A Semi Markov Model for Mammographic Detection of Breast Cancer

by

Keith James Chan

B.Sc, Carleton University 1996

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

February 1999

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Feb. 26/99

Abstract

Mammography is used as a screening tool to detect breast cancer at an early stage. The process of breast cancer detection using mammography is modelled with a semi-Markov process composed of three states: cancer-free (0), preclinical cancer (1) and clinical cancer (2). It is assumed that screening provides the ability to detect disease while it is still in the preclinical state before it enters the clinical state, however screening measurement is subject to error. The sojourn time in the preclinical detectable phase is thus of particular interest and it plays an important role in the design and assessment of screening programmes.

In previous work, the transition rate into the preclinical detectable phase has been modelled by an age-specific step function based on age at first screen. This does not lead to an increasing incidence of breast cancer with age but observations in several populations indicate that incidence increases at approximately the third power of age. This relationship is induced in the model by introducing a smooth age dependent transition rate into the preclinical detectable phase.

The model is applied to data provided by The Screening Mammography Programme of British Columbia (SMPBC). A Quasi-Newton algorithm is used to minimize the negative log likelihood function to obtain maximum likelihood estimates of the model parameters. Comparisons will be made with other published results and the effect of various assumptions examined.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Andy Coldman for his helpful comments and guidance throughout the writing of this thesis and I would also like to thank Harry Joe for his computer expertise.

This thesis would not have been possible without the data provided by the Screening Mammography Programme of British Columbia and the funding provided by the British Columbia Cancer Agency.

KEITH JAMES CHAN

The University of British Columbia
February 1999

Chapter 1

Introduction

1.1 Background

Mammography has long been advocated as a screening test for breast cancer which is the most commonly diagnosed cancer in women. Breast cancer is easiest to treat in the early stages of the disease and the most effective method of early detection is screening mammography. Other methods of detection such as breast self-examination and physical examination by a physician rely on touch and if a cancer is large enough to be found in this manner, it may be too late to effectively treat.

A screening mammogram is a low dose x-ray of the breast. Screening mammography is performed periodically and consists of x-raying each breast from the side and from the top. The breast is compressed between two flat plates to spread the tissue so that any abnormalities can be more easily identified. As some of these abnormalities are difficult to characterize, a patient's prior (if any) mammograms can be used as a reference to help identify subtle changes over time. If the result of a patient's screening mammogram is abnormal (symptoms of breast problem or presence of breast lump) further investigation of the abnormality is conducted such as diagnostic mammogram, ultrasound or biopsy.

Figure 1.1 is a schematic depicting the general progression of breast cancer. A patient is assumed to be born free of breast cancer. At some point in time a first cancer cell is created and this cell then divides until it is large enough to trigger a diagnosis of cancer. If a cancer is large enough to cause symptoms or be felt by physical touch then it is in the clinical phase. A cancer which is present but is not large enough to be clinical is in the preclinical phase which can be further broken down into preclinical detectable and preclinical undetectable. From this point on, we will refer to any cancer smaller than a preclinical detectable cancer as no cancer. The state in which a cancer is detectable by mammography prior to symptom development is defined as the preclinical detectable phase, and the length of this period is called the sojourn time. Note that due to the intermittent occurrence of screening, the transition of a cancer into the preclinical detectable phase is not directly observable.

Several randomized clinical trials have shown screening mammography to be effective, alone or in combination with breast self examination or professional physical examination, in reducing the mortality rate from breast cancer. However, there is still controversy about the indications for and use of screening mammography.

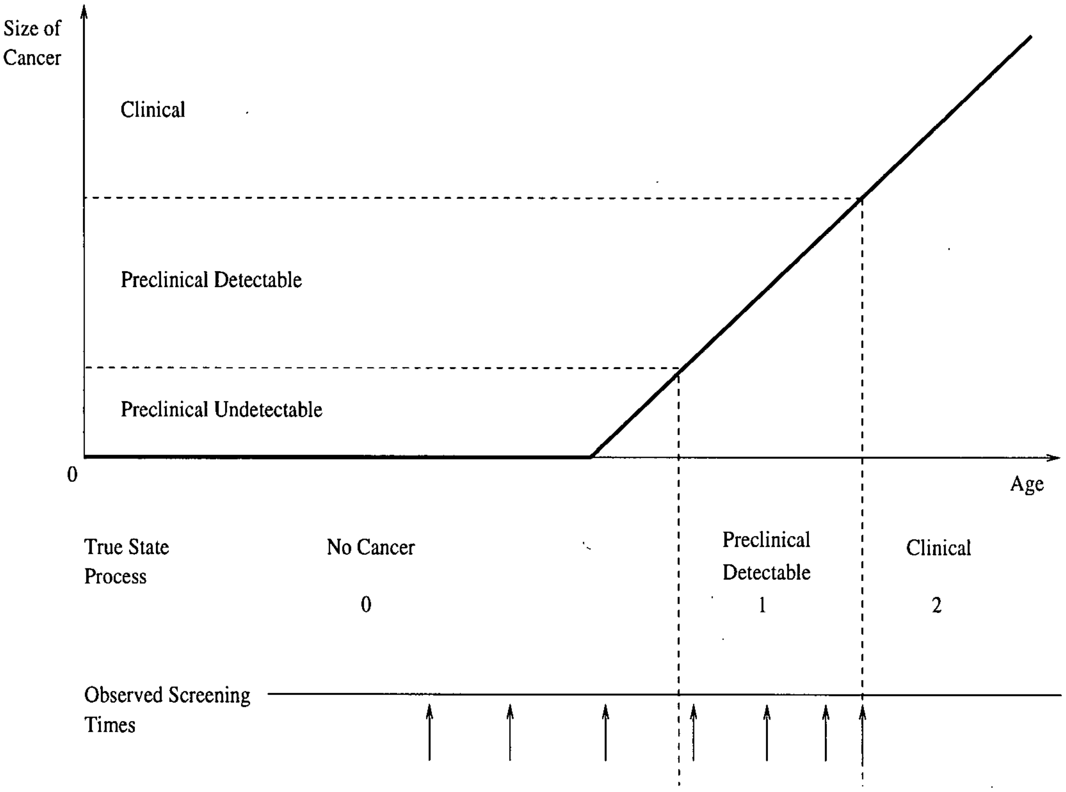

Figure 1.1: A Graph of the Progression of Cancer.

The female breast undergoes changes with age, particularly around the time of menopause and these changes should theoretically affect the sensitivity of mammography to detect breast cancer. Meta-analyses of clinical trial subsets, defined by age at first screen (less than or greater than 50), indicate that the proportional reduction in mortality from breast cancer of women undergoing mammography increases with age. This, coupled with the lower incidence of breast cancer in women less than 50 years of age has resulted in controversy about recommending mammographic screening in this age group. The whole question of screening is further complicated by the fact that screening requires repeated testing so that the effect of frequency (how often a woman goes for screening) as well as age are important. The frequency of mammography is highly dependent upon the sojourn time distribution and we will now examine previous models which have been used to estimate this distribution.

1.2 Previous Work

The optimal time between screens has been a long argued issue particularly because it is directly related to the effectiveness of a program. If screens occur too frequently then mammography will not be cost efficient and there will also be unnecessary exposure to radiation. On the other hand, if the time between screens is too long, preclinical cancers may be missed, which defeats the purpose of screening. To account for exposure, costs and benefits, the time between screens should be close to the mean sojourn time (MST). Hence, we wish to estimate the sojourn time distribution. Before developing the model, we will quickly define some notation.

1.2.1 Definition of Continuous Time Markov Chains

A stochastic process is a family of random variables $\{X(t);\ t \in T\}$ where T is called the index set of the process. The index t will be interpreted as time and we refer to $X(t)$ as the state of the process at time t. We call $\{X(t)\}$ a Markov process if, given the value of $X(t)$, the values of $X(t+s)$ are not influenced by the values of $X(t-u)$, $s > 0$, $u > 0$, i.e., the conditional probability of any future state given the present state and the past states depends only on the current state and is independent of the past states. For our purposes, T will be the non-negative real numbers and the state space of $X(t)$ will be discrete. This process is called a continuous time Markov chain and is written as $\{X(t);\ t \geq 0\}$.

1.2.2 Markov Modelling

Several published articles have used Markov modelling (under various assumptions) on breast cancer screening data. The general model is a continuous time Markov chain with state space $\{0,\ 1,\ 2\}$ (refer to Figure 1.1). At time $t = 0$, it starts in state 0 (no cancer). As time progresses it will move through state 0 to state 1 (preclinical detectable cancer) and is eventually absorbed in state 2 (clinical cancer). The process is specified by the parameters $\lambda_0(t)$ and $\lambda_1(t)$ where $\lambda_0(t)$ is the rate of transition from 0 to 1 at time t and $\lambda_1(t)$ is the rate of transition from 1 to 2 at time t. Pictorially, we have

$$0 \xrightarrow{\lambda_0(t)} 1 \xrightarrow{\lambda_1(t)} 2.$$

In 1995, Duffy et al. [1] used the simplest model of this class which has transition rates from 0 to 1 and from 1 to 2 constant over time for all ages, i.e.,

$$\lambda_0(t) = \lambda_0$$

$$\lambda_1(t) = \lambda_1.$$

This is an unrealistic assumption for breast cancer as incidence of breast cancer is known to increase with age [2]. The estimation of transition rates was also made by Duffy et al. assuming 100% sensitivity and 100% specificity of mammography, i.e., the observed state at screen is the true state. They used an estimation algorithm in which once transition rate estimates were found, they are assumed to be fixed and then the sensitivity is estimated. This two-step method of estimation is not optimal.

The following year, the same authors [3] improved on their model by simultaneously estimating transition rates, sensitivity and specificity. Their approach was to analyze the data in accordance to age at first screen, i.e., patients are assigned to one of the age groups 40-49, 50-59, 60-69, 70-74 based on their age at first screen, and the model was fitted to each of these data sets separately. A result of this splitting of the data into subsets is a smaller number of preclinically and clinically diagnosed cancers which may lead to unstable estimates [3]. The model in which $\lambda_0(t) = \lambda_0$, $\lambda_1(t) = \lambda_1$ and the sensitivity and specificity are assumed constant within age groups will be referred to as the Duffy model.

1.2.3 Data in Previous Work

Two sets of data are analyzed in [1], [3] and [4]. The Swedish Two-county Trial, appearing in all three articles, is a randomized trial that began in 1977. It consists of 133065 participants between the ages of 40 and 74 randomized to:

- invitation for screening with mammography every 2-3 years for approximately 8 years (active study population), or

- no screening during this period until the end of the 8 years (passive study population).

77080 women were invited for screening while 55985 were not invited. The other data set is from the screening programme in Florence from 1975 to 1986 where the screening interval was approximately 2 years for the 40-49 year age group [4].

1.2.4 Results of Previous Works

The results of fitting the Duffy model for the active study population in the Swedish Two-county Trial [3] are displayed in Table 1.1. Note that sensitivity is the complement of β, the false negative rate, and specificity is the complement of γ, the false positive rate.

In the next chapter, we present a more general model which simultaneously estimates transition rates, sensitivity and specificity where the transition rate into the preclinical detectable phase is a function of time.

Parameter	Results for the following age groups:		
	40-49 years	50-59 years	60-69 years
λ_0 ($\times 10^2$)	0.089	0.155	0.236
	(0.084-0.095)	(0.148-0.163)	(0.227-0.244)
λ_1	0.407	0.267	0.236
	(0.350-0.472)	(0.245-0.291)	(0.223-0.250)
MST$(= 1/\lambda_1)$	2.459	3.745	4.234
	(2.120-2.854)	(3.441-4.077)	(3.997-4.485)
β, false negative rate	0.168	0.000	0.000
	(0.107-0.264)	(0-0.000)	(0-0.000)
γ, false positive rate	0.0002	0.0000	0.0000

Table 1.1: Duffy's results of estimation for parameters from the Swedish Two-county Trial, (Respective 95% Confidence Intervals in parentheses).

Chapter 2

Model and Techniques

2.1 Introduction

Our model development will be done in two parts. The first part is to define transition probabilities for the 'true' state $\{T(a)\}$ of cancer where a is age. The second part is to define probabilities for the 'observed' state $\{O(a)\}$ of cancer. We will also describe the data provided by the SMPBC and the construction of the associated likelihood function.

2.2 True State Process

The model for the true state process was described briefly in Section 1.2.2. We continue with the description of the process.

At age a, the true state process is in one of three states:

$$T(a) = \begin{cases} 0, & \text{if no cancer at age } a; \\ 1, & \text{if preclinical detectable cancer at } a; \\ 2, & \text{if clinical cancer at } a \text{ .} \end{cases}$$

It is impossible to regress to a less severe cancer state and 2 is an absorbing state. We have the following properties:

1. $P\{T(a+\Delta a) = k|T(a) = k\} = 1 - \lambda_k(a)\Delta a + o(\Delta a)$, $k = 0, 1$
2. $P\{T(a+\Delta a) = k+1|T(a) = k\} = \lambda_k(a)\Delta a + o(\Delta a)$, $k = 0, 1$
3. $P\{T(a+\Delta a) = j|T(a) = k\} = 0$, $j < k$.

For age a, we will assume that the rates of transition may be approximated by polynomials so that for state 0 to 1 we have $\lambda_0(a) = \lambda_0 a^n > 0$ and, for state 1 to 2, $\lambda_1(a) = \lambda_1 a^m > 0$, where λ_0, λ_1, m and n are positive unknown parameters.

Define $P_{ij}(t; a) = P\{T(t+a) = j|T(a) = i\}$. Differential equations for these transition probabilities are derived from Kolmogorov's Forward Equations (Appendix A):

- $\frac{dP_{00}(t;a)}{dt} = -\lambda_0(a+t)^n P_{00}(t;a)$
- $\frac{dP_{01}(t;a)}{dt} = \lambda_0(a+t)^n P_{00}(t;a) - \lambda_1(a+t)^m P_{01}(t;a)$
- $\frac{dP_{02}(t;a)}{dt} = \lambda_1(a+t)^m P_{01}(t;a)$
- $\frac{dP_{11}(t;a)}{dt} = -\lambda_1(a+t)^m P_{11}(t;a)$

- $\frac{dP_{12}(t;a)}{dt} = \lambda_1(a+t)^m P_{11}(t;a)$.

and we also have $P_{ij}(t;a) \equiv 0$ for $j < i$, $P_{22}(t;a) \equiv 1$ for all $a \geq 0$, $t \geq 0$. We do not need to completely solve these equations as we only need some of the quantities for use with observed data. In particular, we obtain:

$$P_{00}(t;a) = exp\{-\lambda_0 \int_0^t (a+u)^n du\} \tag{2.1}$$

$$P_{01}(t;a) = \frac{\int_0^t exp\{\lambda_1 \int_0^u (a+v)^m dv\}\lambda_0(a+u)^n exp\{-\lambda_0 \int_0^u (a+v)^n dv\}du}{exp\{\lambda_1 \int_0^t (a+u)^m du\}} \tag{2.2}$$

$$P_{11}(t;a) = exp\{-\lambda_1 \int_0^t (a+u)^m du\} \tag{2.3}$$

with boundary conditions $P_{00}(0;a) = 1$, $P_{01}(0;a) = 0$ and $P_{11}(0;a) = 1$.

2.2.1 Selection of Functional Form of Transition Rates

The general estimation of m and n using expressions such as (2.2) is not easily performed. It is thus necessary to simplify the model. From Table 1.1, the rate of transition from 0 to 1 increases with age while the rate of transition from 1 to 2 decreases with age. If we assume these trends to be the truth, we will have $n > 0$ and $m < 0$. Cancer is believed to be a multi-step process so that there is more likely to be greater age variation prior to the preclinical state rather than after. Furthermore examination of Table 1.1 shows that when λ_0 and λ_1 are assumed constant within age ranges, λ_0 varies more across age groups than λ_1, i.e., $0.236/0.089 > 0.407/0.236$, suggesting that this quantity is more age dependent. We therefore follow Duffy and assume $m = 0$.

The incidence rate of breast cancer at age a is defined as

$$I(a) = \frac{dP_{02}(a;0)}{da} \frac{1}{1 - P_{02}(a;0)}$$

but $P_{02}(a;0) < 0.1$ for all ages of interest and hence

$$\begin{aligned} I(a) &\approx \frac{dP_{02}(a;0)}{da} \\ &= \lambda_1 P_{01}(a;0). \end{aligned}$$

In [1], [3] and [4], $m = 0$ and $n = 0$ which give constant transition rates that are independent of age. The above relationship and the solution of equation (2.2) shows that the Duffy model implicitly implies a non-increasing incidence rate, i.e.,

$$I(a) = \frac{\lambda_1 \lambda_0 (exp\{-\lambda_0 a\} - exp\{-\lambda_1 a\})}{\lambda_1 - \lambda_0},$$

which is not a realistic assumption for breast cancer as incidence of breast cancer is known to increase with age.

The data sets by age group are based on age at first screen and not age at current screen. The Duffy model thus "freezes" incidence rates based on age at first screen. It is possible for a patient in one age group to be diagnosed in the next age group. This is anomalous and we would like to have a common n and a common λ_0 across all age groups.

The observed incidence of breast cancer in British Columbia from 1988-1996 is shown in Table 2.1. Rate of Incidence versus age for $n = 0, 1, 2, 3$ are plotted in Figure 2.1 to determine the integer value of n that gives incidence rates close to the observed incidence of breast cancer. The values for $n = 2$ appear to be best suited for our purposes.

Age Group (years)	Number of Cases	B.C. Population	Average Incidence Rate per Year
0-4	0	1008666	0.00000
5-9	1	1027940	0.00000
15-19	1	997574	0.00000
20-24	11	1113083	0.00001
25-29	93	1276730	0.00007
30-34	283	1391295	0.00020
35-39	680	1341207	0.00051
40-44	1368	1209928	0.00113
45-49	1814	996669	0.00182
50-54	1616	775611	0.00208
55-59	1734	684900	0.00253
60-64	2034	671334	0.00303
65-69	2462	666514	0.00369
70-74	2449	587576	0.00417
75-79	1940	447535	0.00433
80-84	1260	296150	0.00425
85-89	628	155574	0.00405
90+	335	84832	0.00395

Table 2.1: Observed Incidence of Breast Cancer in British Columbia by five year age groups from 1988-1996.

Therefore, the solution to the true state process for general n is:

$$P_{00}(t; a) = exp\left\{\frac{\lambda_0 a^{n+1} - \lambda_0(a+t)^{n+1}}{n+1}\right\} \tag{2.4}$$

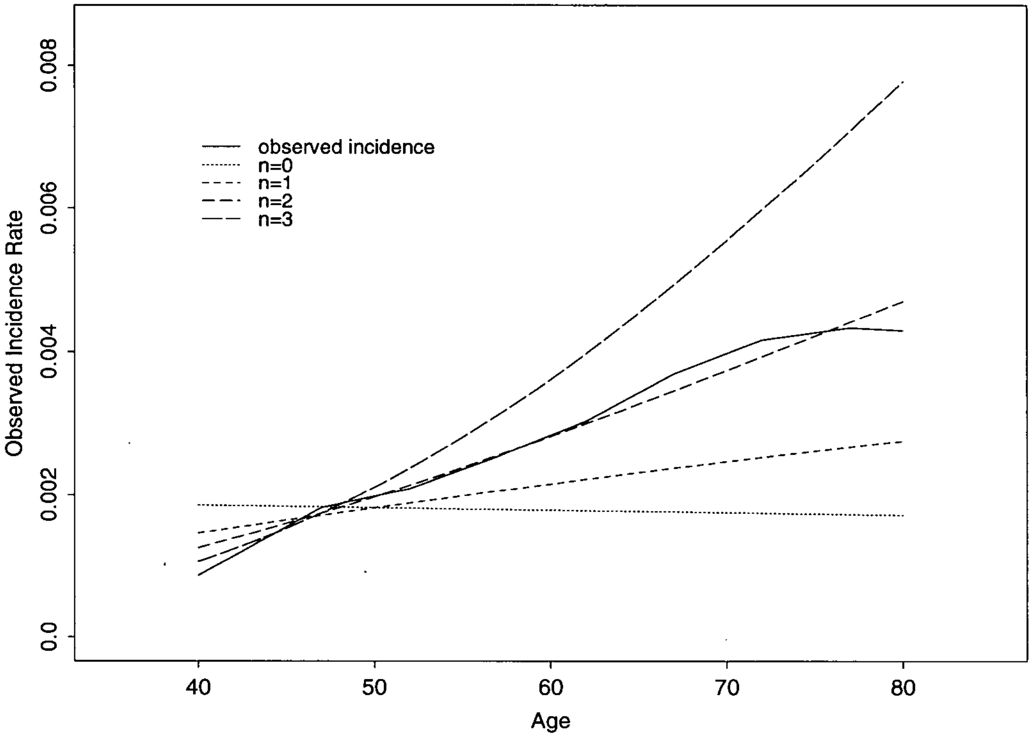

Figure 2.1: Observed Incidence versus Age overlaid by Incidence Rates for various n, $\lambda_1 = 0.4$, $\lambda_0 45^n \approx 0.002$.

$$P_{01}(t;a) = \lambda_0 exp\left\{\frac{\lambda_0 a^{n+1}}{n+1} - \lambda_1 t\right\} \int_0^t (a+u)^n exp\left\{\lambda_1 u - \frac{\lambda_0 (a+u)^{n+1}}{n+1}\right\} du \tag{2.5}$$

$$P_{11}(t;a) = P_{11}(t) = exp\{-\lambda_1 t\}. \tag{2.6}$$

Based on these probabilities, the model infers that a cancer's sojourn time in state 0 with $a = 0$, S_0, has a Weibull distribution function $P(S_0 \leq t) = 1 - exp\left\{-\frac{\lambda_0 t^{n+1}}{n+1}\right\}$, while the sojourn time distribution in state 1, S_1, is Exponential, i.e., $P(S_1 \leq t) = 1 - exp\{-\lambda_1 t\}$. The structure of the model also implies that all cancers start in state 0 and pass through state 1 before entering state 2.

2.3 Observed State Process

The difference between the observed state process and the true state process occurs as a result of screening errors. If screening was perfect, i.e., a true no cancer is identified as no cancer detected and a true preclinical cancer is identified as a screen detected cancer, then the observed state process would be equivalent to the true state process. In reality, this is not the case and must be accounted for.

Let $\{O(a)\}$ be the observed state process where at age a, a patient is in one of three states:

$$O(a) = \begin{cases} 0, & \text{if no cancer detected at age} \leq a \\ & \text{and no mammography performed at } a; \\ 1, & \text{if screen detected cancer at age} \leq a; \\ 2, & \text{if clinical cancer detected at age} \leq a \\ & \text{with no preceding screen detected cancer.} \end{cases}$$

A no cancer detection occurs when the screen result is normal or if the screen result is abnormal and work-up does not lead to a cancer diagnosis. A screen detected cancer occurs when the screen result is abnormal and work-up leads to a cancer diagnosis. A clinically diagnosed cancer is one which is made as a result of signs and symptoms. Note that the monitoring of preclinical and clinical cancer are done on different time scales. Each patient is constantly monitored for clinical cancer and it is assumed that this is done without error, i.e., $P(O(a) = 2|T(a) = 2) = 1$ and $P(O(a) = 2|T(a) \neq 2) = 0$. Periodically, each patient is monitored for preclinical cancer by screening mammography which is done with error. The error can be one of two types. For age a,

$$P(O(a) = 1|T(a) = 0) = \gamma \text{ (false positive rate)}$$

$$P(O(a) = 0|T(a) = 1) = \beta \text{ (false negative rate)}$$

A false positive is defined as a definitively diagnosed cancer that would never have arisen clinically in the absence of screening. A false negative occurs when a preclinical detectable cancer is present and mammography does not identify it. The complement of the false negative rate is sensitivity and that of the false positive rate is specificity.

Now suppose the study runs over a certain period of time t^* which will be called the end of the study period. For a subject aged a^* at t^* we define a new state as

$$O(a^*) = 3, \quad \begin{array}{l} \text{if no screen detected cancer or clinical cancer prior} \\ \text{to } a^* \text{ and no mammography performed at } a^*. \end{array}$$

A description of the data provided by the SMPBC will be given in the next section. This, in addition to the two sections after it, will relate the true state process to the observed state process.

2.4 Description of Observational Data

The SMPBC is a routine service screening programme that started in 1988 where the data available is up to the end of 1996. A woman is eligible to enter the study if she resides in British Columbia in this period of time and has never had a prior clinically diagnosed breast cancer. Suppose we call this woman the i^{th} patient. Every patient is assumed to be born breast cancer free and will first go for screening at age a_{i1}. If no cancer is detected at a_{i1} they will continue to go for screening until one of the following events occur:

1. screen detected cancer
2. clinical cancer
3. end of the study period (1996/12/31).

Every patient starts with an observation period from birth to 1996/12/31. If a patient has a screen detected cancer or a clinically diagnosed cancer prior to the end of the study period then their observation period is truncated to the time of diagnosis.

Suppose the i^{th} patient goes for screening at ages a_{i1}, a_{i2}, $\ldots$, a_{iK_i}, where the event at a_{iK_i} may not be the result of a screen but may be a clinical cancer or the end of the study period. By definition, every event but the last will result in a no cancer observation. Let O_i^* be their observed state at the last event. Then the i^{th} patient's contribution to the likelihood will be

$$L(O_i^*) = P\{O(a_{i1}) = 0, \ldots, O(a_{i,K_i-1}) = 0, O(a_{iK_i}) = O_i^* | T(a_{i1}) \neq 2\}$$

where a_{ij} is age at event j and K_i is the number of events. The conditioning event is $\{T(a_{i1}) \neq 2\}$ because patients who have a no cancer result or a screen detected cancer at first screen are from a cohort of women followed from birth who have never had a prior clinical breast cancer. Hence, a clinical cancer diagnosis is only observed after one or more screens.

2.5 Patient History with One Screen

Suppose K_i=1. Then the i^{th} patient's contribution to the likelihood will be

$$L(O_i^*) = P\{O(a_{i1}) = O_i^* | T(a_{i1}) \neq 2\}.$$

- If O_i^*=0 (no cancer detected at the end of observation period) then

  $$L(O_i^*) = \frac{P_{00}(a_{i1};0)(1-\gamma) + P_{01}(a_{i1};0)\beta}{P_{00}(a_{i1};0) + P_{01}(a_{i1};0)}.$$

  This situation will only arise if the i^{th} patient enters the study at exactly the end of the study period.

- If O_i^*=1 (screen detected cancer at the end of observation period) then

  $$L(O_i^*) = \frac{P_{00}(a_{i1};0)\gamma + P_{01}(a_{i1};0)(1-\beta)}{P_{00}(a_{i1};0) + P_{01}(a_{i1};0)}.$$

2.6 Patient History with Multiple Screens

Suppose $K_i > 1$. Then the i^{th} patient's contribution to the likelihood becomes quite complicated. To aid discussion consider the following example.

2.6.1 Example.

If K_i=5 then the observed sequence is:

Time (age)	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}
Observed	0	0	0	0	O_i^*

Table 2.2: Observed Sequence, $K_i = 5$.

For events prior to the last, the true state process $\{T(t)\}$ can only take values 0 or 1, where possible values can only be consecutive 0's followed by consecutive 1's. To simplify the notation, drop the i subscript and let $a_0 = 0$, $t_j = a_j - a_{j-1}$ for $j \geq 1$, and define

$$P^*_{0k}(a_j; 0) = \frac{P_{0k}(a_j; 0)}{P_{00}(a_1; 0) + P_{01}(a_1; 0)}$$

for $j \geq 0$, $k = 0, 1$.

Now consider all screens but the last.

Time (age)	a_1	a_2	a_3	a_4	
Observed	0	0	0	0	
					Probability of True Sequence
Possible	0	0	0	0	$P^*_{00}(a_4; 0)$
	0	0	0	1	$P^*_{00}(a_3; 0)P_{01}(t_4; a_3)$
True (T)	0	0	1	1	$P^*_{00}(a_2; 0)P_{01}(t_3; a_2)P_{11}(a_4 - a_3)$
	0	1	1	1	$P^*_{00}(a_1; 0)P_{01}(t_2; a_1)P_{11}(a_4 - a_2)$
Sequences	1	1	1	1	$P^*_{01}(a_1; 0)P_{11}(a_4 - a_1)$

Table 2.3: Probabilities of True Sequences.

Time (age)	a_1	a_2	a_3	a_4	
Observed	0	0	0	0	
					Probability of Observing the True Sequence
Possible	0	0	0	0	$(1-\gamma)^4 P^*_{00}(a_4;0)$
	0	0	0	1	$(1-\gamma)^3 P^*_{00}(a_3;0)P_{01}(t_4;a_3)\beta$
True (T)	0	0	1	1	$(1-\gamma)^2 P^*_{00}(a_2;0)P_{01}(t_3;a_2)\beta^2 P_{11}(a_4-a_3)$
	0	1	1	1	$(1-\gamma) P^*_{00}(a_1;0)P_{01}(t_2;a_1)\beta^3 P_{11}(a_4-a_2)$
Sequences	1	1	1	1	$P^*_{01}(a_1;0)\beta^4 P_{11}(a_4-a_1)$

Table 2.4: Probabilities of Observing the True Sequences.

2.6.2 Generalization from Example.

For an observed sequence of 0's we have two types of true sequences; one that has 0 as the true state at age a_{K_i-1} and one that has 1. For $T(a_{K_i-1}) = 0$, define

$$\begin{aligned} F_0(a_{K_i-1}) &= P\{O(a_1)=0,...,O(a_{K_i-1})=0, T(a_{K_i-1})=0\} \\ &= (1-\gamma)^{K_i-1}P^*_{00}(a_{K_i-1};0), \end{aligned}$$

and for $T(a_{K_i-1}) = 1$,

$$\begin{aligned} F_1(a_{K_i-1}) &= P\{O(a_1)=0,...,O(a_{K_i-1})=0, T(a_{K_i-1})=1\} \\ &= \sum_{j=0}^{K_i-2} P^*_{00}(a_j;0)(1-\gamma)^j P_{01}(t_{j+1};a_j)\beta^{K_i-j-1}P_{11}(a_{K_i-1}-a_{j+1}). \end{aligned}$$

Therefore the i^{th} patient's contribution to the likelihood is:

- if O^*_i=0 (no cancer detected at the end of observation period),

$$\begin{aligned} L(O^*_i) &= F_0(a_{K_i-1})[P_{00}(t_{K_i};a_{K_i-1})(1-\gamma) + P_{01}(t_{K_i};a_{K_i-1})\beta] + \\ & F_1(a_{K_i-1})P_{11}(t_{K_i})\beta. \end{aligned}$$

This situation will only occur if the i^{th} patient has multiple screens and a no cancer resulting screen at exactly the end of the study period.

- if O_i^*=1 (screen detected cancer at the end of observation period),

$$\begin{aligned} L(O_i^*) &= F_0(a_{K_i-1})[P_{00}(t_{K_i}; a_{K_i-1})\gamma + P_{01}(t_{K_i}; a_{K_i-1})(1-\beta)] + \\ &\quad F_1(a_{K_i-1})P_{11}(t_{K_i})(1-\beta). \end{aligned}$$

- if O_i^*=2 (clinical cancer at the end of observation period),

$$L(O_i^*) = F_0(a_{K_i-1})\lambda_1 P_{01}(t_{K_i}; a_{K_i-1}) + F_1(a_{K_i-1})\lambda_1 P_{11}(t_{K_i}).$$

  We assume a patient transitions from state 0 to state 1 (or from state 1 to state 1) in time t_{K_i} and then instantaneously transitions from 1 to 2. This assumption follows from the error-free constant monitoring of clinical cancer.

- if O_i^*=3 (no screen detected cancer or clinical cancer prior to the end of study period),

$$L(O_i^*) = F_0(a_{K_i-1})[P_{00}(t_{K_i}; a_{K_i-1}) + P_{01}(t_{K_i}; a_{K_i-1})] + F_1(a_{K_i-1})P_{11}(t_{K_i}).$$

  This situation arises if the i^{th} patient's last screen detects no cancer and it occurs prior to the end of the study period.

As usual the likelihood function is given by

$$L(\beta, \gamma, \lambda_0, \lambda_1, n|\mathbf{a}) = \prod_{i=1}^{N} L(O_i^*)$$

where N is the number of patients.

Chapter 3

Data Generation for Simulation and Results

The parameters are estimated via minimization of the negative log likelihood function which will be accomplished using a Quasi-Newton algorithm [5]. In order to test that this routine works properly and that the likelihood is 'well behaved', simulations on generated data from known distributions were carried out. The distributions that will be important for generating data are the Uniform, Exponential, Weibull and Bernoulli distributions.

3.1 Review of Distributions

A random variable X has a Bernoulli distribution with probability of success $p > 0$ if

$$P(X = x) = p^x(1-p)^{1-x}, \;\; x = 0, 1.$$

A random variable X has a uniform distribution on (0,1) if

$$f_X(x) = \begin{cases} 1, & 0 < x < 1; \\ 0, & \text{otherwise.} \end{cases}$$

A random variable X has an Exponential distribution with scale $\frac{1}{\lambda_1} > 0$ if

$$f_X(x) = \lambda_1 exp\{-\lambda_1 x\}, \;\; x > 0.$$

A random variable X has a Weibull distribution with scale $b > 0$ and shape $a > 0$ if

$$f_X(x) = \frac{ax^{a-1}}{b^a} exp\{-(x/b)^a\}, \;\; x > 0.$$

Another form of the Weibull distribution is arrived at by letting $b^a = \frac{a}{\lambda_0}$, i.e.,

$$f_X(x) = \lambda_0 x^{a-1} exp\{-\lambda_0 x^a/a\}.$$

3.2 True State Model Sojourn Times

For data generation and simulation purposes, let $n = 2$. The history of every patient starts at $T(0)$=0. Let S_0 be the sojourn time in state 0. Then for time $t > 0$,

$$\begin{aligned} P(S_0 > t) &= P(T(t) = 0 | T(0) = 0) \\ &= exp\{-\lambda_0 t^3/3\}, \end{aligned}$$

i.e., S_0 has distribution function $F_{S_0}(t) = 1 - exp\{-\lambda_0 t^3/3\}$. Note that this distribution function is Weibull. Now, if u is a random sample from a Uniform (0,1) distribution, then sojourn times in state 0 can be generated by setting $F_{S_0}(t) = u$ and solving for t which gives $t = \{-3\ln(1-u)/\lambda_0\}^{1/3}$.

Similarly, let S_1 be the sojourn time in state 1. Then for time $t > 0$ and age $a > 0$,

$$\begin{aligned} P(S_1 > t) &= P(T(t+a) = 1 | T(a) = 1) \\ &= exp\{-\lambda_1 t\}, \end{aligned}$$

i.e., S_1 has distribution function $F_{S_1}(t) = 1 - exp\{-\lambda_1 t\}$ which is that of an Exponential.

3.3 Data Generation

Step 1. Choose λ_0 and λ_1 and generate S_0 and S_1 for patient history.

Step 2. Specify screening and end of study times and generate true states at screening times based on patient history.

Step 3. Generate observed states at screening times. This is done in the following way:

- if the true state at some time t is 0, then $P(O(t) = 0 | T(t) = 0) = 1 - \gamma$. A random sample of size 1 from a Bernoulli distribution with $p = \gamma$ is taken to get an observed state at t of 0 or 1.

- if the true state at some time t is 1, then $P(O(t) = 1 | T(t) = 1) = 1 - \beta$. A random sample of size 1 from a Bernoulli distribution with $p = 1 - \beta$ is taken to get an observed state at t of 0 or 1.

Step 4. The observed state sequence for each patient will be a sequence of 0's, 1's and 2's but we only want the observed sequence up to the first 1 or

2. A 'valid' observed screening history is created by truncating the observed sequence at the earliest 1 or 2.

An end of study period will also be defined. If a patient has not entered state 1 or 2 by the end of the study period, then their observed history will end in a 3. This implies that their true state at this time is either 0 or 1.

3.4 Examples

Let the screen times = {40,45,50,55,60} and the end of study period = 59.

1. If the generated sojourn times are S_0=47, S_1=7 then the true state sequence would be {0,0,1,2,2}.

- Suppose an observed sequence of {0,1,0,2,2} is generated. Then we would truncate the sequence to {0,1} with time={40,45}.
- Suppose an observed sequence of {0,0,0,2,2} is generated. Then we would truncate the sequence to {0,0,0,2} with time={40,45,50,54}.
- Suppose an observed sequence of {0,0,1,2,2} is generated. Then we would truncate the sequence to {0,0,1} with time={40,45,50}.

2. If the generated sojourn times were S_0=54, S_1=7 then the true state sequence would be {0,0,0,1,1}.

- Suppose an observed sequence of {0,0,0,0,0} is generated. Then we would truncate the sequence to {0,0,0,0,3} with time={40,45,50,55,59}.

- Suppose an observed sequence of {0,0,0,1,0} is generated. Then we would truncate the sequence to {0,0,0,1} with time={40,45,50,55}.

- Suppose an observed sequence of {0,0,0,0,1} is generated. Then we would truncate the sequence to {0,0,0,0,3} with time={40,45,50,55,59}.

3. If the generated sojourn times were S_0=54, S_1=4 then the true state sequence would be {0,0,0,1,2}.

- Suppose an observed sequence of {0,0,0,0,2} is generated. Then we would truncate the sequence to {0,0,0,0,2} with time={40,45,50,55,58}.

- Suppose an observed sequence of {0,0,0,1,2} is generated. Then we would truncate the sequence to {0,0,0,1} with time={40,45,50,55}.

Therefore, all observed screening histories will end in a 1, 2 or 3. An observed history ending in 0 will only occur if there is a no cancer resulting screen at exactly the end of study period. For the above example, if $S_0 + S_1 \leq 40$, which is the age at first screen, then we do not consider this sojourn time pair in the analysis as state 2 is only observed subsequent to one or more screens.

3.5 Simulation Results

The maximum likelihood estimates of the model parameters are obtained using a Quasi-Newton algorithm which is an iterative procedure and hence starting values must be specified.

We did several simulations each consisting of generating 1000 samples of size 500. Each sample consisted of 500 observed sequences created as described in Section 3.3 and Section 3.4. The parameters are set to specific values for data generation and starting values. The first simulation was the simplest situation of error-free screening. This occurs when the observed states are the true states, i.e., $\beta = 0$ and $\gamma = 0$. Hence, rather than minimizing the negative log likelihood with respect to four parameters, it is minimized with respect to just λ_0 and λ_1. The starting values were set at $\lambda_0 = 0.00001$ and $\lambda_1 = 0.3$.

Parameter	True Value	mean(MLE)	SD(MLE)
λ_0	0.00001	0.00000953	0.00000046
λ_1	0.3	0.29890707	0.00444783

Table 3.1: Simple Model Simulation Results for 1000 Samples of Size 500, $\beta = 0$, $\gamma = 0$.

The next simulations involved minimizing the negative log likelihood function with respect to all four parameters. We set β, γ, λ_0 and λ_1 to several different values for data generation and starting values. It should be noted that sparse number of screen detected cancers and clinical cancers lead to estimates with higher variability.

In conclusion, the algorithm produced appropriate parameter estimates and appeared satisfactorily robust. We have applied it to SMPBC data as detailed in the next chapter.

Parameter	True Value	mean(MLE)	SD(MLE)
β	0.2	0.2000	0.0001
γ	0.1	0.1001	0.0004
λ_0 ($\times 10^3$)	0.01	0.0094	0.0006
λ_1	0.3	0.3001	0.0003
β	0.1	0.1050	0.1277
γ	0.01	0.0104	0.0104
λ_0 ($\times 10^6$)	6.0	5.9454	1.0281
λ_1	0.3	0.3439	0.1740

Table 3.2: Full Model Simulation Results for 1000 Samples of Size 500.

Chapter 4

Results for SMPBC Data

Summaries of the SMPBC data by age groups specified by age at first screen are presented in Tables 4.1, 4.2 and 4.3. A summary of the frequency of events at the end of observation period and average time between screens is contained in Table 4.1. The frequency of type of transition is in Table 4.2 and the frequency of number of screens per patient is in Table 4.3.

Table 4.4 shows the results of simultaneous estimation of transition rates, sensitivity and specificity under Duffy's assumptions for transition rates.

Table 4.5 shows the results of simultaneous estimation of transition rates, sensitivity and specificity for our model with common n and λ_0 across age groups. The maximum likelihood estimates of γ for the 40-49 and 50-59 year age groups was of the order 10^{-10} and was not significantly different from 0. Hence, the optimization was done with $\gamma = 0$ for these two age groups.

Event at	Frequencies for the following age groups:			
last screen	40-49 years	50-59 years	60-69 years	70-74 years
	(N=116526)	(N=76403)	(N=59022)	(N=18425)
No Cancer (O^*=0)	41 (0.0004)	22 (0.0003)	17 (0.0003)	6 (0.0003)
Screen Detected Cancer (O^*=1)	506 (0.0043)	640 (0.0084)	840 (0.0142)	312 (0.0169)
Clinical Cancer (O^*=2)	353 (0.0030)	264 (0.0035)	223 (0.0038)	79 (0.0043)
End of Study Period (O^*=3)	115626 (0.9923)	75477 (0.9879)	57942 (0.9817)	18028 (0.9785)
Average Time Between Screens (Years)	1.365	1.291	1.256	1.240

Table 4.1: Event Frequencies at last screen and Average Time Between Screens by Age Group, SMPBC (Respective proportions in parentheses).

	Transition	To State				
Age Group	From State	0	1	2	3	Total
40-49 years	0	274221	506	353	115626	390706
50-59 years	0	200898	640	264	75477	277279
60-69 years	0	160284	840	223	57942	219289
70-74 years	0	46865	312	79	18028	65284

Table 4.2: Frequency of Transitions by Age Group, SMPBC.

Number of	Frequencies for the following age groups:			
Screens	40-49 years	50-59 years	60-69 years	70-74 years
1	275	348	442	185
2	48206	26932	19294	6495
3	26192	16366	12565	4024
4	16387	11231	8869	2784
5	11599	8923	7195	2149
6	7798	6624	5639	1675
7	4145	3972	3498	880
8	1292	1276	980	163
9	559	636	447	58
10	73	95	93	12

Table 4.3: Frequency of Number of Screens by Age Group, SMPBC.

Parameter	Results for the following age groups:			
	40-49 years	50-59 years	60-69 years	70-74 years
λ_0 ($\times 10^2$)	0.171 (0.150-0.191)	0.259 (0.229-0.290)	0.310 (0.268-0.352)	0.378 (0.308-0.448)
λ_1	0.879 (0.649-1.110)	0.587 (0.512-0.661)	0.453 (0.359-0.547)	0.377 (0.271-0.483)
MST($= 1/\lambda_1$)	1.137 (0.901-1.542)	1.705 (1.513-1.954)	2.206 (1.828-2.783)	2.653 (2.070-3.693)
β, false negative rate	0.110 (0-0.220)	0.015 (0-0.064)	0.067 (0-0.179)	0.096 (0-0.258)
γ, false positive rate	0.0004 (0-0.001)	0.0001 (0-0.001)	0.0010 (0-0.002)	0.0008 (0-0.002)
$-\log L(.)$	6478.94	6322.81	6922.83	2445.90

Table 4.4: Results of estimation for parameters from SMPBC, Duffy assumptions, $n = 0$ (Respective 95% Confidence Intervals in parentheses).

Parameter	Results for the following age groups:			
	40-49 years	50-59 years	60-69 years	70-74 years
n	1.861 (1.613-2.109)			
$\lambda_0(\times 10^6)$	1.499 (1.239-1.759)			
λ_1	0.623 (0.479-0.767)	0.531 (0.461-0.602)	0.413 (0.337-0.489)	0.356 (0.259-0.453)
MST$(= 1/\lambda_1)$	1.604 (1.304-2.085)	1.882 (1.660-2.171)	2.422 (2.046-2.966)	2.810 (2.207-3.866)
β, false negative rate	0.175 (0.047-0.303)	0.027 (0-0.093)	0.062 (0-0.158)	0.090 (0-0.223)
γ, false positive rate	0	0	0.0003 (0-0.0008)	0.0001 (0-0.0010)
$-\log L(.)$	22135.75			

Table 4.5: Results of estimation for parameters from SMPBC, common n and λ_0 across all age groups (Respective 95% Confidence Intervals in parentheses).

Chapter 5

Discussion

The advancement of the model we introduce is a smooth age-dependent transition rate into the preclinical detectable phase. The likelihood becomes extremely complicated as it is composed of many integrals that do not have a closed form and must be calculated using numerical methods [6]. Our model indicated that women who go for their first screen prior to fifty years of age not only have a shorter mean sojourn time but also poorer sensitivity of the screening tool than women who go for their first screen subsequent to fifty years of age. The screening interval should be shorter for younger age groups and longer for older age groups. In general, our maximum likelihood estimates for λ_1 increase with age which follows the same trend as previously documented rates.

The most drastic difference in parameter estimates occurs at age 60. The 95% confidence intervals for λ_1 from the 40-49 and 50-59 year age groups do not contain the estimates of λ_1 from the 60-69 and 70-74 age groups and vice versa.

For our model, the estimates of specificity in age groups 40-49 and 50-59 were not significantly different from 1, i.e., there is little to no over diagnosis in 40-59 year old women being screened. The estimates of specificity in the 60-69 and 70-74 year age groups are close to 1 as well. The sensitivity does not seem to follow any sort of trend from age group to age group.

The difference between the results of our model and Duffy's model can be seen by comparing Table 4.4 and Table 4.5. Our model has longer mean sojourn times, smaller false positive rates and larger false negative rates (except for the 70-74 year age group). This can partly be explained by the following. Suppose the time between screens is fixed and λ_1 is allowed to vary. If the mean sojourn time increases then the probability that a patient is screened at an age for which $T(a) = 1$ increases, i.e., a longer mean sojourn time implies a higher number of screen detected cancers, but the number of observed screen detected cancers is fixed. This quantity can remain fixed if the mean sojourn time increases and the false negative rate increases so that more screen detected cancers occur due to the mean sojourn time increasing but a drop in the number of screen detected cancers will occur due to the false negative rate increasing. In all age groups, the false positive rate, γ, is so small that it can be effectively set to be zero.

The sum of negative log likelihood's across age groups is smaller for our model (22135.75) compared to that of Duffy's model (22170.48) which is an indicator of a better fitting model. In addition to this, we also have fewer parameters across all age groups, 12 parameters vs. 16 parameters. An incidence by age plot based on the parameters from both models is shown in Figure 5.1. Our model clearly gives incidence rates more closely related to the observed incidence in the general British Columbia population than Duffy's model.

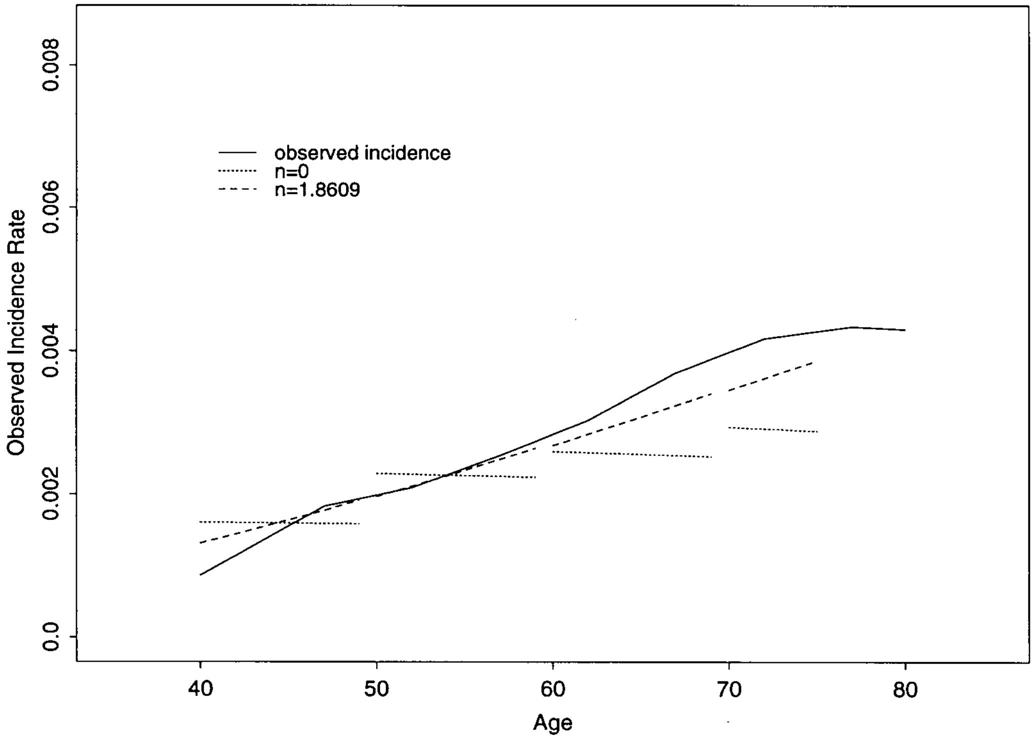

Figure 5.1: Observed Incidence versus Age overlaid by Incidence Rates calculated from Parameter Estimates for Our Model and for Duffy's Model.

By comparing Table 1.1 with Table 4.4 , it appears that women attending the SMPBC in the 90's differ from women in Sweden in the 80's. The difference in all the parameter estimates may be attributed to several things. It may be due to improvements in the screening tool over the years and it may have to do with Duffy's assumption of same time between screens for each patient.

The current policy for the SMPBC is women less than 50 years of age are invited for screening with mammography annually while women greater than 50 years of age are invited for screening with mammography every two years. Based on our results, the time between screens could be lengthened several months for women less than 50 years of age and shortened several months for women 50 to 60 years of age.

Appendix A

Kolmogorov's Forward Equations

The following is an illustration of how the differential equations were derived from Kolmogorov's Forward Equations.

Definition A.1 *For all states i, j and times $a \geq 0$, $s \geq 0$, $t \geq 0$, the Chapman-Kolmogorov equations are:*

$$P_{ij}(t+s;a) = \sum_{k=0}^{\infty} P_{ik}(t;a)P_{kj}(s;a+t). \tag{A.1}$$

From (A.1), we get

$$\begin{aligned} P_{ij}(t+\Delta t;a) - P_{ij}(t;a) &= \sum_{k=0}^{\infty} P_{ik}(t;a)P_{kj}(\Delta t;a+t) - P_{ij}(t;a) \\ &= \sum_{k \neq j} P_{ik}(t;a)P_{kj}(\Delta t;a+t) - \\ &\quad P_{ij}(t;a)[1 - P_{jj}(\Delta t;a+t)] \end{aligned}$$

and thus

$$\begin{aligned}\lim_{\Delta t\to 0}\frac{P_{ij}(t+\Delta t;a)-P_{ij}(t;a)}{\Delta t} &= \lim_{\Delta t\to 0}\sum_{k\neq j}P_{ik}(t;a)\frac{P_{kj}(\Delta t;a+t)}{\Delta t}-\\ &\lim_{\Delta t\to 0}P_{ij}(t;a)\frac{1-P_{jj}(\Delta t;a+t)}{\Delta t}.\end{aligned}$$

As the limit and the summation can be interchanged [7] we obtain

$$\frac{dP_{ij}(t;a)}{dt}=\sum_{k\neq j}P_{ik}(t;a)\nu_{kj}(a+t)-P_{ij}(t;a)\mu_j(a+t) \qquad \text{(A.2)}$$

where

$$\nu_{kj}(t)=\lim_{\Delta t\to 0}\frac{P_{kj}(\Delta t;t)}{\Delta t}$$

and

$$\mu_j(t)=\lim_{\Delta t\to 0}\frac{1-P_{jj}(\Delta t;t)}{\Delta t}.$$

We refer to (A.2) for all states i, j and times $a \geq 0$, $t \geq 0$ as Kolmogorov's Forward Equations.

A.1 Example

Kolmogorov's Forward Equation when $i = 0$ and $j = 1$ is

$$\begin{aligned}\frac{dP_{01}(t;a)}{dt} &= \sum_{k\neq 1}P_{0k}(t;a)\nu_{k1}(a+t)-P_{01}(t;a)\mu_1(a+t)\\ &= P_{00}(t;a)\nu_{01}(a+t)-P_{01}(t;a)\mu_1(a+t).\end{aligned}$$

Now from properties 1, 2 and 3 of Section 2.2 with $\lambda_0(a) = \lambda_0 a^n$ and $\lambda_1(a) = \lambda_1 a^m$, we have

$$\begin{aligned}
\nu_{01}(a+t) &= \lim_{\Delta t \to 0} \frac{P_{01}(\Delta t; a+t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\lambda_0 (a+t)^n \Delta t + o(\Delta t)}{\Delta t} \\
&= \lambda_0 (a+t)^n
\end{aligned}$$

and

$$\begin{aligned}
\mu_1(a+t) &= \lim_{\Delta t \to 0} \frac{1 - P_{11}(\Delta t; a+t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\lambda_1 (a+t)^m \Delta t + o(\Delta t)}{\Delta t} \\
&= \lambda_1 (a+t)^m
\end{aligned}$$

and hence

$$\frac{dP_{01}(t;a)}{dt} = P_{00}(t;a)\lambda_0(a+t)^n - P_{01}(t;a)\lambda_1(a+t)^m.$$

Bibliography

[1] Duffy, S. W., Chen, H. H., Tabar, L., Day, N. E. (1995). Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statist. Med.*, **14**, 1531-1543.

[2] Bryant, H. E., Brasher, P. (1994). Risk and probabilities of breast cancer: short-term versus lifetime probabilities. *Can. Med. Assoc. J.*, **150**, 211-216.

[3] Chen, H. H., Duffy, S. W., Tabar, L. (1996). A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *The Statistician*, **45**, 307-317.

[4] Duffy, S. W., Chen, H. H., Tabar, L., Fagerberg, G., Paci, E. (1996). Sojourn time, sensitivity and positive predictive value of mammography screening for breast cancer in women aged 40-49. *Int. J. Epidem.*, **25**, 1139-1145.

[5] Nash, J. C. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation, 2nd edition.* New York: Hilger.

[6] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (1992). *Numerical Recipes in C : The Art of Scientific Computing, 2nd edition.* New York: Cambridge University Press.

[7] Ross, S. M. (1993). *Introduction to Probability Models, 5th edition.* Boston: Academic Press.

[8] Taylor, H. M., Karlin, S. (1984). *An Introduction to Stochastic Modelling.* Boston: Academic Press.

[9] Schechter, M. T., Miller, A. B., Baines, C. J., Howe, G. R. (1986). Selection of women at high risk of breast cancer for initial screening. *J. Chron. Dis.*, **39**, 253-260.

[10] Shapiro, S. (1977). Evidence on screening for breast cancer from a randomized trial. *Cancer*, **39**, 2772-2782.