Revista Colombiana de Estadística
July 2017, Volume 40, Issue 2, pp. 263 to 278
DOI: http://dx.doi.org/10.15446/rce.v40n2.60642

A Review of Estimation of Key Parameters and Lead Time in Cancer Screening

Una revisión de la estimación de los parámetros claves y el tiempo de ventaja en la búsqueda de cáncer

Ruiqi Liu[a], Jeremy T. Gaskins[b], Ritendranath Mitra[c], Dongfeng Wu[d]

Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, United States

Abstract

Early detection combined with effective treatment are the only ways to fight against cancer, and cancer screening is the primary technique for early detection. Although mass cancer screening has been carried out for decades, there are many unsolved problems, and the statistical theory of cancer screening is still under developed. Screening sensitivity, time duration in the preclinical state, and time duration in the disease free state are the three key parameters, which are critical in cancer screening, since all other estimates are functions of the three key parameters. Lead time is the diagnosis time advanced by screening, and it serves as a measurement of effectiveness of screening programs. In this article, we provide a review for major probability models and statistical methodologies that have been developed on the estimation of the three key parameters and the lead time distributions. These methods can be applied to screening of other chronic diseases after slight modifications.

***Key words*:** Cancer, Lead time, Sensitivity, Sojourn time, Transition density.

Resumen

Detección temprana combinada con la efectividad de los tratamientos son las únicas formas de combatir en contra del cáncer, y el examen de búsqueda temprana es la técnica principal para detección temprana. A pesar de que la búsqueda temprana de la masa cancerígena se ha realizado pro décadas,

[a]PhD. E-mail: ruiqi.liu@louisville.edu
[b]PhD. E-mail: jeremy.gaskins@louisville.edu
[c]PhD. E-mail: ritendranath.mitra@louisville.edu
[d]PhD. E-mail: dongfeng.wu@louisville.edu

hay muchos problemas sin resolver, y la teoría estadística de la búsqueda del cáncer está todavía en desarrollo. Los tres parámetros claves: sensibilidad de la búsqueda, la duración en tiempo en el estado pre-clínico, y la duración en tiempo de la enfermedad en estado libre, son críticos en la búsqueda de cáncer; esto es porque todos los otros estimadores son funciones de estos tres parámetros claves. El tiempo de ventaja es el tiempo de diagnóstico avanzado por la búsqueda, y sirve como una medida de la efectividad de los programas de búsqueda. En este artículo, presentamos una revisión de los modelos de probabilidad principales y las metodologías estadísticas que han sido desarrolladas en la estimación de los tres parámetros claves y las distribuciones del tiempo de ventaja. Estos métodos pueden ser aplicados a la búsqueda de otras enfermedades crónicas con modificaciones menores.

***Palabras clave*:** búsqueda de cáncer, densidad de transición, sensibilidad, tiempo de estadía, tiempo de ventaja.

1. Introduction

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 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.

The commonly followed progressive model used in cancer screening and its parameters are outlined below. A cohort of apparently healthy individuals are enrolled in a screening program to detect the presence of a specific disease. The disease progression stochastic model was first proposed by Zelen & Feinleib (1969) and has been used since then. In this model, the disease progresses through 3 states: $S_0 \to S_p \to S_c$ (See Figure 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 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.

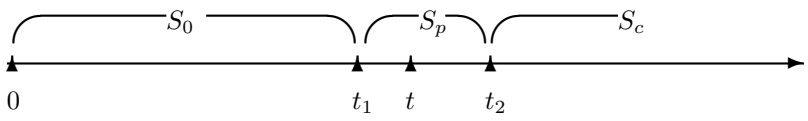

FIGURE 1: Disease progressive states and the lead time.

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, $\beta = P(X = 1 \mid D = 1)$. *Specificity* is the probability of correctly identifying those who do *not* have the disease, that is, $\alpha = P(X = 0 \mid D = 0)$. 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 take 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.

TABLE 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, the sensitivity is $\beta = n_{11}/(n_{11} + n_{21})$, and the specificity is $\alpha = n_{22}/(n_{12} + n_{22})$, 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, although their sum is observed. Hence, β and α cannot be obtained from data directly. Also, a screened negative individual who has been followed and found to be positive later may fall into one of two cases: either it was a false negative on the previous screening exams, or it is a newly developed case. However, the sensitivity can be estimated by likelihood-based estimation from mass screening data (Shen & Zelen 1999, Wu, Rosner & Broemeling 2005, Wu, Wu, Banicescu & Cariño 2005).

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). 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, previous analyses have shown that the preclinical state of breast cancer may last from 1 to 4 years (Shen & Zelen 1999, Shen, Wu & Zelen 2001, Wu, Rosner & Broemeling 2005, Wu, Wu, Banicescu & Cariño 2005), and sojourn time may last longer for colorectal cancer (Wu, Erwin & Rosner 2009*b*). Hence, cancers with longer sojourn time are more likely to 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. It is commonly assumed that the sojourn time

and the transition time are independent (Wu, Rosner & Broemeling 2005, Wu, Wu, Banicescu & Cariño 2005). 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, 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. In the next few sections, we will 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. Finally, we close the article with a brief discussion of the variations of these methods as applied to different cancer sites, such as breast, colon, and lung cancer.

2. Estimation of the Three Key Parameters

We first introduce some notation used in the remainder of the article. 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 often modeled as 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 pieces of information from each screening cycle: n_i is the total number of individuals examined at ith screening (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 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 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
⋮										

2.1. Likelihood Function in Stable and Nonstable Disease Models

Shen & 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 screening given at age t_{i-1}. For an individual who is diagnosed at the ith screening ($i > 1$), the person has to be tested as negative at all previous $(i-1)$ screening exams and stay in preclinical state at least till t_{i-1}. It can be calculated by

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

Let I_i be the probability of an individual being incident in the ith interval. The person has failed to be detected at i previous exams (true negative or false negative), and develops clinical cancer at time point t after t_{i-1}. The person can enter the preclinical state at anytime before t. It is given by

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

Thus, the full likelihood function was derived as

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

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

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

In both papers (Shen & Zelen 1999, Shen et al. 2001), the data was not stratified by age, which means Table 2 could be collapsed into a vector. Two breast cancer screening datasets, the Health Insurance Plan (HIP) study and the Canadian National Breast Screening study were used in both stable and nonstable model. In the nonstable model, estimates of the transition rate w for each five year interval 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.

2.2. Estimation of age-dependent sensitivity and transition probability

Wu, Rosner & Broemeling (2005) 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 incident case may be detected. Let $(n_{i,t_0}, s_{i,t_0}, r_{i,t_0})$ be the data for the ith screening for the strata with starting age t_0. Then the likelihood for the individuals aged t_0 at study entry is proportional to

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

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. Here, we work with a single testing modality, so there are no α terms as in the previous section. When $k = 1, 2, \ldots, K$, D_{k,t_0} can be calculated by

$$
\begin{aligned}
D_{1,t_0} =& \beta(t_0) \int_0^{t_0} w(x) Q(t_0 - x)\, dx \\
D_{k,t_0} =& \beta(t_{k-1}) \left\{ \sum_{i=1}^{k-2} \left\{ [1-\beta(t_i)] \cdots [1-\beta(t_{k-2})] \int_{t_{i-1}}^{t_i} w(x) Q(t_{k-1} - x)\, dx \right\} \right. \\
& \left. + \int_{t_{k-2}}^{t_{k-1}} w(x) Q(t_{k-1} - x)\, dx \right\}, \text{for } k = 2, \ldots, K
\end{aligned}
$$

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

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

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

$$
L = \prod_{t_0} L(\cdot \mid t_0)
$$

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:

$$
\begin{aligned}
\beta(t) &= \frac{1}{1 + \exp\{-b_0 - b_1(t - \bar{t})\}} \\
w(t) &= w_{max} \cdot \frac{1}{\sqrt{2\pi}\sigma t} \exp\{-(\log t - \mu)^2/(2\sigma^2)\}, \quad t > 0 \\
q(x) &= \frac{\kappa x^{\kappa-1} \rho^{\kappa}}{(1 + (x\rho)^{\kappa})^2}
\end{aligned}
$$

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, and for heavy smokers in lung cancer, it was 0.3 (Wu, Rosner & Broemeling 2005, Liu, Levitt, Riley & Wu 2015). 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. Note that although the integrals in D_{k,t_0} and I_{k,t_0} are not available in closed-form, methods for numerical integration can be applied. Simulations were carried to evaluate the reliability of the proposed likelihood, and the detailed procedure can be found in Wu, Wu, Banicescu & Cariño (2005). 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.

2.3. Key Parameters Estimation When Sensitivity Depends on Sojourn Time

Wu, Cariño & Wu (2008) argued that the screening sensitivity should be a function of both 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 progressing 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 $\beta = \beta(t, s \mid S)$, 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

$$D_{1,t_0} = \int_0^{t_0} w(x) \int_{t_0-x}^{\infty} q(t)\beta(t_0, t_0 - x \mid t)\, dt\, dx \tag{2}$$

$$\begin{aligned} D_{k,t_0} = & \\ & \sum_{i=0}^{k-2} \left\{ \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{\infty} q(t) \left(\prod_{j=i}^{k-2} [1 - \beta(t_j, t_j - x \mid t)] \right) \beta(t_{k-1}, t_{k-1} - x \mid t)\, dt\, dx \right\} \\ & + \int_{t_{k-2}}^{t_{k-1}} w(x) \int_{t_{k-1}-x}^{\infty} q(t)\beta(t_{k-1}, t_{k-1} - x \mid t)\, dt\, dx, \text{ for } k = 2, \ldots, K \end{aligned} \tag{3}$$

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

$$\begin{aligned} I_{k,t_0} = & \sum_{i=0}^{k-1} \left\{ \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{t_k - x} q(t) \left(\prod_{j=i}^{k-1} [1 - \beta(t_j, t_j - x \mid t)] \right) dt\, dx \right\} \\ & + \int_{t_{k-1}}^{t_k} w(x)[1 - Q(t_k - x)]\, dx, \text{ for } k = 2, \ldots, K \end{aligned} \tag{4}$$

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

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

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 & 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

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

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 2, 3 and 4. This method combined with the likelihood in Equation 1 was applied to the Johns Hopkins Lung Project data in Kim & Wu (2016).

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 typically 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 screened group and the control group, the lead time must be adjusted for the screened group, so accurate estimation of the lead time is necessary.

Many researchers have proposed methods to estimate the lead time (Kafadar & Prorok 1994, Kafadar & Prorok 1996, Kafadar & Prorok 2003, Straatman, Peer & Verbeek 1997). Most of these methods assume that the sojourn time follows an exponential distribution. 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.

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 ith 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 screening, say $(j+1)$th screening, 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 screening (at time t_j) can be defined by

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

where $f_{ij}(l)$ is the lead time distribution for ith generation who are detected at $(j+1)$th screening but not before. This $f_{D_j}(l)$ 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

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

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 length of time from entering S_p to being detected at screening t_i, which is a random variable.

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

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

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 screening exams. 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 screening exams, given the screening interval was fixed at 1 year. This suggested a possible stopping rule when designing the screening programs, since continued screenings are not expected to yield any additional benefit. 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.

3.2. Global Lead Time Distribution When Lifetime is Fixed

Wu, Rosner & Broemeling (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 $P(L = 0 \mid D = 1)$ and the conditional density function $f_L(z \mid D = 1)$, for $z \in (0, T - t_0)$:

$$P(L = 0 \mid D = 1) = \frac{P(L = 0, D = 1)}{P(D = 1)} \tag{5}$$

$$f_L(z \mid D = 1) = \frac{f_L(z, D = 1)}{P(D = 1)} \tag{6}$$

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

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

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

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

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

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

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

The validity of the probability calculation can be proved by

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

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 performed 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, Wu, Erwin & Rosner 2011, Wu, Erwin & Rosner 2009*a*).

3.3. Global Lead Time Distribution When Lifetime Is a Random Variable

Wu, Kafadar, Rosner & Broemeling (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

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

where $P(L = 0 \mid D = 1, T = t)$ and $f_L(z \mid D = 1, T = t)$ can be calculated by Equations 5 and 6, and $f_T(t \mid T \geq t_0) = f_T(t)/P(T \geq t_0)$ is the conditional lifetime distribution. The validity of this mixed probability distribution can be proved by

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

The actuarial life table from the United States Social Security Administration was used to estimate the lifetime distribution $f_T(t \mid T \geq t_0)$ (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 $b_N = P(T < N + 1 \mid T \geq N), N = 0, 1, \ldots, 119$. The conditional density can be approximated by

$$f_T(t = t_0 + N \mid T \geq t_0) = b_{t_0+N} \prod_{i=1}^{N} (1 - b_{t_0+i-1})$$

The final lifetime distribution was approximated by a step function, $f_T(t \mid T \geq t_0) \approx f_T(N \mid T \geq t_0)$, 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. Hence, the distribution of the lead time is a weighted average across different lengths of lifetimes. Additional simulations were done in Kendrick, Rai & Wu (2015).

4. Discussion

Accurate estimation of the three key parameters in cancer screening lays a foundation for evaluating the effectiveness of a screening protocol. In particular, all the interesting terms (lead time, rate of over diagnosis, survival benefits, etc.) can be expressed as a function of the sensitivity, sojourn time distribution, and transition density. Estimation of the unobserved lead time is another important topic in cancer screening, as lead time is essential in evaluating the survival benefit of cancer screenings.

In this paper, we reviewed three existing methods for estimating the three key parameters. The stable and nonstable models proposed by Shen & Zelen (1999) provide a way to estimate the key parameters using likelihood-based methods. Under the nonstable model, the transition probability is not constant across different age groups, but assumed the same within each age group. Conversely, the stable model treated the transition probability as a constant over ages. In this approach, the sensitivity was fixed over different age groups and sojourn time was estimated using an exponential distribution. To perform inference allowing sensitivity to vary by age, Wu, Rosner & Broemeling (2005) model sensitivity using a logistic function of age, while assuming a log-normal density for the transition time from S_0 to S_p. While it has been argued that the sensitivity is negatively correlated with the sojourn time (Walter & Day 1983), Wu et al. (2008) extend the sensitivity model in Wu, Rosner & Broemeling (2005) by modeling the sensitivity using both age and the ratio of the time already spent in the preclinical state to the full sojourn time.

Since breast cancer screening programs began before screening of other cancer sites, the earliest developed models are known to be quite accurate for breast cancer. It is now commonly known that sensitivity of mammogram increases as a woman's age increases. The medical explanation is that the breast tissue of younger women is denser and more fibrous compared to that of older women, whose breast tissue is relatively softer and fattier. The previous probability models also

showed the trend (Wu, Rosner & Broemeling 2005, Wu, Wu, Banicescu & Cariño 2005, Chen, Brock & Wu 2010). This may not be generally true for screenings of other cancer sites, such as the widely recommended lung cancer screening using low-dose computed tomography (Liu et al. 2015). With this screening method, the sensitivity does not seem to be affected by age. Whether the sensitivity of the fecal occult blood test for colorectal cancer depends on age is currently uncertain (Prevost, Launoy, Duffy & Chen 1998, Wu et al. 2009*b*).

We also reviewed three methods for estimating the lead time distribution in cancer screening. Prorok (1982) derived the lead time distribution for those detected at the ith exam and used it to determine the stopping rules when designing the screening program. However, the method only considers the screen-detected cases when the lead time is positive. Wu et al. (2007) derived the lead time distribution for the whole cohort, by considering both screen-detected cases and interval cases when the lifetime is fixed. The distribution of the lead time is a mixture of a point mass at zero (for the interval cases) and a piece-wise continuous density function (for the screen-detected cases); the probability calculation was dramatically simplified in this model, and it includes the result of Prorok's as a special case. Wu et al. (2012) extended the model in Wu et al. (2007) to consider lifetime as a random variable. In this circumstance, the lead time distribution is a weighted average of the lead time distribution under different lifetimes. This is the first prospective study: for people at current age and based on the existing data, one can make predictive inference on the distribution of the lead time under different future screening schedules. Thus, one can use this method to infer future outcomes, such as the possibility of early detection, and how early it could be if it is a screen-detected case, and the possibility of no-benefit if it is an interval case, under various future screening schedules. These methods have been applied to estimate the lead time distribution of breast cancer screening (Wu et al. 2007, Shows & Wu 2011, Wu et al. 2012), lung cancer screening (Wu et al. 2011, Jang, Kim & Wu 2013) and colorectal cancer (Wu et al. 2009*a*).

Effectiveness of screening is constantly debated. Questions regarding the efficient design of cancer screening programs have arisen, such as at what age to start a screening exam and how frequently patients should be re-screened. For example, there has been recent controversy about whether mammography in breast cancer screening benefits women in their 40s. Here, we reviewed several statistical methods and hope effective and novel statistical evaluation of screening protocols will be an integral part of this debate.

Acknowledgements

We authors want to thank Dr. Adriana Pérez for inviting us to submit this review article, and we also thank her for translating our abstract into Spanish.

[Received: October 2016 — Accepted: April 2017]

References

Chen, Y., Brock, G. & Wu, D. (2010), 'Estimating Key Parameters in Periodic Breast Cancer Screening-Application to the Canadian National Breast Screening Study Data', *Cancer Epidemiology* **34**(4), 429–433.

Jang, H., Kim, S. & 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. & 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. & 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. & 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. & 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. & 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., 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 and Biostatistics* **6**(5).

Prevost, T. C., Launoy, G., Duffy, S. W. & Chen, H. H. (1998), 'Estimating Sensitivity and Sojourn Time in Screening for Colorectal Cancer: A Comparison of Statistical Approaches', *American Journal of Epidemiology* **148**(6), 609–619.

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. & Zelen, M. (2001), 'Testing the Independence of Two Diagnostic Tests', *Biometrics* **57**(4), 1009–1017.

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

Shows, J. & Wu, D. (2011), 'Inferences for the Lead Time in Breast Cancer Screening Trials under a Stable Disease Model', *Cancers* **3**(2), 2131 – 2140.

Straatman, H., Peer, P. G. M. & 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

Walter, S. D. & Day, N. E. (1983), 'Estimation of the Duration of A Preclinical Disease State Using Screening Data', *American Journal of Epidemiology* **118**(6), 865–886.

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

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

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

Wu, D., Rosner, G. L. & 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. & Cariño, R. L. (2005), 'Simulation Procedure in Periodic Cancer Screening Trials', *Journal of Modern Applied Statistical Methods* **4**(2), 522–527.

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