Dynamic Scheduling of the Upcoming Exam in Cancer Screening

Dongfeng Wu\textsuperscript{1} and Karen Kafadar\textsuperscript{2}

\textsuperscript{1}Department of Bioinformatics and Biostatistics  
University of Louisville

\textsuperscript{2}Department of Statistics  
University of Virginia

\textit{ICHPS, January 12, 2018}
Efficient design of cancer screening is not available until now: at what age to start screening? and how frequently?

For an individual who has gone through a few screening exams in the past and get negative results so far, when should s/he come back for the next exam?

Some research has been done to study the problem of a fixed budget that allows only \( n \) exams in a fixed age interval, using some utility functions (Zelen 1993, Lee and Zelen 1998). The problem: utility functions are subjective, and the methods cannot be applied directly by diagnostic radiologists.

We will use conditional probability of incidence before the next exam, to control the risk of incidence.

The method can be applied to any kinds of screening; it will be applied to the women’s breast cancer using the Health Insurance Plan for Greater New York (HIP) data.
The HIP data

- The Health Insurance Plan of Greater New York (HIP) study, which began at the end of 1963, was the first randomized clinical trial to examine mammogram screenings for breast cancer.
- Asymptomatic women without a history of breast cancer, with initial age: 40−64, average age = 51.2, and with 15 years of follow-up.
- About 60,000 participants were equally randomized into 2 arms: Study and Control.
- Study group: mammogram + clinical exam in each screening, and 4 annual screenings.
- Control group: usual care without screening.
The HIP Study Group Data

\[\begin{align*}
& s_1 \quad r_1 \quad s_2 \quad r_2 \quad s_3 \quad r_3 \quad s_4 \quad r_4 \\
& t_0 \quad t_1 \quad t_2 \quad t_3 \quad T \\
& n_i \quad 20166 \quad 15936 \quad 13679 \quad 11971
\end{align*}\]

- \( t_0 < t_1 < \cdots < t_{k-1} < t_k \): \( k \) ordered screening exam times.
- \( n_i \): the number of individuals examined at \( t_{i-1} \)
- \( s_i \): screening detected cases at the exam given at \( t_{i-1} \)
- \( r_i \): interval cases, the number of cases found in the clinical state \( (S_c) \) within \( (t_{i-1}, t_i) \).
- \((n_i, s_i, r_i)\): data stratified by initial age in the \( i \)-th interval.
### Table 1. HIP study group data: 4 annual screenings

<table>
<thead>
<tr>
<th>Age</th>
<th>$n_1$</th>
<th>$s_1$</th>
<th>$r_1$</th>
<th>$n_2$</th>
<th>$s_2$</th>
<th>$r_2$</th>
<th>$n_3$</th>
<th>$s_3$</th>
<th>$r_3$</th>
<th>$n_4$</th>
<th>$s_4$</th>
<th>$r_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>985</td>
<td>1</td>
<td>0</td>
<td>850</td>
<td>1</td>
<td>0</td>
<td>782</td>
<td>2</td>
<td>2</td>
<td>687</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>1014</td>
<td>3</td>
<td>0</td>
<td>887</td>
<td>2</td>
<td>1</td>
<td>833</td>
<td>1</td>
<td>0</td>
<td>744</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>47</td>
<td>933</td>
<td>2</td>
<td>1</td>
<td>808</td>
<td>2</td>
<td>0</td>
<td>747</td>
<td>0</td>
<td>2</td>
<td>658</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>978</td>
<td>1</td>
<td>0</td>
<td>849</td>
<td>0</td>
<td>0</td>
<td>777</td>
<td>0</td>
<td>0</td>
<td>683</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>915</td>
<td>1</td>
<td>0</td>
<td>817</td>
<td>0</td>
<td>0</td>
<td>765</td>
<td>1</td>
<td>0</td>
<td>677</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: a characteristic of periodic cancer screening is that data were collected repeatedly.
The progressive disease model assumes that all clinical cancer will go through three states (Zelen and Feinleib in 1969):

- $S_0$ is the **disease-free** state or the state in which the disease can not be detected.
- $S_p$ is the **preclinical** state, in which an asymptomatic individual unknowingly has the disease that a screening exam can detect.
- $S_c$ is the **clinical** state at which the disease manifests itself in clinical symptoms.
Sojourn time, transition probability and lead time

- Illustration of disease progression and the lead time:

Let $t_i$ represent a person’s age.
- **sojourn time**: $(t_2 - t_1)$, the time duration in the preclinical state.
- **transition probability density**: measures the time duration in the disease free state, i.e. the distribution of $t_1$.
- **lead time**: $(t_2 - t)$, the time interval between the diagnosis time $t$ and the onset of $S_c$ if not screened, i.e. the length of time the diagnosis is advanced by screening.
The three key parameters

- sensitivity at age $t$: $\beta(t) = P(X = 1|D = 1, t)$.

\[ T_1 \sim w(t) \quad T_2 \sim q(x) \quad S_c \]

- $w(t)$: Probability Density Function (PDF) of the time spent in the disease-free state $S_0$.
- $q(x)$: PDF of the sojourn time (time duration in the preclinical state $S_p$).
- $Q(z) = Pr(T > z) = \int_z^\infty q(x)dx$, survivor function of the sojourn time.
- The three key parameters: sensitivity $\beta(t)$, transition density $w(t)$, and sojourn time distribution $q(x)$. Any other term/probability is a function of the three key parameters.
Define events

History

\[ 0 \quad t_0 \quad t_1 \quad \cdots \quad t_{K-1} \quad (t_{K-1} + t_x) \]

Current age

Suppose one has taken \( K = \) exams at ages \( t_0 < t_1 < \cdots < t_{K-1} \) and is asymptomatic at current age \( t_{K-1} \). Define events:

\[
\begin{align*}
H_K &= \{ \text{one is asymptomatic in } [0, t_{K-1}] \text{ after taking } K \text{ exams at ages } t_0 < t_1 < \cdots < t_{K-1} \}, \\
I_K &= \{ \text{one will be a clinical incident case first time in } (t_{K-1}, t_{K-1} + t_x) \}, \\
D_K &= \{ \text{one will be diagnosed at } (t_{K-1} + t_x) \text{ for the first time} \}, \\
A_K &= \{ \text{one will be asymptomatic in } (t_{K-1}, t_{K-1} + t_x) \},
\end{align*}
\]

The mutually exclusive events \((I_K, D_K, A_K)\) is a partition of the whole sample space:

\[ I_K \cup D_K \cup A_K = \Omega. \]
Let $t_K = t_{K-1} + t_x$, the probability of incidence before the next screening exam among people at risk ($I_K$ or $D_K$) is:

$$P(I_K|I_K \cup D_K, H_K) = \frac{P(I_K|H_K)}{P(I_K \cup D_K|H_K)} = \frac{P(I_K|H_K)}{1 - P(A_K|H_K)}$$

$$= \frac{P(I_K \cap H_K)/P(H_K)}{1 - P(A_K \cap H_K)/P(H_K)} = \frac{P(I_K \cap H_K)}{P(H_K) - P(A_K \cap H_K)}$$

We need to calculate the probabilities: $P(H_K), P(I_K \cap H_K)$ and $P(A_K \cap H_K)$. 


Probability calculation

Let \( t_{-1} = 0 \), then

\[
P(H_K) = P(\text{remained in } S_0 \text{ in } (0, t_{K-1})) + P(\text{entered and remained in } S_p \text{ and not being detected})
\]

\[
= 1 - \int_0^{t_{K-1}} w(x)dx + \sum_{i=0}^{K-1} (1 - \beta_i) \cdots (1 - \beta_{K-1}) \int_{t_i}^{t_{i-1}} w(x)Q(t_{K-1} - x)dx
\]
Probability calculation - cont.

\[ P(I_K \cap H_K) = \sum_{i=0}^{K-1} (1 - \beta_i) \cdots (1 - \beta_{K-1}) \int_{t_i}^{t} w(x)[Q(t_{K-1} - x) - Q(t_K - x)]dx \]

\[ + \int_{t_{K-1}}^{t_K} w(x)[1 - Q(t_K - x)]dx, \quad (3) \]

And

\[ P(A_K \cap H_K) = 1 - \int_0^{t_K} w(x)dx \]

\[ + \sum_{j=0}^{K} (1 - \beta_j) \cdots (1 - \beta_K) \int_{t_{j-1}}^{t_j} w(x)Q(t_K - x)dx, \quad (4) \]

And in fact \( P(A_K \cap H_K) = P(H_{K+1}) \).
This probability of incidence, \( P(I_K|I_K \cup D_K, H_K) \), is monotonically increasing as the upcoming screening time interval \( t_x \) increases. Therefore, for any pre-selected small value \( \alpha \), there exists a unique numerical solution \( t^* \), that satisfies

\[
P(I_K|I_K \cup D_K, H_K) = \alpha.
\]

That is, with probability \( (1 - \alpha) \), she will NOT be a clinical cancer case before her next screening exam at her age \( (t_{K-1} + t^*) \), where \( t_{K-1} \) is her current age.

One may choose \( \alpha \) at a risk level that she is comfortable with, such as 0.05 or 0.10 (5% or 10%).
Lead time and overdiagnosis

After $t^*$ is found (based on one’s screening history and other parameters), we can make more inferences if one were diagnosed with cancer at $(t_{K-1} + t^*)$:

- Calculate the conditional distribution of the lead time at $(t_{K-1} + t^*)$.
- Calculate the probability of overdiagnosis (and true-early-detection) at $(t_{K-1} + t^*)$.

These provide predictive information.
We let $L$ be the lead time (the time diagnosis is advanced by screening), and let $t_K = t_{K-1} + t^*$, then the conditional probability density function (PDF) of the lead time given the event $D_K$ (i.e., one will be diagnosed at $t_K$ for the first time) is

$$f_L(z|D_K) = \frac{f_L(z, D_K)}{P(D_K)}, \quad \text{for } z \in (0, \infty).$$ \hspace{1cm} (6)

Where the denominator

$$P(D_K) = \beta_K \left\{ \sum_{i=0}^{K-1} (1 - \beta_i) \cdots (1 - \beta_{K-1}) \int_{t_{i-1}}^{t_i} w(x) Q(t_K - x) \, dx \right.$$

$$\left. + \int_{t_{K-1}}^{t_K} w(x) Q(t_K - x) \, dx \right\}. \hspace{1cm} (7)$$
And the numerator

\[ f_L(z, D_K) = \sum_{i=0}^{K} f_L(z, D_K, \text{and she is from the } i\text{-th generation}) \]

\[ = \beta_K \left\{ \sum_{i=0}^{K-1} (1 - \beta_i) \cdots (1 - \beta_{K-1}) \int_{t_{i-1}}^{t_i} w(x)q(t_K + z - x)dx \right. \]

\[ + \int_{t_{K-1}}^{t_K} w(x)q(t_K + z - x)dx \} \].

This is a valid probability density function (pdf), since

\[ \int_{0}^{\infty} f_L(z|D_K)dz = 1. \]
Probability of overdiagnosis and true-early-detection

To calculate the probability, we first let the lifetime $T$ to be a fixed value, then let it to be random. Given one would be diagnosed at $t_K = t_{k-1} + t^*$ and one's fixed lifetime $T = t (> t_K)$, the probability of overdiagnosis and true-early-detection can be derived:

\[
P(\text{OverD}|D_K, T = t) = \frac{P(\text{OverD}, D_K|T = t)}{P(D_K|T = t)} ,
\]

\[
P(\text{TrueED}|D_K, T = t) = \frac{P(\text{TrueED}, D_K|T = t)}{P(D_K|T = t)} .
\]

where $P(D_K|T = t) = P(D_K)$ as in equation (7).
Setting 

\[
P(\text{OverD}, D_K \mid T = t) 
= \beta_K \left\{ \sum_{i=0}^{K-1} (1 - \beta_i) \cdots (1 - \beta_{K-1}) \int_{t_{i-1}}^{t_i} w(x) Q(t - x) dx 
\right. 
\left. + \int_{t_{K-1}}^{t_K} w(x) Q(t - x) dx \right\}. 
\] (10)

\[
P(\text{TrueED}, D_K \mid T = t) 
= \beta_K \left\{ \sum_{i=0}^{K-1} (1 - \beta_i) \cdots (1 - \beta_{K-1}) \int_{t_{i-1}}^{t_i} w(x)[Q(t_K - x) - Q(t - x)] dx 
\right. 
\left. + \int_{t_{K-1}}^{t_K} w(x)[Q(t_K - x) - Q(t - x)] dx \right\}. 
\] (11)

And it can be verified that 

\[
P(\text{TrueED} \mid D_K, T = t) + P(\text{OverD} \mid D_K, T = t) = 1.
\]
Probability of Overdiagnosis - cont.

Now we allow human lifetime $T$ to be random, and let $f_T(t|T > t_k)$ be the conditional PDF of the lifetime $T$, derived from the actuarial life table (US Social Security Administration, http://www.ssa.gov/OACT/STATS/table4c6.html), then

$$P(\text{OverD}|D_K, T > t_K) = \int_{t_K}^{\infty} P(\text{OverD}|D_K, T = t)f_T(t|T > t_K)dt,$$

$$P(\text{TrueED}|D_K, T > t_K) = \int_{t_K}^{\infty} P(\text{TrueED}|D_K, T = t)f_T(t|T > t_K)dt.$$

Where

$$f_T(t|T \geq t_K) = \begin{cases} \frac{f_T(t)}{P(T > t_K)} = \frac{f_T(t)}{1 - F_T(t_k)}, & \text{if } t \geq t_K, \\ 0, & \text{otherwise.} \end{cases}$$
A period life table is based on the mortality experience of a population during a relatively short period of time. Here we present the 2007 period life table for the Social Security area population. For this table, the period life expectancy at a given age represents the average number of years of life remaining if a group of persons at that age were to experience the mortality rates for 2007 over the course of their remaining life.

<table>
<thead>
<tr>
<th>Exact age</th>
<th>Male</th>
<th></th>
<th></th>
<th>Female</th>
<th></th>
<th></th>
<th>Life expectancy</th>
<th>Life expectancy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Death probability</td>
<td>Number of lives</td>
<td>Life expectancy</td>
<td>Death probability</td>
<td>Number of lives</td>
<td>Life expectancy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.007379</td>
<td>100,000</td>
<td>75.38</td>
<td>0.006096</td>
<td>100,000</td>
<td>80.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.000494</td>
<td>99,262</td>
<td>74.94</td>
<td>0.000434</td>
<td>99,390</td>
<td>79.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.000317</td>
<td>99,213</td>
<td>73.98</td>
<td>0.000256</td>
<td>99,347</td>
<td>78.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.000241</td>
<td>99,182</td>
<td>73.00</td>
<td>0.000192</td>
<td>99,322</td>
<td>77.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.000200</td>
<td>99,158</td>
<td>72.02</td>
<td>0.000148</td>
<td>99,303</td>
<td>76.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.000179</td>
<td>99,138</td>
<td>71.03</td>
<td>0.000136</td>
<td>99,288</td>
<td>76.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.000166</td>
<td>99,120</td>
<td>70.04</td>
<td>0.000128</td>
<td>99,275</td>
<td>75.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.000152</td>
<td>99,104</td>
<td>69.05</td>
<td>0.000122</td>
<td>99,262</td>
<td>74.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.000133</td>
<td>99,089</td>
<td>68.06</td>
<td>0.000115</td>
<td>99,250</td>
<td>73.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.000108</td>
<td>99,075</td>
<td>67.07</td>
<td>0.000106</td>
<td>99,238</td>
<td>72.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: The conditional PDF of lifetime for females when $t_K = 60, 70, 80$
Simulation set up

- Historic screening interval $\Delta = 1, 1.5, 2$ years, from $t_0 = 50$ to current age $t_{K-1} = 62$.
- Screening sensitivity $\beta = 0.7, 0.8, 0.9$.
- Sojourn time: Log logistic distribution
  \[ Q(x) = \left[ 1 + (x\rho)^\kappa \right]^{-1}, \quad \kappa > 0, \rho > 0. \]
  Let mean sojourn time be 2, 5, 10 and 15 years, i.e., $\kappa = 2.5, \rho = 0.661, 0.264, 0.132, 0.088$.
- Transition density
  \[ w(t|\mu, \sigma^2) = \frac{0.2}{\sqrt{2\pi}\sigma t} \exp \left\{ -\frac{(\log t - \mu)^2}{2\sigma^2} \right\}, \sigma > 0. \]
  with mode around 60 years old, i.e. $\mu = 4.2, \sigma^2 = 0.1$.
- Probability (risk) of incidence before next screening: $\alpha = 0.05, 0.1, 0.15, 0.2$. 

Wu
Optimal Scheduling
Table 2: Simulated optimal screening time $t^*$

<table>
<thead>
<tr>
<th></th>
<th>Beta = 0.7</th>
<th>Beta = 0.8</th>
<th>Beta = 0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\Delta$</td>
<td>$\alpha$</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.13</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>0.1</td>
<td>0.33</td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td>0.15</td>
<td>0.58</td>
<td>0.45</td>
<td>0.39</td>
</tr>
<tr>
<td>0.2</td>
<td>0.83</td>
<td>0.69</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MST = 5 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Beta = 0.7</td>
<td>Beta = 0.8</td>
<td>Beta = 0.9</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\Delta$</td>
<td>$\alpha$</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.87</td>
<td>0.53</td>
<td>0.38</td>
</tr>
<tr>
<td>0.1</td>
<td>1.65</td>
<td>1.28</td>
<td>1.00</td>
</tr>
<tr>
<td>0.15</td>
<td>2.25</td>
<td>1.94</td>
<td>1.67</td>
</tr>
<tr>
<td>0.2</td>
<td>2.78</td>
<td>2.52</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MST = 10 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Beta = 0.7</td>
<td>Beta = 0.8</td>
<td>Beta = 0.9</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\Delta$</td>
<td>$\alpha$</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>0.05</td>
<td>2.69</td>
<td>2.22</td>
<td>1.74</td>
</tr>
<tr>
<td>0.1</td>
<td>4.03</td>
<td>3.66</td>
<td>3.29</td>
</tr>
<tr>
<td>0.15</td>
<td>5.11</td>
<td>4.79</td>
<td>4.47</td>
</tr>
<tr>
<td>0.2</td>
<td>6.09</td>
<td>5.81</td>
<td>5.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MST = 15 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Beta = 0.7</td>
<td>Beta = 0.8</td>
<td>Beta = 0.9</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\Delta$</td>
<td>$\alpha$</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>0.05</td>
<td>4.43</td>
<td>4.02</td>
<td>3.56</td>
</tr>
<tr>
<td>0.1</td>
<td>6.35</td>
<td>6.02</td>
<td>5.66</td>
</tr>
<tr>
<td>0.15</td>
<td>7.91</td>
<td>7.61</td>
<td>7.30</td>
</tr>
<tr>
<td>0.2</td>
<td>9.33</td>
<td>9.06</td>
<td>8.78</td>
</tr>
</tbody>
</table>

$\Delta$ is the historical screening interval (in years) in the corresponding row.
$\alpha$ = probability of incidence before the next screening exam in the 1st column.
MST = mean sojourn time.
Summary of the simulated optimal scheduling time

- Sensitivity will affect the next screening time interval in a positive way: higher sensitivity means longer time interval to maintain the same incidence risk $\alpha$.

- Mean Sojourn Time (MST) affects the next screening interval in a positive way: long MST (slow growing cancer or low risk people) means she can wait longer time to carry out the next screening.

- Screening history plays an important role: shorter screening interval in the past means longer screening interval for the upcoming test, and vise versa.

- Lower probability of incidence $\alpha$, means shorter screening interval.
Figure 2: Lead time distribution using $t^*$ under the four factors: $\alpha, \beta, \Delta$ and MST.
The density curve of the lead time changes MORE with the risk $\alpha$ and the mean sojourn time (MST), and it changes LESS with the screening sensitivity $\beta$ and the past screening interval $\Delta$ if the optimal scheduling time $t^*$ is adopted.

- Smaller $\alpha$ means larger mean/median/mode of the lead time, and smaller standard deviation of the lead time.
- Longer MST means larger mean/median/mode, and larger standard deviation of the lead time.
Table 3: Estimated probability of overdiagnosis (in percentage) using the $t^*$

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$\beta = 0.7$</th>
<th>$\beta = 0.8$</th>
<th>$\beta = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$MST = 2$ yrs</td>
<td>$MST = 5$ yrs</td>
<td>$MST = 10$ yrs</td>
</tr>
<tr>
<td>0.05</td>
<td>1.58</td>
<td>1.58</td>
<td>1.59</td>
</tr>
<tr>
<td>0.10</td>
<td>1.58</td>
<td>1.60</td>
<td>1.60</td>
</tr>
<tr>
<td>0.15</td>
<td>1.60</td>
<td>1.62</td>
<td>1.62</td>
</tr>
<tr>
<td>0.20</td>
<td>1.62</td>
<td>1.64</td>
<td>1.64</td>
</tr>
</tbody>
</table>

$^a$ MST = 2 yrs

$^b$ MST = 5 yrs

$^c$ MST = 10 yrs

$^d$ MST = 15 yrs
Figure 3: Percentage of overdiagnosis vs. next screening time $t$

Percentage of overdiagnosis under different mean sojourn time

- $\beta = 0.7$ Delta = 1yr
- $\beta = 0.7$ Delta = 2yr
- $\beta = 0.9$ Delta = 1yr
- $\beta = 0.9$ Delta = 2yr
Summary of the over-diagnosis under the $t^*$

- Mean sojourn time (MST) plays the most important role in overdiagnosis: longer MST means larger probability of overdiagnosis.

- Probability of overdiagnosis will slightly increase as the screening sensitivity $\beta$ increases.

- For the same fixed MST, the probability of overdiagnosis won’t change much as the risk $\alpha$ increases.

- We plot the probability of overdiagnosis versus future scheduling time in Figure 3. It shows that the probability of overdiagnosis is monotonic increasing as $t^*$ increases, and it won’t change much with sensitivity $\beta$ and past screening interval $\Delta$ for a fixed MST.
Application to the HIP Data

- All methods that we derived before, are functions of the three key parameters: $\beta(t)$, $q(x)$, $w(t)$, so we need to extract these information from the HIP data first.

- Wu, Rosner & Broemeling (2005) developed statistical inference procedures to estimate the sojourn time $q(x)$, the age-dependent sensitivity $\beta(t)$, and the age-dependent transition density $w(t)$ using parametric link and likelihood function.

- Sensitivity: $\beta(t) = \left[1 + \exp(-b_0 - b_1 * (t - m))\right]^{-1}$.

- Transition density $S_0 \rightarrow S_p$: 0.2 * lognormal pdf
  \[
  w(t|\mu, \sigma^2) = \frac{0.2}{\sqrt{2\pi}\sigma t} \exp\left\{-\frac{(\log t - \mu)^2}{2\sigma^2}\right\}, \sigma > 0.
  \]

- Sojourn time: Log logistic distribution
  \[
  q(t) = \frac{\kappa t^{\kappa - 1} \rho^\kappa}{[1 + (t\rho)^\kappa]^2}, \quad \kappa > 0, \rho > 0.
  \]
Let \( HIP = \) HIP data, and \( \theta = (b_0, b_1, \mu, \sigma^2, \kappa, \rho) \). Using the likelihood function and Markov Chain Monte Carlo (MCMC), 2000 Bayesian posterior samples \( (\theta_j^*) \) were generated; for details, see Wu et al. (2005). Using each \( \theta_j^* \), and \( P(I_K|I_K \cup D_K, H_K, \theta_j^*) = \alpha \), we conducted Bayesian inference on hypothetical cohorts of asymptomatic women with current age \( t_{K-1} = 62 \) (or 72), assuming that they have started their first screening at age \( t_0 = 50 \) (or 60), and with different screening intervals \( \Delta = 1, 2 \) and 3 years in the 12 years.
Table 4: Optimal scheduling time using the HIP data

<table>
<thead>
<tr>
<th>α</th>
<th>$\Delta = 1.0$</th>
<th>$\Delta = 2.0$</th>
<th>$\Delta = 3.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t_0 = 50$, $t_{K-1} = 62$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.30 (0.19)</td>
<td>0.23 (0.15)</td>
<td>0.22 (0.15)</td>
</tr>
<tr>
<td>0.10</td>
<td>0.56 (0.31)</td>
<td>0.45 (0.19)</td>
<td>0.43 (0.19)</td>
</tr>
<tr>
<td>0.15</td>
<td>0.78 (0.42)</td>
<td>0.67 (0.24)</td>
<td>0.64 (0.21)</td>
</tr>
<tr>
<td>0.20</td>
<td>0.99 (0.52)</td>
<td>0.88 (0.32)</td>
<td>0.85 (0.25)</td>
</tr>
<tr>
<td></td>
<td>$t_0 = 60$, $t_{K-1} = 72$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.42 (0.40)</td>
<td>0.32 (0.25)</td>
<td>0.30 (0.20)</td>
</tr>
<tr>
<td>0.10</td>
<td>0.68 (0.55)</td>
<td>0.59 (0.40)</td>
<td>0.55 (0.31)</td>
</tr>
<tr>
<td>0.15</td>
<td>0.91 (0.67)</td>
<td>0.82 (0.53)</td>
<td>0.78 (0.43)</td>
</tr>
<tr>
<td>0.20</td>
<td>1.11 (0.79)</td>
<td>1.04 (0.65)</td>
<td>1.00 (0.55)</td>
</tr>
</tbody>
</table>
Table 5: Estimated mean/median/mode/standard deviation of the lead time at $t^*$ using HIP data

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\Delta = 1.0$</th>
<th>$\Delta = 2.0$</th>
<th>$\Delta = 3.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t_0 = 50, t_{K-1} = 62$</td>
<td>$t_0 = 50, t_{K-1} = 62$</td>
<td>$t_0 = 50, t_{K-1} = 62$</td>
</tr>
<tr>
<td>0.05</td>
<td>1.84,0.97,0.49(3.42) 1.88,0.95,0.47(3.61) 1.92,0.95,0.46(3.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>1.80,0.92,0.42(3.42) 1.85,0.93,0.41(3.58) 1.89,0.93,0.41(3.72)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>1.78,0.89,0.35(3.43) 1.83,0.90,0.35(3.58) 1.87,0.91,0.35(3.70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>1.77,0.87,0.30(3.46) 1.82,0.88,0.30(3.59) 1.85,0.89,0.30(3.70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$t_0 = 60, t_{K-1} = 72$</td>
<td>$t_0 = 60, t_{K-1} = 72$</td>
<td>$t_0 = 60, t_{K-1} = 72$</td>
</tr>
<tr>
<td>0.05</td>
<td>1.83,0.96,0.48(3.38) 1.87,0.96,0.47(3.53) 1.90,0.96,0.46(3.67)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>1.78,0.91,0.40(3.39) 1.83,0.92,0.40(3.52) 1.86,0.93,0.40(3.63)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>1.77,0.88,0.34(3.41) 1.81,0.89,0.34(3.52) 1.84,0.90,0.34(3.62)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>1.76,0.86,0.28(3.44) 1.79,0.87,0.29(3.54) 1.82,0.88,0.29(3.63)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6: Estimated mean probability of overdiagnosis with 95% CI (in percentage) at $t^*$ using HIP data

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\Delta = 1.0$</th>
<th>$\Delta = 2.0$</th>
<th>$\Delta = 3.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t_0 = 50, t_{K-1} = 62$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>2.25(0.57,10.97)</td>
<td>2.38(0.56,12.47)</td>
<td>2.52(0.56,13.80)</td>
</tr>
<tr>
<td>0.10</td>
<td>2.26(0.53,11.35)</td>
<td>2.37(0.54,12.61)</td>
<td>2.49(0.54,13.65)</td>
</tr>
<tr>
<td>0.15</td>
<td>2.29(0.52,11.37)</td>
<td>2.39(0.52,12.84)</td>
<td>2.50(0.52,13.67)</td>
</tr>
<tr>
<td>0.20</td>
<td>2.34(0.51,11.54)</td>
<td>2.43(0.51,12.91)</td>
<td>2.52(0.51,13.80)</td>
</tr>
<tr>
<td></td>
<td>$t_0 = 60, t_{K-1} = 72$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>4.65(1.40,19.98)</td>
<td>4.76(1.40,20.48)</td>
<td>4.89(1.40,21.45)</td>
</tr>
<tr>
<td>0.10</td>
<td>4.68(1.33,20.37)</td>
<td>4.78(1.33,21.39)</td>
<td>4.89(1.33,21.42)</td>
</tr>
<tr>
<td>0.15</td>
<td>4.76(1.30,20.76)</td>
<td>4.84(1.31,21.86)</td>
<td>4.94(1.31,21.73)</td>
</tr>
<tr>
<td>0.20</td>
<td>4.87(1.30,21.49)</td>
<td>4.93(1.30,22.07)</td>
<td>5.02(1.31,22.57)</td>
</tr>
</tbody>
</table>
The next screening time is slightly decreasing if the past screening interval increases from 1 year to 3 years; it follows the same pattern as in our simulation study.

The mean/median/mode of the lead time is decreasing as the risk $\alpha$ increases.

The mean/median/mode of the lead time is increasing as the past screening interval $\Delta$ increases if other factors are the same.

The probability of overdiagnosis is low in the HIP study (2-5%), although it slightly increasing with age, past screening interval $\Delta$. 
Conclusion

- We developed a probability method to dynamically schedule one's upcoming screening exam, based on one's past screen history, risk tolerance $\alpha$, screening sensitivity, sojourn time distribution, etc.

- The method can handle any screening history $t_0 < t_1 < \ldots t_{K-1}$, including unequally-spaced screening intervals.

- Simulations show that longer screening interval in the past means shorter interval for the upcoming test.

- We provide predictive information on the lead time and overdiagnosis if one were diagnosed with cancer at the future time. This may be the first step towards personalized screening schedule in the near future.

- The modeling approach is just one way of thinking about the problem. Other models and approaches are possible. The important point is to recognize that screening has outcomes & consequences that one should consider, especially for policy purposes.
Thank You!!

Any questions or comments??
The actuarial life table of SSA

- The distribution of the lifetime $f_T(t)$ was derived from the actuarial life table, Social Security Administration.
  http://www.ssa.gov/OACT/STATS/table4c6.html
- The period life table is based on population mortality from all Social Security area, including 50 states, DC, and surrounding islands of the US.
- It provides the conditional probability of death within one year given one’s current age $P(T < N + 1|T \geq N)$, from age $N = 0$ to 119.
- Let $\alpha_N = P(T \geq N + 1|T \geq N) = 1 - P(T < N + 1|T \geq N)$. 
Transfer the life table to a valid pdf

- For any integer age $t_0$,

$$P(T \geq t_0 + N | T \geq t_0) = \prod_{i=1}^{N} a_{t_0+i-1}, \quad \forall N = 1, 2, \ldots, 120 - t_0.$$

- Using a density approximation, we have

$$f_T(t_0 + N | T \geq t_0) = \lim_{\epsilon \to 0} \frac{P(t_0 + N < T \leq t_0 + N + \epsilon | T \geq t_0)}{\epsilon}$$

$$\approx P(t_0 + N < T \leq t_0 + N + 1 | T \geq t_0) = (1 - a_{t_0+N}) \prod_{i=1}^{N} a_{t_0+i-1}.$$

- Finally, for any $t \in (N, N + 1)$ ($N < 120$), we use a step function to approximate: $f_T(t | T \geq t_0) \approx f_T(N | T \geq t_0)$.

- It is a valid pdf because $\sum_{N=0}^{120-t_0} f_T(t_0 + N | T \geq t_0) = 1$. 
Likelihood function and probability of detection at $k$th exam

\[ L = \prod_{t_0=40}^{64} \prod_{k=1}^{4} D_{k,t_0}^{s_{k,t_0}} l_{k,t_0}^{r_{k,t_0}} (1 - D_{k,t_0} - l_{k,t_0})^{n_{k,t_0}-s_{k,t_0}-r_{k,t_0}} \]

\[ D_{1,t_0} = \beta(t_0) \int_0^{t_0} w(x) Q(t_0 - x) dx. \]

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

\[ + \int_{t_{k-2}}^{t_{k-1}} w(x) Q(t_{k-1} - x) dx \} , \quad \text{for } \forall k = 2, \cdots, K. \]
Probability of incidence within the $k$th interval

$$l_{k,t_0} = \sum_{i=0}^{k-1} [1 - \beta(t_i)] \cdots [1 - \beta(t_{k-1})]$$

$$\times \int_{t_{i-1}}^{t_i} w(x)[Q(t_{k-1} - x) - Q(t_k - x)]dx$$

$$+ \int_{t_{k-1}}^{t_k} w(x)[1 - Q(t_k - x)]dx, \quad \text{for } \forall k = 1, \cdots, K.$$