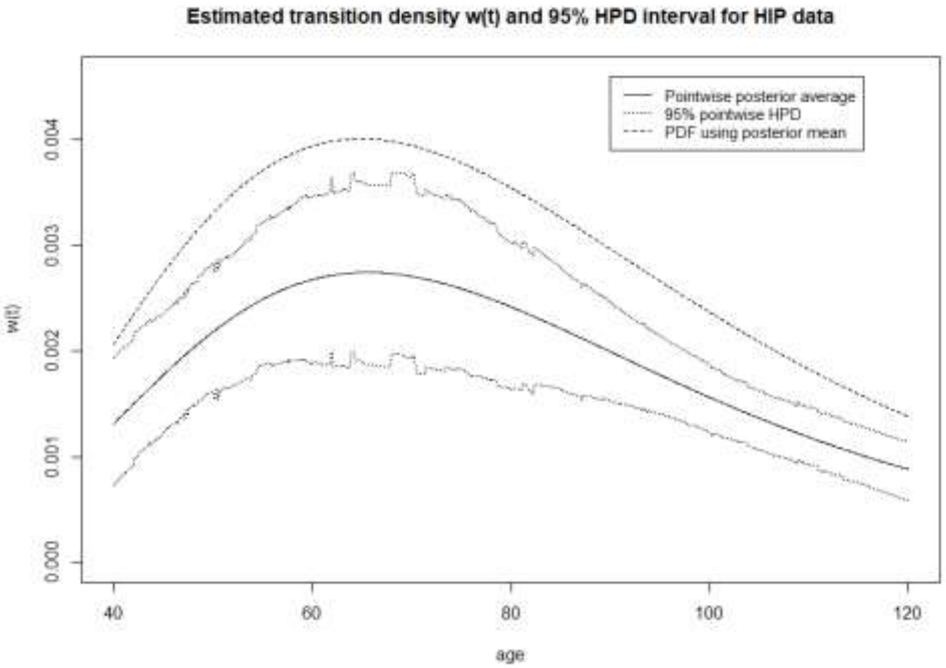

Figure 4: Posterior quantities (2.5%, 50%,97.5%) of transition probabilities

We can estimate the mean sojourn time from the estimates of (λ, α). Since the sojourn time follows the Weibull distribution, the r th moment is $E(X^r) = \frac{\Gamma\left(1+\frac{r}{\alpha}\right)}{\lambda^{\frac{r}{\alpha}}}$ for any $r > \frac{1}{\alpha}$; the median is $(ln2/\lambda)^{\frac{1}{\alpha}}$; and the mode is $\left(\frac{\alpha-1}{\lambda\alpha}\right)^{\frac{1}{\alpha}}$ if $\alpha > 1$. If we use the MLE of (λ, α) = (0.231, 1.388), the mean, median and mode are 2.62, 2.21 and 1.15 years correspondingly. If we use all 600 posterior samples, each pair (λ, α) provides a density curve *q(x)*, from which we can obtain the mean, median, and mode, then we take the average and find the corresponding 95% C.I. and summarize in Table 2. The mean, median and mode of the sojourn time are 2.00, 1.34, and 0.66 years, shorter than that using the MLE estimate. Similarly, from each density curve, we can obtain a 95% HPD interval for the sojourn time itself, and taking the average, the 90%, 95% and 99% HPD intervals for the sojourn time in the preclinical state are (0.19, 3.96), (0.16, 5.53), and (0.10, 11.79) years correspondingly. The density curve *q(x)* using the posterior mean, using the pointwise posterior average and the 95% pointwise HPD band were plotted in Figure 5.

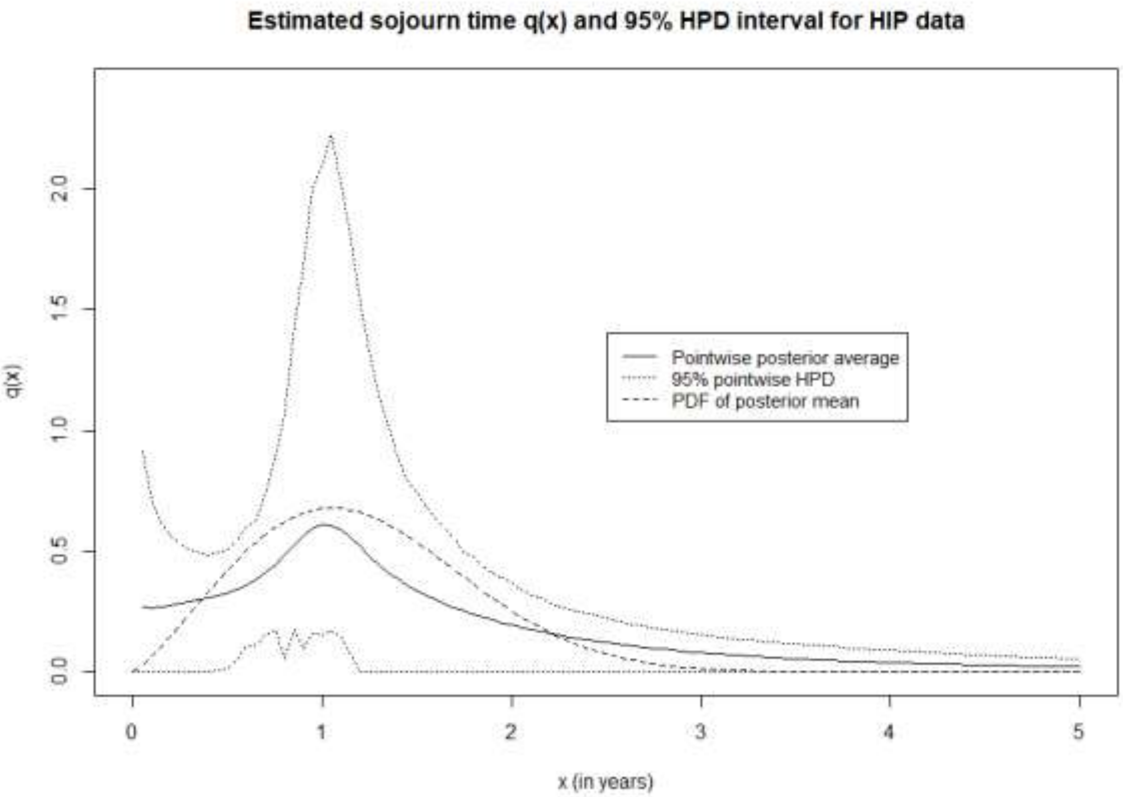

Figure 5: Posterior quantities (2.5%,50%,97.5%) of sojourn time probabilities

4. Discussion and Conclusion

We applied the likelihood method with a new link function to the HIP data, to get the MLE and Bayesian posterior samples for the model parameters θ, and then using the model parameters to estimate the three key parameters: sensitivity, transition density and sojourn time.

The result shows that the mean sensitivity is 0.73 and median sensitivity is 0.75 at the onset of the preclinical state with the 95% C.I. (0.54, 0.88), and it is 0.82 (mean) and 0.84 (median) at the end of the preclinical state (or the onset of the clinical state), with the 95% C.I. (0.67, 0.88). This is compatible with the epidemiology estimate of the sensitivity of 0.76, which is a rough estimate of the overall sensitivity [8]. There was many research using the HIP data to estimate screening sensitivity for breast cancer, with the assumption that either sensitivity was fixed, or it depends on one's age [6]. This project is the first one using the HIP data to estimate the sensitivity as a function of time in the preclinical state relative to the total sojourn time. It enables us to estimate the sensitivity at the onset and at the end of the preclinical state, which is an improvement over the existing ones. This improvement also makes it possible to improve other modeling in cancer screening, such as how to estimate the lead time and overdiagnosis, and how to schedule the future screening exams, since all other terms are functions of the three key parameters.

Bayesian posterior samples make it easy to estimate credible intervals for the time duration in the disease-free state and the sojourn time in the preclinical state. Using the 600 posterior samples of (μ, σ^2), the 75% and the 90% HPD interval for the disease-free state time duration are (38.89, 109.88) and (31.33, 137.61) years respectively. This shows that the onset age for breast cancer could span a much larger age interval than we thought. The peak transition age (i.e., the mode) from the disease-free to the preclinical state is 65.07 years, with the 95% credible interval (55.76, 73.02) years, which is compatible to the result in Wu et al. (2005) [6], although that paper did not estimate this credible interval. The estimated probability of making a transition from the disease-free state to the preclinical state in the age interval [50, 74] is approximately 58%, very low, given that this is the age interval currently recommended for breast cancer screening by the US Preventive Services Task Force.

Using the posterior samples, the mean sojourn time is 2.00 years, and the median sojourn time is 1.34 years. The 90% and 95% credible interval for the sojourn time in the preclinical state is (0.19, 3.96) and (0.16, 5.53) years respectively. This implies that breast cancer has a relatively large variation regarding the time duration in the preclinical state: some fast-growing tumor may have a very shorter sojourn time and hard to get detected by screening; some slow-growing tumor may have a longer sojourn time and hence easier to get detected by screening. Overall, the time duration in the preclinical state is relatively large to carry out screening and catch the disease early.

Finally, the HIP study is the earliest study on breast cancer screening carried out in the 1960s. Since then, mammogram for breast imaging has improved a lot, and digital mammogram has gradually replaced traditional technology [9], hence the screening sensitivity has increased dramatically. However, this project still provides very useful information regarding the time duration in the disease-free state and sojourn time in the preclinical state, both are very important in the planning of future screening. Currently the US Preventive Services Task Force recommends biennial screening for women's breast cancer from 50 to 74 years old. Our estimate shows that the probability to enter the preclinical state in the age interval (50,74) is less than 60% among those at risk, which is low, and a larger age interval maybe more appropriate. On the other hand, other factors such as cost or risk, must be balance in practice as well. We are also considering improving our likelihood function in future research [10-12].

Acknowledgements

D.W. was partially supported by NIH/NCI 1R15CA242482. S.K. was partially supported by the United States Public Health Service Cancer Center Support Grant P30 CA022453 and National Institute of General Medical Sciences Grant R21GM140352.

References

[1] https://www.breastcancer.org/symptoms/understand_bc/statistics. Accessed 10/19/2021.

[2] https://www.cancer.net/cancer-types/breast-cancer/statistics. Accessed 10/19/2021.

[3] https://www.cancer.net/cancer-types/breast-cancer-metastatic/statistics. Accessed 11/2/2021.

[4]https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening. Accessed 11/20/2021.

[5] https://www.cdc.gov/nchs/data/vsrr/VSRR10-508.pdf. Accessed 11/20/2021.

[6] Wu D, Rosner GL, Broemeling L. (2005). MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. *Biometrics.* 2005; 61(4): 1056-1063.

[7] Wu D, Rai SN, and Seow A. (2021). Estimation of preclinical state onset age and sojourn time for heavy smokers in lung cancer. Statistics and Its Interface. Accepted. NIHMS ID: 1734205.

[8] Walter SD, Day NE. Estimation of the duration of a pre-clinical disease state using screening data. *Am J Epidemiol*. 1983; 118(6): 865-886.

[9] Joe, BN. And Sickles, EA. (2014). The evolution of breast imaging: past to present. *Radiology*. 273 (2). S23-S44.

[10] Chen Y, Brock GN and Wu D. (2010). Estimating key parameters in periodic breast cancer screening - application to the Canadian National Breast Screening Study data. *Cancer Epidemiology*. 34, 429-433. DOI: 10.1016/j.canep.2010.04.001.

[11] Kim S, and Wu D. (2016). Estimation of sensitivity depending on sojourn time and time spent in preclinical state. Statistical Methods in Medical Research. 2016, Vol. 25(2), 728-740.
DOI: 10.1177/0962280212465499.

[12] Wu D, Kim S (2020). Problems in the estimation of the key parameters using MLE in lung cancer screening. *J Clin Res Rep.* 2020; 5(3).