Statistica Sinica **24** (2014), 815-831
doi:http://dx.doi.org/10.5705/ss.2012.067

INFERENCE OF LONG TERM EFFECTS AND OVERDIAGNOSIS IN PERIODIC CANCER SCREENING

Dongfeng Wu, Karen Kafadar and Gary L. Rosner

University of Louisville, Indiana University and
The Sidney Kimmel Comprehensive Cancer Center

Abstract: We develop a probability model for evaluating long-term effects due to regular screening. People who take part in cancer screening are divided into four mutually exclusive groups: *True-early-detection, No-early-detection, Overdiagnosis*, and *Symptom-free-life*. For each case, we derive the probability formula. Simulation studies using the HIP (Health Insurance Plan for Greater New York) breast cancer study's data provide estimates for these probabilities and corresponding credible intervals. These probabilities change with a person's age at study entry, screening frequency, screening sensitivity, and other parameters. We also allow human lifetime to be subject to a competing risk of death from other causes. The model can provide policy makers with important information regarding the distribution of individuals participating in a screening program who eventually fall into one of the four groups.

Key words and phrases: Overdiagnosis, sensitivity, sojourn time, symptom free life, transition probability, true early detection.

1. Introduction

Cancer screening programs can be effective in detecting tumors early, before symptoms are present. A challenge remains as to how to evaluate the long-term effects due to continued screening. For example, how should the probability of true early detection be estimated in regular screening? Will regular screening exams result in a greater chance of overdiagnosis? How should the probabilities of no-early-detection and the probability of overdiagnosis be estimated?

Some research has been done in the area of overdiagnosis, the diagnosis of cancer that never would have become symptomatic during a person's lifetime (Badgwell et al. (2008), Duffy et al. (2008), Davidov and Zelen (2004), Gotzsche et al. (2009), Jorgensen and Gotzsche (2009), Zahl, Mahlen, and Welch (2008)). Most of the research has been based on observational studies, however; results from the research may be biased due to inadequate probability modeling. The estimated percentage of breast cancer overdiagnosis in women varies widely, from 7% (Zackrisson et al. (2006)) to 52% (Jorgensen and Gotzsche (2009)). With

Table 1. Definition of long-term outcomes/events in screening.

diagnosis status	ultimate lifetime disease status	
	no symptom before death	symptoms before death
not-screen-detected	Symptom-free-life	No-early-detection
screen-detected	Over-diagnosis	True-early-detection

controversy concerning the benefit of regular screening, we need a better understanding of the factors affecting the risk of overdiagnosis.

Instead of modeling the effect of overdiagnosis alone, we want to address the long-term effect attributable to regular screening for the whole cohort. We assume that a woman is asymptomatic and without a history of breast cancer before she takes her first screening exam, and we divide all women into four mutually exclusive groups: *True-early-detection, No-early-detection, Overdiagnosis*, and *Symptom-free-life* based on their diagnosis status and whether symptoms would have appeared before death, see Table 1. Eventually every participant falls into one of the four groups, and here are the definitions.

- *Group* 1: **Symptom-free-life (SympF)**. A woman in Group 1 took part in screening exams, but breast cancer was never detected and ultimately she died of other causes.
- *Group* 2: **No-early-detection (NoED)**. A woman in Group 2 took part in screening exams, but disease manifested itself clinically and was not detected by scheduled screening exams.
- *Group* 3: **True-early-detection (TrueED)**. A woman in Group 3 was diagnosed with breast cancer at a scheduled screening exam and her clinical symptoms would have appeared before her death.
- *Group* 4: **Overdiagnosis (OverD)**. A woman in Group 4 was diagnosed with breast cancer at a scheduled screening exam but her clinical symptoms would NOT have appeared before her death.

The more familiar term "interval case" falls in *Group 2*; however, as defined above, *Group 2* includes all possible interval cases with the lifetime being treated as a random variable.

The remainder of the paper is organized as follows. In Section 2 we propose a probability model and derive the probabilities for each of the four cases, treating the duration of human lifetime as a random variable, with the cause of death subject to other competing risks. In Section 3, we present extensive simulation results for different scenarios, and in Section 4, we apply our method to the Health Insurance Plan for Greater New York (HIP) breast cancer screening data, and present simulation results for different screening schedules. Policy makers

may use these probabilities to help assess different screening strategies, such as changing screening frequencies, determining age to start screening, etc. We conclude with a discussion in Section 5.

2. Probability and Calculations

We propose a probability model and derive the probability for cases in each of the four groups. We assume the commonly followed disease progression model in which the disease develops through three states $S_0 \to S_p \to S_c$: S_0 refers to the disease-free state or the state in which the disease cannot be detected; S_p refers to the preclinical disease state, in which an asymptomatic individual unknowingly has disease that a screening exam can detect; and S_c refers to the disease state at which the disease manifests itself in clinical symptoms.

Consider a cohort of initially asymptomatic individuals who enroll in a screening program. Let $\beta(t)$ be the sensitivity at age t, the probability that the screening exam is positive given that the individual is in the preclinical state. Take $w(t)dt$ 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 survivor function of the sojourn time. Throughout, the time variable t represents an individual's age at time of screening, and T represents a person's lifetime, a continuous random variable with a probability density function $f_T(t)$. Let

$$A = \{\text{A woman is asymptomatic of breast cancer before and at } t_0\}.$$

We can calculate the conditional probability that *no* breast cancer was found before age t_0, given that one's lifetime T exceeds t_0, which can arise as one of only two mutually exclusive events: either (i) she remains in the disease-free state through age t_0, the probability of which is $1 - \int_0^{t_0} w(x)dx$; or (ii) she enters state S_p before t_0 but remains in S_p long enough that no symptoms present before t_0, the probability of which is $\int_0^{t_0} w(x)Q(t_0 - x)dx$. Then, the probability of A is the sum of the two probabilities (i) and (ii):

$$P(A|T \geq t_0) = 1 - \int_0^{t_0} w(x)dx + \int_0^{t_0} w(x)Q(t_0 - x)dx. \tag{2.1}$$

Keeping in mind that a woman is asymptomatic at t_0 (event A), we first derive the probability of each case when there is only one screening exam during a woman's lifetime, given that the lifetime $T = t$ is a fixed value; then we extend this model to the case when her lifetime is a random variable, $T \sim f_T(t)$. Finally, we derive the general results for these probabilities when there are multiple screening exams and the lifetime T is a random variable.

2.1. The probability of cases in each group: one exam only

Suppose a woman undergoes one screening exam at age t_0. We first derive the conditional probability of each case given her (fixed) lifetime $T = t(> t_0)$.

For a Group 1 case, a woman who never has detectable breast cancer during her lifetime can follow one of three trajectories: (a) she never progressed out of the disease-free state S_0 throughout her lifetime; (b) she entered the preclinical state S_p before t_0, her cancer was not detected, and her sojourn time was so long that no clinical symptom appeared before her death; (c) she entered the preclinical state S_p after t_0 and had a long sojourn time, so that no symptom appeared before her death. Hence the conditional probability given her lifetime $T = t(> t_0)$ is:

$$P(\textit{Case } 1\text{: SympF}, A|T = t) = 1 - \int_0^t w(x)dx + (1-\beta_0)\int_0^{t_0} w(x)Q(t-x)dx + \int_{t_0}^t w(x)Q(t-x)dx. \quad (2.2)$$

We are assuming that the sojourn time distribution does not depend on the age of entry into S_p.

For a Group 2 case, a woman whose cancer became symptomatic and was thereby found in (t_0, T), either (a) she entered S_p before t_0 and was missed by the screening exam, or (b) she entered the preclinical state after t_0. In either situation, her sojourn time in S_p was shorter than $(t - x)$, where x is her age entering S_p. Hence the conditional probability is

$$P(\textit{Case } 2\text{: NoED}, A|T = t) = (1-\beta_0)\int_0^{t_0} w(x)[Q(t_0-x) - Q(t-x)]dx + \int_{t_0}^t w(x)[1 - Q(t-x)]dx. \quad (2.3)$$

For a Group 3 case, a woman is truly detected early by taking scheduled exam, her cancer must have been diagnosed at t_0, and her symptoms would have appeared before death. That is, she must have entered S_p at some age x before t_0, and her sojourn time was between $(t_0 - x)$ and $(T - x)$. Hence,

$$P(\textit{Case } 3\text{: TrueED}, A|T = t) = \beta_0 \int_0^{t_0} w(x)[Q(t_0-x) - Q(t-x)]dx. \quad (2.4)$$

For a Group 4 case, the case of overdiagnosis, she was diagnosed at t_0 but her symptoms would not have appeared before death. That is, she must have entered S_p at some age $x(< t_0)$, but her sojourn time extended to beyond time $(T - x)$. Hence,

$$P(\textit{Case } 4\text{: OverD}, A|T = t) = \beta_0 \int_0^{t_0} w(x)Q(t-x)dx. \quad (2.5)$$

The probability of each case when the lifetime T is a random variable and $T \geq t_0$ can be obtained as

$$P(\text{Case } i, A|T \geq t_0) = \int_{t_0}^{\infty} P(\text{Case } i, A|T = t) f_T(t|T \geq t_0) dt, \quad i = 1, 2, 3, 4. \tag{2.6}$$

where the conditional pdf $f_T(t|T \geq t_0)$ is

$$f_T(t|T \geq t_0) = \begin{cases} \frac{f_T(t)}{P(T>t_0)} = \frac{f_T(t)}{1-F_T(t_0)}, & \text{if } t \geq t_0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

By adding(2.2) to (2.5), one can verify that for any $t > t_0$,

$$\sum_{i=1}^{4} P(\text{Case } i, A|T = t) = 1 - \int_0^{t_0} w(x)dx + \int_0^{t_0} w(x)Q(t_0 - x)dx = P(A|T \geq t_0). \tag{2.8}$$

Since the right hand side of (2.8) does not depend on t, we have

$$\begin{aligned} \sum_{i=1}^{4} P(\text{Case } i, A|T \geq t_0) &= \int_{t_0}^{\infty} [\sum_{i=1}^{4} P(\text{Case } i, A|T = t)] f_T(t|T \geq t_0) dt \\ &= P(A|T \geq t_0). \end{aligned} \tag{2.9}$$

This implies

$$\sum_{i=1}^{4} P(\text{Case } i|A, T \geq t_0) = \sum_{i=1}^{4} \frac{P(\text{Case } i, A|T \geq t_0)}{P(A|T \geq t_0)} = 1. \tag{2.10}$$

2.2. The probability of cases in each group: multiple exams

We generalize this idea to any number of screening exams. Suppose an initially asymptomatic individual undergoes K screening exams, occurring at ages $t_0 < t_1 < \cdots < t_{K-1}$. We define $t_{-1} = 0$. The conditional probability of a case in any one of the four groups, given that her lifetime is $T = t_K (> t_{K-1})$, can be generalized as follows.

A Group 1 case where clinical breast cancer never occurs in her lifetime, can arise as any one of $(K + 2)$ disjoint events: (a) she remained in the disease-free state S_0 throughout her lifetime, the probability of which is $1 - \int_0^{t_K} w(x)dx$. (b) she entered the preclinical state S_p when she was between ages t_{j-1} and $t_j, j = 0, \ldots, K - 1$, was not detected by the following $(K - j)$ exams, and had a long sojourn time, so no symptom appeared before her death (K disjoint events). (c) she entered S_p after t_{K-1} with no symptoms before her death. When we add the probability of these events together, the probability is

$$
\begin{aligned}
&P(\text{Case } 1, A|T = t_K) \\
&\quad = 1 - \int_0^{t_K} w(x)dx + \int_{t_{K-1}}^{t_K} w(x)Q(t_K - x)dx \\
&\qquad + \sum_{j=0}^{K-1} (1-\beta_j)\cdots(1-\beta_{K-1}) \int_{t_{j-1}}^{t_j} w(x)Q(t_K - x)dx. \qquad (2.11)
\end{aligned}
$$

For a Group 2 case, we calculate the probability of no early detection by defining $I_{K,j}$ as the probability of being an interval case in the interval (t_{j-1}, t_j) in a sequence of K screening exams. Thus

$$
P(\text{Case } 2, A|T = t_K) = I_{K,1} + I_{K,2} + \cdots + I_{K,K}, \qquad (2.12)
$$

where

$$
\begin{aligned}
I_{K,j} &= \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)[Q(t_{j-1} - x) - Q(t_j - x)]dx \\
&\quad + \int_{t_{j-1}}^{t_j} w(x)[1 - Q(t_j - x)]dx, \qquad \text{for all } j = 1, \ldots, K. \qquad (2.13)
\end{aligned}
$$

For more details, see Wu, Rosner, and Broemeling (2007).

A Group 3 case, true early detection, can arise as one of K disjoint events depending on her age at diagnosis by screening, namely, at $t_j, j = 0, 1, \ldots, K-1$. If she is diagnosed at t_j, then she must have entered the preclinical state S_p before t_j, and missed the previous exams, and her sojourn time must have been at least $(t_j - x)$ and at most $(t_K - x)$, where x represent the onset time of the preclincal state. Therefore

$$
\begin{aligned}
&P(\text{Case } 3, A|T = t_K) \\
&= \sum_{j=1}^{K-1} \beta_j \Big\{ \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)[Q(t_j - x) - Q(t_K - x)]dx \\
&\quad + \int_{t_{j-1}}^{t_j} w(x)[Q(t_j - x) - Q(t_K - x)]dx \Big\} + \beta_0 \int_0^{t_0} w(x)[Q(t_0 - x) - Q(t_K - x)]dx. \qquad (2.14)
\end{aligned}
$$

A Group 4 case, overdiagnosis, also can arise as one of K disjoint events. She might have been diagnosed at the jth exam, but her symptoms did not appear before her death. Hence,

$$
\begin{aligned}
&P(\text{Case } 4, A|T = t_K) \\
&\quad = \sum_{j=1}^{K-1} \beta_j \Big\{ \sum_{i=0}^{j-1} (1-\beta_i)\cdots(1-\beta_{j-1}) \int_{t_{i-1}}^{t_i} w(x)Q(t_K - x)dx \\
&\qquad + \int_{t_{j-1}}^{t_j} w(x)Q(t_K - x)dx \Big\} + \beta_0 \int_0^{t_0} w(x)Q(t_K - x)dx. \qquad (2.15)
\end{aligned}
$$

We can verify that for any screening number $K \geq 1$, it is still true that

$$\sum_{i=1}^{4} P(\text{Case } i, A|T = t_K) = 1 - \int_0^{t_0} w(x)dx + \int_0^{t_0} w(x)Q(t_0 - x)dx = P(A|T \geq t_0). \tag{2.16}$$

For an individual currently at age t_0, her lifetime is not fixed but random, so it is unrealistic to consider the future number of exams K to be a fixed value. However, if she plans to follow a future screening schedule, such as $t_0 < t_1 < \ldots$, then $K = n$ if $t_{n-1} < T < t_n$, the screening number $K = K(T)$ is a random variable, changing with the lifetime T. The probability of each case when her lifetime T is longer than t_0 can be obtained as the weighted average

$$P(\text{Case } i, A|T \geq t_0) = \int_{t_0}^{\infty} P(\text{Case } i, A|K = K(T), T = t) f_T(t|T \geq t_0)dt, i = 1, 2, 3, 4, \tag{2.17}$$

here $f_T(t|T \geq t_0)$ was defined in (2.7). The probability $P(\text{Case i}, A|K = K(T), T = t)$ was derived in (2.11)−(2.15).

Again, it is easy to verify by (2.16) that for any future screening schedule when the lifetime T is random,

$$\sum_{i=1}^{4} P(\text{Case } i|A, T \geq t_0) = 1. \tag{2.18}$$

3. Simulation Study

We conducted extensive simulation studies using the method derived in Section 2. Since the probability of each case is a function of age at the initial screening, the screening interval, the sensitivity, the sojourn time in the preclinical state, the transition probability from the disease-free to the preclinical state, and the human lifetime, we want to explore the effects of these factors on the probability of each outcome, and also explore how the proportion of true-early-detection and over-diagnosis change among the screen-detected cases due to these factors. We selected the following scenarios for simulation: age at initial screening $t_0 = 40, 50, 60$, screening interval $\Delta = 6, 12, 24$ months, screening sensitivity $\beta = 0.3, 0.7, 0.9$, we picked $\beta = 0.3$ assuming some screening test has very low sensitivities (Davidov and Zelen (2004)). The transition probability density were chosen to be either a Gamma or a log Normal pdf, with a single mode at about 60 years old and an upper limit of 20%, which is applicable to different kinds of cancer. The sojourn time distribution were chosen to be either an exponential or a log-logistric pdf, the parameters were carefully chosen so that the mean sojourn time was 2, 5, 10, and 20 years.

The number of screens $K = K(T) = \lceil (T - t_0)/\Delta \rceil$ (the largest integer that is less than or equal to $(T - t_0)/\Delta$) is a function of the lifetime T, and hence is a random vaiable in the simulation. For the lifetime distribution, we used the acturial life table from the Social Security Administration (SSA), published online at `http://www.ssa.gov/OACT/STATS/table4c6.html`. The Period Life Table was made available in 2006, and was reviewed and updated on April 19, 2010. It is based on mortality, and it provides the probability of death within one year from age 0 to age 119 for both males and females. We derived the conditional lifetime distribution $f_T(t|T > t_0)$ based on the current life table (See Section 4 and Figure 1 in Wu et al. (2012)).

The results for different initial age group are very similar, so we only report the case of $t_0 = 50$. The results are also similar if the transition probability density is a Gamma or a log Normal pdf, so we only report the case of log Normal. The results have little difference when using male or female lifetime density, so we reported the case of females here. However, there were obvious differences when the sojourn time was an exponential or a log logistic pdf, even though the mean sojourn times were the same. We report the results in Tables 1 and 2 in the appendix in the supplementary website, the corresponding results when the sojourn time distribution is either the log logistic or the exponential.

In the simulation results, we can see clearly that the mean sojourn time plays the most important rule in the case of overdiagnosis. For example, in the last column of Table 1, the proportion of overdiagnosis could be as high as 38% among the screen-diagnosed cases if the mean sojourn time is 20 years long, and it is around 20% if the mean sojourn time is 10 years long, and it is only about 4-9% when the mean sojourn time changes from 2 to 5 years. In Table 2 in the web-appendix, when the sojourn time is exponentially distributed this pattern is more dramatic, the probability of overdiagnosis could be as high as 43%.

The screening sensitivity affects the ratio of the no-early-detection and the true-early-detection: when sensitivity is higher, the probability of true-early-detection is highter, and the probability of no-early-detection is lower. However, the sensitivity only has a small effect in the percentage of true-early-detection and overdiagnosis among the screen-detected cases, when the sojourn times are the same: the percentage of overdiagnosis increases slightly (<0.1%) when the sensitivity increases from 0.3 to 0.9.

The screening interval also plays a role in these proabilities: when the screening interval is longer, the probability of no-early-detection is larger, the probability of true-early-detection is smaller, and the probability of overdiagnosis is slightly smaller (<0.1% in 5th column). The case of symptom-free-life is pretty stable in all the simulations, it is about 86-88% for the whole population.

The transition proability density $w(t)$ is surely important, but in this simulation, we limit it to the situation where the density has a single peak around

age 60, based on common sense. We consider the log logistic pdf a more suitable candidate for the sojourn time distribution compared with the exponential, because the exponential density has its mode at 0 and with a constant hazard rate that is not realistic.

4. A Projection of Benefits Using the HIP Data

We applied our method to the Health Insurance Plan for the Greater New York (HIP) data (Shapiro et al. (1988)).

4.1. Bayesian inference of the probability

The probability for each of the four cases is a function of the sensitivity $\beta(t)$, the transition probability density $w(t)$, the sojourn time distribution $q(t)$, a person's age at first screening t_0, and her future screeing interval Δ, according to the results in Section 2. The age-dependent sensitivity $\beta(t)$, the age-dependent transition probability $w(t)$, and the sojourn time distribution $q(\cdot)$ were estimated from the HIP data in Wu, Rosner, and Broemeling (2005). The parametric models for $\beta(t), w(t)$, and $q(x)$ were

$$\beta(t) = \frac{1}{1 + \exp\{-b_0 - b_1(t-m)\}}, \tag{4.1}$$

$$w(t) = \frac{0.2}{\sqrt{2\pi}\sigma t} \exp\left\{-\frac{(\log t - \mu)^2}{2\sigma^2}\right\}, \tag{4.2}$$

$$q(x) = \frac{\kappa x^{\kappa-1}\rho^{\kappa}}{[1 + (x\rho)^{\kappa}]^2}, \quad \kappa > 0, \rho > 0, \tag{4.3}$$

where m is the average age of women at the study entry. The 0.2 in the $w(t)$ is the upper limit of making a transition from the disease-free state to the preclinical state. The unknown parameters in this model are $\theta = (b_0, b_1, \alpha_1, \alpha_2, \kappa, \rho)$. We generated a posterior random sample of 2000 by Markov Chain Monte Carlo (MCMC) from the posterior distribution (Wu, Rosner, and Broemeling (2005)). We used these Bayesian posterior samples (θ_j^*) in the inference.

Using the HIP data, the posterior predictive probability of each case can be estimated as

$$\begin{aligned} P(\text{Case } i | T \geq t_0, A, HIP) &= \int P(\text{Case } i, \theta | T \geq t_0, A, HIP) d\theta \\ &= \int P(\text{Case } i | T \geq t_0, A, \theta) f(\theta | HIP) d\theta \\ &\approx \frac{1}{n} \sum_{j=1}^{n} P(\text{Case } i | T \geq t_0, A, \theta_j^*), \end{aligned} \tag{4.4}$$

Table 2. A projection of breast cancer screening effects using the HIP data.

Δ^a	P^b(SympF)	P(NoED)	P(TrueED)	P(OverD)
Age at initial screen $t_0 = 40$				
6 mo.	89.66(1.34)	0.99(0.47)	8.83(1.58)	0.37(0.25)
12 mo.	89.71(1.35)	2.48(0.70)	7.34(1.31)	0.32(0.24)
18 mo.	89.75(1.35)	3.79(0.94)	6.03(1.01)	0.28(0.23)
24 mo.	89.78(1.35)	4.77(1.12)	5.04(0.80)	0.25(0.22)
30 mo.	89.81(1.36)	5.51(1.25)	4.31(0.67)	0.23(0.21)
Age at initial screen $t_0 = 50$				
6 mo.	91.15(1.28)	0.74(0.47)	7.70(1.46)	0.38(0.25)
12 mo.	91.21(1.28)	1.95(0.71)	6.49(1.25)	0.33(0.24)
18 mo.	91.25(1.29)	3.06(0.91)	5.38(0.99)	0.29(0.23)
24 mo.	91.28(1.29)	3.91(1.07)	4.52(0.81)	0.26(0.22)
30 mo.	91.30(1.29)	4.55(1.18)	3.88(0.69)	0.23(0.21)
Age at initial screen $t_0 = 60$				
6 mo.	93.15(1.02)	0.52(0.43)	5.90(1.16)	0.39(0.26)
12 mo.	93.21(1.02)	1.39(0.63)	5.03(1.02)	0.33(0.25)
18 mo.	93.25(1.02)	2.21(0.78)	4.21(0.85)	0.29(0.24)
24 mo.	93.28(1.02)	2.86(0.89)	3.57(0.72)	0.26(0.23)
30 mo.	93.30(1.03)	3.34(0.96)	3.08(0.63)	0.24(0.22)

[a] $\Delta = t_i - t_{i-1}$ is the time interval between screens.
[b] The mean probability and its standard error (in parenthesis) are reported as percentages in the table.

where θ_j^* is the random sample drawn from the posterior distribution $f(\theta|HIP)$ and n is the posterior sample size (Wu, Rosner, and Broemeling (2005)). The last step is the Monte Carlo simulation.

4.2. Results

We applied (4.4) to the 2000 MCMC posterior samples, to conduct Bayesian inference in the case of a program consisting of periodic screening exams for three hypothetical cohorts of asympomatic women. The three cohorts had initial ages of 40, 50, and 60 in the first screening exam.

For each group, we examined various screening frequencies, with screening interval $\Delta = 6$, 12, 18, 24, and 30 months. The number of screens $K = K(T) = \lceil (T - t_0)/\Delta \rceil$ is again a function of the lifetime T, therefore it is a random variable in the simulation. For the lifetime distribution, we used the conditional lifetime density derived from the acturial life table from the SSA for females as in Section 3. The conditional probabilities of each of the four cases $P(\text{Case } i|A, T \geq t_0, HIP)$ are reported in Table 2.

For all three age groups the probability of "Overdiagnosis" is very small. For the 12-month screening interval, it is 0.32%, 0.33%, and 0.33%, respectively,

for ages at first screening exam 40, 50, and 60 years old. This probability barely changes when age at initial screening exam increases. It decreases as the screening time interval (Δ) increases.

The probability of "True-early-detection" is 7.34%, 6.49%, and 5.03% respectively for these cohorts, if screening is annual. This probability also decreases as the screening time interval increases. The probability of "True-early-detection" is slightly lower when the initial screen age is 60; however, there is very little difference for the other age groups.

The probability of "No-early-detection" is 2.48%, 1.95%, and 1.39% for the 12-month screening schedule if the woman initiates screening at ages 40, 50, and 60. It increases as the screening interval increases; it decreases slightly when age at the initial screen increases.

The probability of "Symptom-free-life" is very high. It increases from about 89% to 93% when the initial screening age increases from 40 to 60. It is comparatively stable when the screening interval changes within each age group. Overall, the difference of the corresponding probabilities is smaller between the age groups 40 and 50 than that between the age groups 50 and 60.

Boxplots of the results for the probability of each case when $t_0 = 50$ are given in Figure 1. The boxplots for age groups 40 and 60 are very similar to those in age 50, so we omit them. In Figure 1, we see that the probability of "Symptom-free-life" and the probability of "Over-diagnois" are either stable, or barely change with the screening time interval. The probability of "No-early-detection" increases monotonically with the screening time interval; while the probability of "True-early-detection" decreases monotonically with the length of the screening time interval.

If we calculate the conditional probability for cases 2, 3, and 4, given that she is a diagnosed cancer case, either an interval clinical incident case or a screen-detected case, the percentage of overdiagnosis is 3.93%, 4.62%, and 6.02% for the 6-months screening group if starting age is 40, 50, and 60. The conditional probability of "True-early-detection" given it is a diagnosed case decreases dramatically when the screening interval Δ increases; it changes from 86% to 43% in the 40-year-old group, from 87% to 45% in the 50-year-old group, and from 86% to 46% in the 60-year-old group. For the same screening interval, the probability of "true-early-detection" slightly increases with the initial screening age. The conditional probability of "No-early-detection" increases within each age group as the screening interval increases, while the conditional probability of "Over-diagnosis" decreases slightly within each age group. See Table 3 for details.

The probabilities and 95% HPD intervals of "True-early-detection" and "Overdiagnosis" given it is a screen-detected case, are listed in Table 4. The percentage of "Overdiagnosis" increases from 4.41% to 5.05% in the 40-year-old

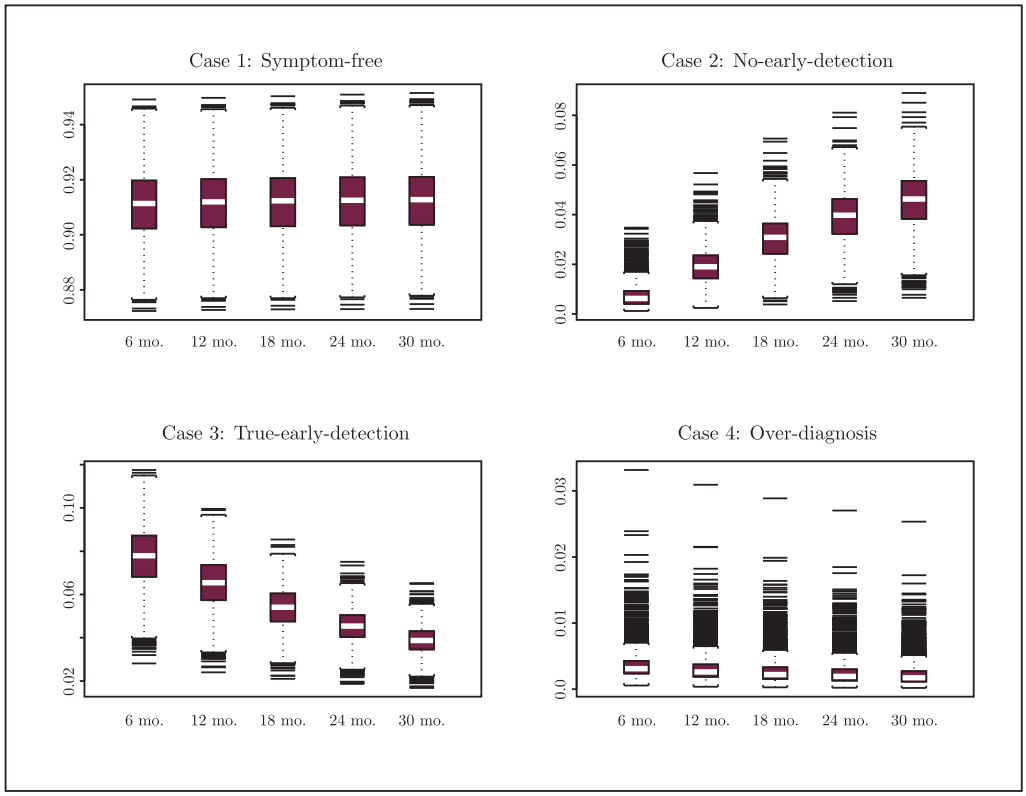

Figure 1. The boxplot of the estimated probability for each case with $t_0 = 50$.

age group. This percentage increases from 5.09% to 5.69% in the group whose initial screening exam is at age 50, and it increases from 6.56% to 7.08% in the 60-year-old group. In summary, the probability of "Overdiagnosis" is much lower than we had expected; while the probability of "True-early-detection" is often above 93% and is higher than we had expected. The length of the 95% HPD interval for these two probabilities (percentages) increases as the screening interval increases.

5. Discussion

This study provides a way to assess the overall performance of the screening program in the long-term. We separated all initially asymptomatic participants in a screening program into four mutually exclusive groups: symptom-free-life, no-early-detection, true-early-detection, and overdiagnosis. Analyses such as this one can provide policy makers with estimates of the probability of true-early-detection, overdiagnosis, and other outcomes that result from a periodic screening program. We used a Bayesian approach, because this can incorporate uncertainty

Table 3. The estimated probability given that it is a diagnosed cancer case.

Δ	P^c(NoED$\|D^d$)	P(TrueED$\|D$)	P(OverD$\|D$)
Age at initial screen $t_0 = 40$			
6 mo.	9.87	86.20	3.93
12 mo.	24.44	72.12	3.44
18 mo.	37.21	59.73	3.06
24 mo.	46.97	50.27	2.76
30 mo.	54.29	43.18	2.53
Age at initial screen $t_0 = 50$			
6 mo.	8.49	86.89	4.62
12 mo.	22.17	73.78	4.05
18 mo.	34.72	61.67	3.61
24 mo.	44.48	52.25	3.27
30 mo.	51.88	45.11	3.00
Age at initial screen $t_0 = 60$			
6 mo.	7.73	86.24	6.02
12 mo.	20.48	74.22	5.29
18 mo.	32.60	62.68	4.72
24 mo.	42.20	53.52	4.27
30 mo.	49.55	46.53	3.92

[c]The estimated conditional probability was calculated as $p_i^*/(p_2^* + p_3^* + p_4^*), i = 2, 3, 4$, for each of the 2,000 posterior samples, then averaged. It is in percentage.

[d]The event $D = \{$Diagnosed cases: including both interval-incident and screen-detected cases$\}$.

easily, and make it easy to calculate the variations and the credible invervals of the probability (or percentage).

In November 2009, the U.S. Preventive Services Task Force (USPSTF) announced new recommendations regarding mammography screening: women should start screening at age 50, rather than at age 40, and that women between the ages of 50-74 should undergo screening mammography every other year, instead of every year. Based on the HIP study group data, our probability of "Symptom-free-life" is very high, about 90% for all participants, while the probability of "Over-diagnosis" is very low, less than 0.4% among all participants (Table 2). The estimate of "Over-diagnosis" is 5.5% among those diagnosed early in the 50-year-old cohort if screenings were taken every other year (Table 4, with a 95% HPD interval of (1.45%, 21.84%). These estimates are based on the HIP data, and the HIP study was carried out in the 1960s. The sensitivity of mammography might well have been lower than with today's screening modalities. However, from our simulation study in Section 3, sensitivity only has a slight effect on the percentage of over-diagnosis. The life expectancy of the US women

Table 4. The estimated probability for the screen-detected cases (with 95% credible interval).

Δ	P[d](TrueED$\|ScrD$[e])	P(OverD$\|ScrD$)
	Age at initial screen $t_0 = 40$	
6 mo.	95.59 (82.78, 98.62)	4.41 (1.38, 17.22)
12 mo.	95.47 (81.49, 98.76)	4.53 (1.24, 18.51)
18 mo.	95.29 (80.75, 98.79)	4.71 (1.21, 19.25)
24 mo.	95.12 (79.91, 98.79)	4.88 (1.21, 20.09)
30 mo.	94.95 (79.46, 98.78)	5.05 (1.22, 20.54)
	Age at initial screen $t_0 = 50$	
6 mo.	94.91 (80.76, 98.34)	5.09 (1.66, 19.24)
12 mo.	94.84 (79.38, 98.53)	5.16 (1.47, 20.62)
18 mo.	94.68 (78.97, 98.57)	5.32 (1.43, 21.03)
24 mo.	94.50 (78.16, 98.55)	5.50 (1.45, 21.84)
30 mo.	94.31 (77.65, 98.52)	5.69 (1.48, 22.35)
	Age at initial screen $t_0 = 60$	
6 mo.	93.44 (76.44, 97.74)	6.56 (2.26, 23.56)
12 mo.	93.43 (75.37, 97.97)	6.57 (2.03, 24.63)
18 mo.	93.28 (75.04, 98.04)	6.72 (1.96, 24.96)
24 mo.	93.10 (74.44, 98.05)	6.90 (1.95, 25.56)
30 mo.	92.92 (73.81, 98.04)	7.08 (1.96, 26.19)

[d]The estimated conditional probability was calculated as $p_i^*/(p_3^*+p_4^*), i = 3, 4$, for each of the 2,000 posterior samples, then averaged. It is a percentage.

[e]The event $ScrD = \{$Screen-detected case$\}$.

in 2006 is 80.4 years, versus 73.1 in the 1960s. However, we cannot find similar lifetable of the 1960s. We think that the over-diagnosis rate might be slightly higher if the life expectancy is shorter. We hope to investigate this question with data from more recent studies.

We checked the NIH SEER database, the lifetime risk for breast cancer for both invasive and in-situ is 7.53% for all races, with a 95% CI (7.49%, 7.57%) (National Institute of Health (2010)). Our esimated probability of "Symptom-free-life" is close to one minus the lifetime risk. A generally accepted lifetime risk of breast cancer is 1 in 9 (or 11%), with almost all the risk after age 40. Our estimated probability of "Symptom-free-life" is about 89% for the 40-year-old age group, which is compatible to the accepted lifetime risk.

We also did more simulations, with the upper limit as 0.3 for the $w(t)$ in (4.2). That is, assuming a 30% transition from the disease-free state to the preclinical state, we first ran the MCMC using the HIP data, then applied our probabilty model to the collected posterior samples. The results were similar to

those in Table 2, suggesting that the model is reasonably robust to small changes in $w(t)$.

Zackrisson et al. (2006) compared the cumulative breast cancer incidence rates in the screening and the control groups for the same 15-years follow-up period in the Sweden breast cancer trial, and they estimated that overdiagnosis is about 7% for invasive and 10% for both invasive and in situ. Our results are comparable to theirs. Duffy et al. (2008) showed how complicated it is to estimate the degree of overdiagnosis in breast cancer screening because of the lead time bias and other factors. Jorgensen and Gotzsche (2009) estimated overdiagnosis by comparing the incidence trend before and after the screening. Their pre-screening periods were mainly in the 1970s and 1980s, with one exception of the Norway data (1980-1994), while the post-screening periods were after 1993. Their estimate of overdiagnosis was 52%, with a 95% C.I of (46%, 58%). However, due to a higher prevalence of hormone replacement therapy (HRT) in the 1990s, the breast cancer incidence trend was dramatically increasing in the western world until 2000, and then was decreasing afterward, even without screening. Their inference may be flawed by failing to take this trend into account. Zahl, Mahlen, and Welch (2008) inferred from a comparison of pre- and post-mammography programs that some cancers may regress to normal if left untreated. These authors' conclusions resulted from research using observational studies where problems are well known. Results based on one study cannot be extended to other scenarios, and inferences made on observational studies usually need long follow-up periods to collect incidence data from the study and the control arms. This is not cost effective, and sometimes data may not be available (Wu and Perez (2011)).

Davidov and Zelen (2004) introduced two measures of overdiagnosis: individual overdignosis and schedule overdiagnosis. Individual overdiagnosis represents the risk of overdiagnosis in an upcoming exam, given a person's screening history. Schedule overdiagnosis, on the other hand, is a property of the particular schedule used by the screening program and reflects the overall risk of overdiagnosis with that program. They first used a forward recurrence time model to derive the conditional probability of overdiagnosis for an i-th generation person who was diagnosed at the j-th exam ($j > i$) at age t_j. Next, they summarized these probabilities to get the conditional probability of overdiagnosis at the j-th exam for a fixed screening schedule $\tau = (\tau_1, \ldots, \tau_m)$ with m exams. Then, they applied their method to prostate cancer screening data. Under different assumptions of the mean sojourn time, sensitivity, and the number of screenings, they found that the risk of overdiagnosis for prostate cancer was very high. For example, this risk was over 50% if an individual was diagnosed at age 80, had a mean sojourn time of 10 years, and was screened every 5 years. However, their methods required a fixed lifetime, hence the number of screening fixed as well.

Our model differs from existing work. We developed a systematic approach to evaluate the long-term outcomes in regular screening. Unlike other methods that deals with overdiagnosis alone, we separate all participants in a screening program into four disjoint outcomes to evaluate the whole cohort. Other methods are retrospective, our method is prospective: we use existing data to obtain information on three key parameters: screening sensitivity, sojourn time distribution, and transition probability, then use these parameters and our probability model to predict the probability of true-early-detection, no-early-detection, overdiagnosis and symptom-free-life for different screening frequencies and different age groups in the future. Long-term incidence data is not needed when using the probability modeling. In summary, no assessment tools exist for continued screening today; our model may provide a baseline.

This model can be generalized in practical ways. For example, consider an eighty-year-old woman who has a history of screening and has remained healthy so far. How best to incorporate that screening history into the calculation of the probabilities of overdiagnosis, and true-early-detection? How can we incorporate personal risks, such as a family history of cancer, into the model? We are working on this extension. The model also shows the importance of accurate estimation of the screening sensitivity, the sojourn time distribution, and the transition probability density, because probability of each of the long term outcomes is a function of these parameters.

Possible limitations of our model include that we do not consider the case when the sensitivity and the sojourn time are correlated, and we do not model the sojourn time in S_p as depending on the age at which the woman enters the preclinical state. Accurate estimates of the sensitivity, the sojourn time distribution, and the transition probability are very important, though it is a complicated issue itself. We are looking into relaxing these assumptions when we have access to more recent data. As pointed out by an anonymous referee, true-early-detection may not be translated into benefit: if there is no effective treatment, early-detection could be detrimental, we totally agree. However, we hope our model will help policy makers evaluate a screening program's long-term effects more appropriately.

Supplementary Materials

Detailed simulation results in Section 3 and programming codes in C/C++ are provided in the appendix in the *Statistica Sinica* website.

Acknowledgement

We thank the anonymous AE and referee for comments that improved the article. We would like to thank Dr. Naisyin Wang, the Editor, and Dr. Ylvisaker, the Consulting Editor, for their many helpful suggestions in revising the manuscipt.

References

Badgwell, B. D., Giordano, S. H., Duan, Z. Z., Fang, S., Bedrosian, I., Kuerer, H. M., Singletary, S. E., Hunt, K. K., Nortobagyi, G. N. and Babiera, G. (2008). Mammography before diagnosis among women age 80 years and older with breast cancer. *J. Clinical Oncology* **26**, 2482-2488.

Davidov, O. and Zelen, M. (2004). Overdiagnosis in early detection programs. *Biostatistics* **5**, 603-613.

Duffy, S. W., Lynge, E., Jonsson, H., Ayyaz, S. and Olsen, A. H. (2008). Complexities in the estimation of overdiagnosis in breast cancer screening. *British J. Cancer* **99**, 1176-1178.

Gotzsche, P. C., Jorgensen, K. J., Mahlen, J, and Zahl, P. H. (2009). Estimation of lead time and overdiagnosis in breast cancer screening. *British J. Cancer* **100**, 219.

Jorgensen, K. J. and Gotzsche, P. C. (2009). Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *British Medical J.*. Doi: 10.1136/bmj.b2587.

National Institute of Health (2010). `http://seer.cancer.gov/csr/1975_2007/results_merged/topic_lifetime_risk.pdf`

Shapiro, S., Venet, W., Strax, P. and Venet, L.(1988). *Periodic Screening for Breast Cancer. The Health Insurance Plan Project and its Sequelae*, 1963-1986. The Johns Hopkins University Press, Baltimore.

Social Security Administration (2010). Period Life Table. Actuarial Publications. `http://www.ssa.gov/OACT/STATS/table4c6.html`.

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. *Internat. J. Biostatist.*. DOI: 10.1515/1557-4679.1363.

Wu, D. and Perez, A. (2011). A limited review of over-diagnosis methods and long term effects in breast cancer screening. *Oncology Rev.* **5**, 143-147.

Wu, D., Rosner, G. L. and Broemeling, L. D. (2005). MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. *Biometrics* **61**, 1056-1063.

Wu, D., Rosner, G. L. and Broemeling, L. D. (2007). Bayesian inference for the lead time in periodic cancer screening. *Biometrics* **63**, 873-880.

Zackrisson, S., Andersson, I., Janzon, L., Manjer, J. and Garne, J. P. (2006). Rate of over-diagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study. *BMJ* **332**, 689-692.

Zahl P. H., Mahlen, J. and Welch, H. G. (2008). The natural history of invasive breast cancers detected by screening mammography. *Arch internal Medicine* **168**, 2311-2316.

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA.

E-mail: dongfeng.wu@louisville.edu

Department of Statistics, Indiana University, Bloomington, IN 47408, USA.

E-mail: kkafadar@indiana.edu

Division of Oncology Biostatistics and Bioinformatics, The Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD 21205, USA.

E-mail: grosner1@jhmi.edu

(Received February 2012; accepted March 2013)