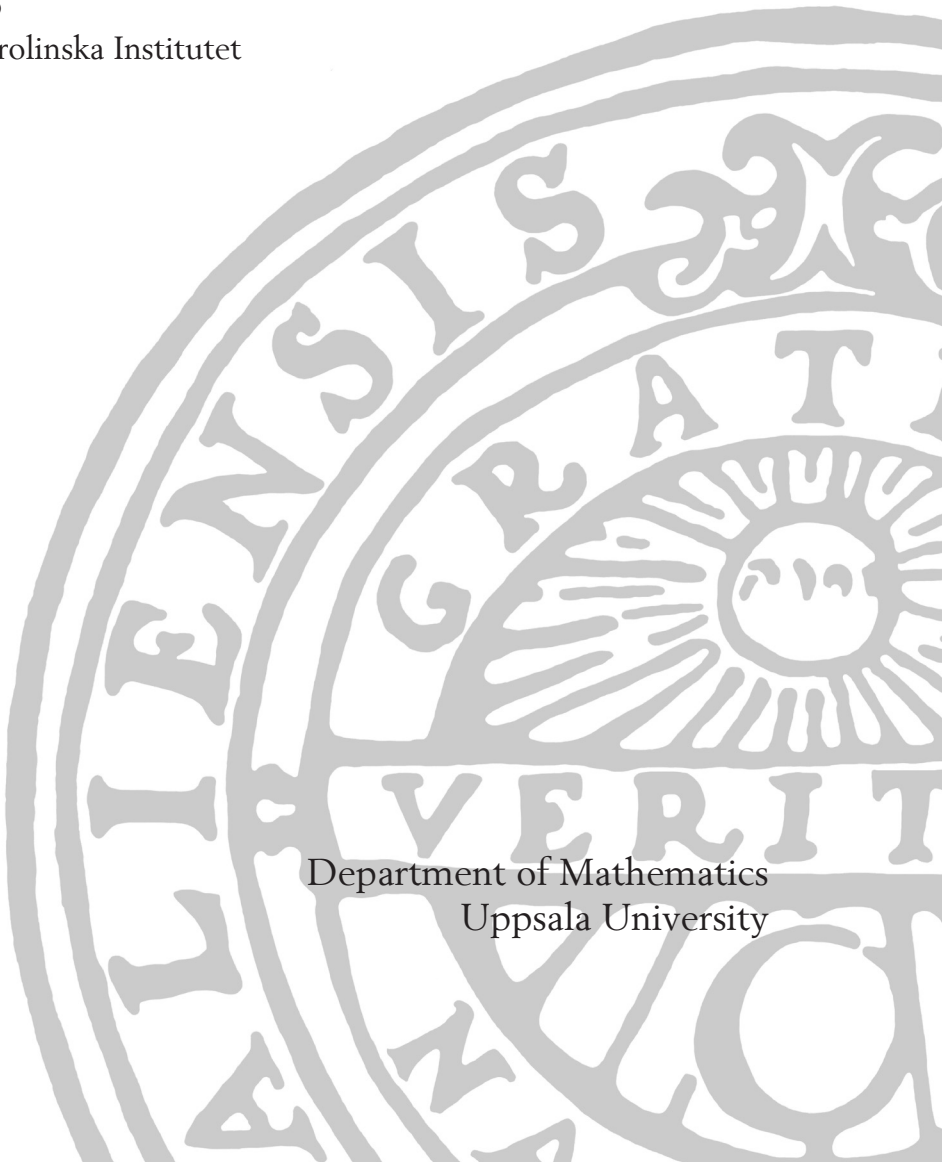
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2012:4

Statistical models of breast cancer tumour growth for mammography screening data

Linda Abrahamsson

Examensarbete i matematik, 30 hp
Handledare: Keith Humphreys, Karolinska Institutet
Examinator: Silvelyn Zwanzig
Februari 2012



Department of Mathematics
Uppsala University

Statistical models of breast cancer tumour growth for mammography screening data

Linda Abrahamsson

February 20, 2012

Abstract

For evaluating breast cancer screening programs it is important to know something about tumour growth rates and the sensitivity of the screening, and also their risk factors. Relevant variables, such as tumour size, are typically only observable at time of diagnosis. How can one estimate tumour growth when the size of each tumour is only measured once? There exists information in differences between tumours found at screening and tumours found clinically. Stochastic models of cancer development and detection can therefore be constructed, which yield the distribution of observable variables at diagnosis. In many studies multi-state Markov models have been used in which the tumour passes through different states. The approach used here is to model tumour growth with a continuous function. Likelihood theory is used to find estimates for the risk factor parameters.

Sammanfattning

För att kunna utvärdera screeningprogram i bröstcancer är det viktigt att ha kunskap om tumörtillväxt, screeningens sensitivitet samt riskfaktorer till dessa. Relevanta variabler, såsom tumörstorlek, är oftast observerbara endast vid diagnostillfället. Hur kan man skatta tumörtillväxtens hastighet när varje tumörs storlek endast är känd vid ett tillfälle? I olikheterna mellan tumörer funna vid screening och tumörer funna kliniskt finns information. Det är därför möjligt att konstruera stokastiska modeller gällande tumörers utveckling och detektion, vilka kan ge fördelningar för observerbara variabler vid diagnos. I många studier har Markov modeller använts, där tumören antas passera genom olika faser. Vårt tillvägagångssätt är att anta en kontinuerlig funktion för tumörtillväxten. Teori om likelihoodfunktioner används för att skatta riskfaktorernas parametrar.

Acknowledgement

I would like to thank my supervisor Keith Humphreys for introducing me to this topic and also for his continuous help and guidance during this thesis.

Contents

List of Figures

List of Tables

1 Introduction

Breast cancer is the most common cancer type for women in Sweden. Around 7.000 women get the diagnosis every year, of these 1.500 die from the disease. The number of diagnosed breast cancers seems to increase with time but the mortality is decreasing. In Sweden it is recommended for women between 40 and 75 years old to attend screening every other year [25] [15].

At screening most tumours are found, but factors like a small tumour size and/or high mammographic density can make it harder to see the tumour at screening. Tumours not found at screening can be found at the next screening or clinically between two screening rounds. There is a possiblity though, that they will never be found. In a population not going to screening all tumours which will be found, are detected clinically. In this paper we refer to women with tumours found between screenings as interval cases, women with tumours found at screening as screening cases and women with tumours found in the absence of a screening program as clinical cases.

1.1 Epidemiology

Many studies have been made in breast cancer patients to find risk factors for the malignancy. One of the strongest risk factors is family history. Having a mother or sister with a breast cancer diagnosis gives a relative risk for breast cancer of 1.5-3.0 compared to if no mother or sister had breast cancer. Some genes have been found which are linked to breast cancer risk. Around 5-10 percent of all breast cancer cases are thought to be inheritable [13]. High hormone levels seem to play an important role in breast cancer since many established risk factors are associated with hormones. Some factors increasing the risk for breast cancer are pregnancy (though in the long-term it gives protection), high estrogen levels, early age at menarche, late age at menopause, late age at first birth, low parity, postmenopausal hormone use, moderate alcohol intake, and adult weight gain [13].

Some factors that decrease the risk are breast-feeding, physical activity and increased intake of fruits and vegetables [13]. Coffee consumption intake has been found to decrease the risk of some subtypes of breast cancer [18]. High mammographic density is also a risk factor for breast cancer[17]. See Figure 1 from [24] for examples of how screening pictures (a selection with different densities) look like and information about the relative risks in different density groups.

Age and mammographic density are associated, with density decreasing with age (particularly at menopause). The density is thought to depend on hormone levels, therefore the lower density for older women. Breast cancer risk increases with age, but younger women tend to have faster growing tumours (Weedon-Fekjær et al. [30]).

It is not known whether mammographic density affects tumour growth, but it does make the sensitivity of the screening lower.

Hormone Replacement Therapy, HRT, influences breast cancer risk. It has been found to give higher incidence, but better survival [4]. It also has an impact on density and possibly on tumour growth.

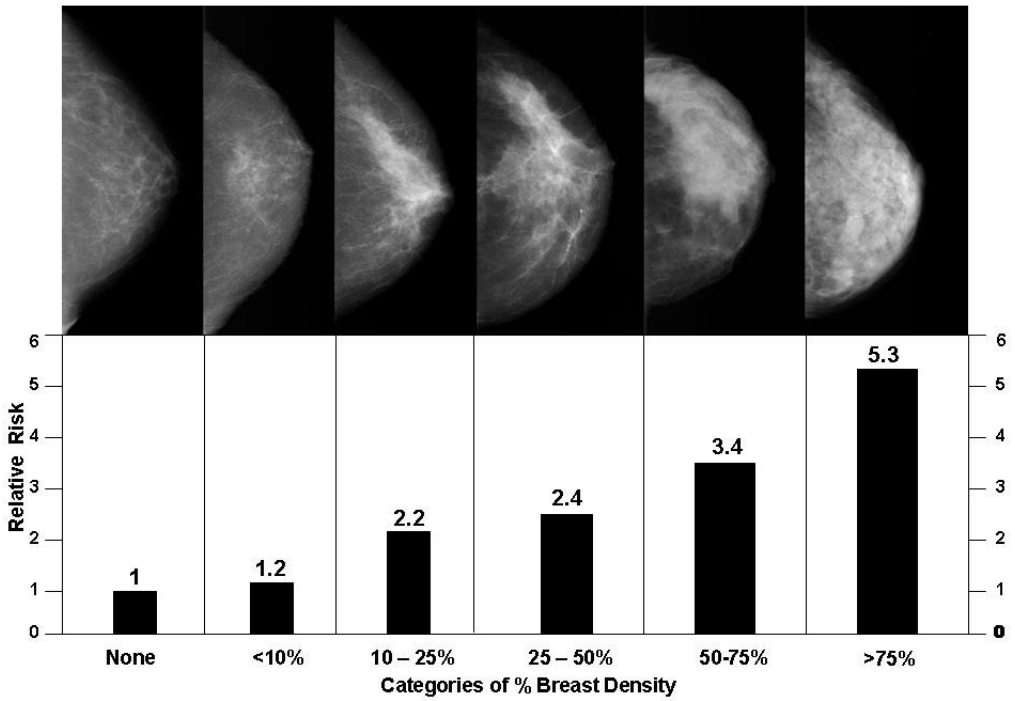


Figure 1: Different densities and the relative risks of getting a tumour. Source: [24].

1.2 Background to this study

To be able to decrease the morbidity and mortality in breast cancer it would be good to know more about the tumour growth. Knowledge about how fast tumours grow can help in the design of good and effective screening programs. Such information has been used by Forastero et al. [11] in a microsimulation study. In their study they examined the effect of different time intervals between screenings and between which ages women should be screened, on the effectiveness of screening.

The exposure to x-rays is thought to be a risk factor for inducing breast cancer, according to the National Health System Service Breast Screening Program (NHS-BSP) [11]. The risk is not that big but still, due to this fact and high costs the number of screenings can not be too many for each person. It is also not efficient to screen the oldest in the population, since they are likely to be able to live the rest of their lives with a tumour without getting problems from it. In their study Forastero and colleagues found through simulations that the best screening program screens women between 50 and 70 every other year. To be able to make more individual screening programs we also want to know more about different risk factors, how they affect growth and the sensitivity of the type of screening that is used.

One of the strongest risk factors for breast cancer is mammographic density. It has been found by Mandelson et al. [19] that a woman with high mammographic density has a higher risk of being an interval case than a woman with low mammographic density. But is this fact due to lower sensitivity at the screening, faster growing tumours, or both, for the women with high density?

Another question easier to answer knowing the tumour growth is whether it is the screening effect or the adjuvant therapy effect which has decreased mortality in breast cancer lately. This has been studied by the Cancer Intervention and Surveillance Modeling Network (CISNET) and a review has been made by Berry et al. [3].

It has been found in studies that the mortality has decreased with around 25 percent by using screening [4]. If screening programs were more effective the mortality could be even lower.

1.2.1 The Karma study

The Karma study, led by Professor Per Hall at Karolinska Institutet, is the biggest breast cancer study so far in Sweden. It is a cohort in which the women will be followed during ten years. Data on about 100.000 women going to mammographic screening is being collected. The goal with the study is to be able to find the individual risk for getting breast cancer, for each woman. It is hoped that knowledge of risk could guide individualized preventive treatment programs and that individual screening programs could be made so that, for example, women with high risk are screened often and women with low risk are screened more seldom. For all the women in the study facts about lifestyle, mammographic density and genetic variations will be sampled so that the risks can be measured. By the sixth of February 2012, 26.071 women had been enrolled in the study. [15]

Already mammographic density is known to be a big risk factor for breast cancer. Since the way of measuring mammographic density lately is much more effective and reliable, this factor is supposed to give even more knowledge about how density affects breast cancer in various ways and knowing the densities for women in the future will tell much about how big the risk is for them to get breast cancer [17].

In the future, data from the Karma study can be used in the estimation of tumour growth.

2 Tumour growth modelling and estimation procedures

Different types of statistical study designs, models and estimation procedures have been used to estimate tumour growth. Two examples of common study designs are cohort studies and case-control studies. Tumour growth can be modeled by a multistate (Markov) model or a continuous model. These models have been applied both in the absence of screening and in the presence of screening.

2.1 Cohort and case-control studies

In a cohort study, a type of longitudinal study, a defined population of persons is followed over a long time period. The cohort can for example consist of all persons free of a particular disease at time of study onset and the disease can be the outcome of interest. If the disease is a common one and the population is big

enough, some persons will most likely get the disease. Then a comparison of healthy and sick persons can be made to assess whether an exposure affects the incidence. If the disease is breast cancer and the cohort follows a screening programme, it is possible to compare different groups of the persons with breast cancer. There exists information in differences between screening cases and clinical cases. From a cohort study it is also possible to calculate incidence rates of a disease if one assumes that the study population is a random sample of the population at interest.

A case-control study is often used when a rare disease is going to be investigated. Two different groups of persons are chosen, one with persons having the disease, the cases, and one with persons not having the disease, the controls. In such a study the incidence rate cannot be calculated but still associations between exposure and outcome can be evaluated. In a breast cancer case control study one could choose cases from both persons detected after screening and at screening. Information about time, for example time since last screening can hopefully be found retrospectively while in a cohort study such information is already available.

It can be possible to make a cohort study out of case-control data in some special cases. This has been done by Biesheuvel et al. in study [4]. There the cases, which are most of the cases in the whole population from January 1993 to April 1995, have been taken out from an earlier case-control study. Persons with a breast cancer diagnosis are used in the study and information about time aspects are found retrospectively. The cases were then divided into different types of cases to be able to evaluate differences between the groups.

2.2 Multistate Markov discrete growth approaches

A common way to mathematically model tumour growth is to use Markov chains in continuous time. Firstly a countable set of states, S, is defined.

Definition 1. *Let $X(t)$, $t \geq 0$, be a family of random variables taking values in a countable set S, called the state space. Then $X(t)$ is a Markov chain if for any collection $k, j_1, \ldots, j_{n-2}, j$ of states, and $t_1 < t_2 < \cdots < t_n$ of times, we have*

$$P(X(t_n) = k | X(t_1) = j_1, \ldots, X(t_{n-1}) = j) = P(X(t_n) = k | X(t_{n-1}) = j).$$

In Definition 1, from [28], the Markov property has been used. This property must hold for all Markov models and it states that: *Given a complete description of the state of the process at time t, its future development is independent of its track record up to t* [28]. If a Markov chain is time homogeneous then for the probabilities in Definition 1, $P(X(t_n) = k | X(t_{n-1}) = j) = p_{jk}(t_n - t_{n-1})$, holds. These probabilities can be summarized in the matrix of transition probabilities, $\mathbf{P} = (p_{jk}(t))$, $j, k \in S$. It is also possible to find the $\mathbf{Q}$-matrix, sometimes called the generator of the Markov chain.

Definition 2. *For $j, k \in S$ and $j \neq k$ $p_{jk}(t) = q_{jk}(t) + o(t)$ for small t. $q_{jk}(t)$ is called the transition rate from state j to state k. Now $\mathbf{Q} = (q_{jk}(t))$ where $q_{jj}(t) = 1 - \sum_{k \neq j} q_{jk}(t)$.*

Theorem 1. *If the state space is finite and* $\mathbf{P}'$ *is the matrix with the derivatives of the transition probabilities in* $\mathbf{P}$*, then* $\mathbf{P}' = \mathbf{PQ}$ *holds.*

In Theorem 1 the so called *Chapman-Kolmogorov equations* have been used to derive the *forward equations* $\mathbf{P}' = \mathbf{PQ}$. To estimate tumour growth parameters Duffy and colleagues [8] [9] propose a basic three state Markov model for the tumour development. The states are 0 = "no disease", 1 = "preclinical but detectable tumour" and 2 = "clinical tumour". All potential tumours, even the ones which don't exist, are in state 0 until they are detectable at mammographic screening, when they remove to state 1. Once the tumour can be found clinically, it is in state 2. There is no possibility to go from a higher state to a lower state. Here is the matrix $\mathbf{Q}$ with the general transition rates for the model, the rows represent the starting states.

$$\boldsymbol{Q} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 \\ 0 & -\lambda_2 & \lambda_2 \\ 0 & 0 & 0 \end{bmatrix}$$

In this Markov model the times for transition from state 0 to state 1 and from state 1 to state 2 are exponentially distributed. In [6] and [29], Day and Walter found that this was a reasonable assumption.

To be able to form a procedure for the estimation of the transition rates, $\boldsymbol{P}$ needs to be known. Using Theorem 1, $\mathbf{P}' = \mathbf{PQ}$ must hold. This results in nine differential equations. Here is the transition probability matrix:

$$\boldsymbol{P} = \begin{bmatrix} e^{-\lambda_1 t} & \frac{\lambda_1(e^{-\lambda_2 t} - e^{-\lambda_1 t})}{(\lambda_1 - \lambda_2)} & 1 + \frac{\lambda_2 e^{-\lambda_1 t}}{(\lambda_1 - \lambda_2)} + \frac{-\lambda_1 e^{-\lambda_2 t}}{(\lambda_1 - \lambda_2)} \\ 0 & e^{-\lambda_2 t} & 1 - e^{-\lambda_2 t} \\ 0 & 0 & 1 \end{bmatrix}.$$

The solutions can be found in Appendix A.

By using the probabilities $p_{jk}(t)$ it is now possible to estimate two different parameters of interest in the Markov model; the screening test sensitivity (STS_M), and the mean sojourn time (MST_M). The following definitions have been stated by Duffy et al. [8].

Definition 3. STS_M *is the fraction of tumours in state 1 which are found at screening.*

Definition 4. MST_M *is the mean time a tumour spends in state 1.*

Other terms/states are also used in the literature (Figure 2). The lead time is the detection time gained because of the use of screening.

For the estimation of MST Duffy et al. [8] propose that an estimate, $\hat{\lambda_2}$, for λ_2 can be used:

$$\hat{MST} = \frac{1}{\hat{\lambda_2}}. \tag{1}$$

Parameters left to estimate are STS, λ_1 and λ_2. These can be estimated in different ways. One way is to use the two-stage estimation procedure used in [8]. Firstly,

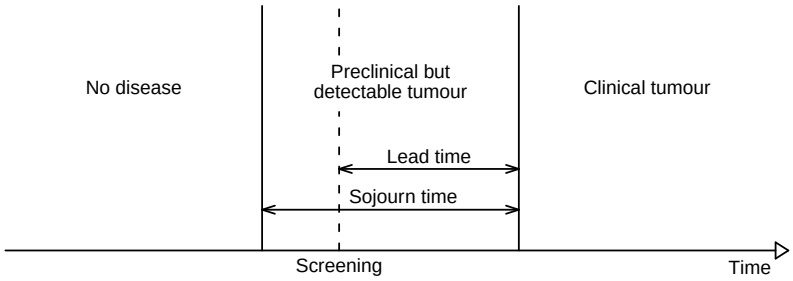


Figure 2: Different states/terms used in the Markov discrete growth model. Adapted from [9].

the transition rates are estimated under the assumption that the sensitivity is 100 percent. Given the transition rates, the STS can subsequently be estimated. Duffy et al. [8] states that this procedure is unsatisfactory. Instead they propose, in [9], that one of the methods by Prevost et al. [22] used on screening for colorectal cancer, should be used. This estimation procedure, presented as model 2 in [22], is also discussed by Weedon-Fekjær et al. [31].

For estimation, a likelihood function $L_M = L_{1M}L_{2M}$ is proposed. L_{1M} is a Poisson distribution with intensity $I(t)$, for the number of interval cases, r_t, in the interval (t-1,t] years after a negative screening. L_{2M} is a Binomial distribution, with parameters n_s and p, for the number of detected tumours at screening, c. In this likelihood function expressions for the intensity $I(t)$ and for the probability p must be derived. Let us start with $I(t)$. The interval cases can be of two different types, either a tumour was in the preclinical stage at screening and was overlooked, or it came to that stage after screening and had no chance to be a screening case. Remember now that the duration time of the preclinical screening detectable phase is exponential with intensity λ_2, this implies that

$$P(\textit{Time to transition to clinical state} \leq t | \textit{In preclinical stage at } t = 0) = 1 - e^{-\lambda_2 t}.$$

If the intensity of preclinical disease, λ_1, is known we can model the number of interval cases during one year (t-1,t] given that the tumour was in state 0 at screening as $\lambda_1(1 - e^{-\lambda_2(t-0.5)})$. Here the mean point of the time interval was used. The parameter λ_1 is estimated as the incidence in an unscreened population.

It is also possible to calculate the number of interval cases which were overlooked at screening. If c is the number of cancers detected at screening and S is the sensitivity then the expected number of cancers overlooked is $\frac{c(1-S)}{S}$. The probability

for such a tumour to be clinically detected in time period (t-1,t] is

$$P(t-1 < \textit{Time to transition to clinical state} \leq t | \textit{ In preclinical stage at } t=0)$$
$$= e^{-\lambda_2(t-1)} - e^{-\lambda_2 t}.$$

It is now possible to calculate the number of interval cases, $I(t)$, in interval (t-1,t] as the sum of the interval cases which were overlooked at the screening and the interval cases with an onset time after the screening:

$$I(t) = \lambda_1(1 - e^{-\lambda_2(t-0.5)}) + \frac{c(1-S)}{S}(e^{-\lambda_2(t-1)} - e^{-\lambda_2 t}). \tag{2}$$

Regarding p, we can start by expressing it in a new way: $p = \frac{Pre}{n_s}$ where Pre is the prevalence (the number of women eith detected tumours at the screening) and n_s is the number of screened women. For the prevalence, Pre, it is possible to use the transition probabilities. We also use the number of screened women, n_s and the age at screening, T.

$$Pre = n_s S \cdot P(\textit{Tumour in preclinical state} | \textit{ Not clinically detected})$$
$$= n_s S \cdot \frac{p_{01}(T)}{p_{00}(T) + p_{01}(T)} = n_s S \cdot \frac{\frac{\lambda_1(e^{-\lambda_2 T} - e^{-\lambda_1 T})}{\lambda_1 - \lambda_2}}{e^{-\lambda_1 T} + \frac{\lambda_1(e^{-\lambda_2 T} - e^{-\lambda_1 T})}{\lambda_1 - \lambda_2}}. \tag{3}$$

If c and r_t are observed, parameter estimates for S and λ_2 can be found through maximation of the following likelihood function:

$$\binom{n_s}{c} p^c (1-p)^{n_s - c} \prod_{i=1}^{k} \frac{I(t_i)^{r_{t_i}} e^{-I(t_i)}}{r_{t_i}}, \tag{4}$$

where i is a specific time interval.

This multi-state model can be made more realistic by adding more states. Duffy et al. [9] describe a five-state model which incorporates the axillary lymph node to be either positive or negative. Oortmarssen et al. [20] use a very complex model with many states to evaluate different screening programs. They divide the tumour sizes into three different intervals which means that there are three preclinical states, three clinical states, and so on.

2.3 Continuous growth approaches

It is possible to estimate tumour growth rates without using Markov models. One can postulate a continuous function for the tumour growth and it is then possible to estimate its parameters using data. Some different growth models have been proposed. Although the growth rate can vary for each tumour, most models assume that the tumour growth follows a smooth increasing function. This assumption is thought to work well on the population level [30].

In the study made by Bartoszyński et al. [2] the tumour's growth is assumed to depend on the cell reproductive rate, through an exponential growth function

with a constant doubling time. When the tumour gets bigger the growth rate can be assumed to decrease as the nutrition becomes limited [30]. Bartoszyński et al. describe different models, one in which the growth is assumed to be exponential regardless of the size of the tumour. The volume (measured in mm^3) for a tumour t years after the onset time point follows the following function:

$$V(t) = V_{cell}e^{t/r} \tag{5}$$

Here V_{cell} is the volume of one cell and r is the inverse growth rate. If r is assumed to be deterministic, this model doesn't fit real data very well. A proposed solution to this problem is to model the inverse growth rate as an outcome from a gamma distributed variable, R, with shape parameter τ_1 and inverse scale parameter τ_2. A gamma distribution includes the gamma function $\Gamma(\tau_1)$ which is defined as

$$\Gamma(\tau_1) = \int_0^\infty t^{\tau_1 - 1}e^{-t}dt.$$

The more general lower incomplete gamma function is defined as

$$\gamma(\tau_1, x) = \int_0^x t^{\tau_1 - 1}e^{-t}dt.$$

The density function for R is as follows:

$$f_R(r) = \frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)}r^{\tau_1 - 1}e^{-\tau_2 r},\ r \geq 0. \tag{6}$$

This model has been used by Plevritis et al. [21]. To allow for the growth rate to decrease for bigger tumours Bartoszyński et al. propose the use of a Gompertz or logistic function. Those curves are sigmoidal functions which are quite similar.

Spratt et al. [26] [27] compare the general logistic function, the Gompertz function and the exponential function to see which one has the best fit to real data. They, however, never used the gamma distribution for the inverse growth rate. An assumption made by Spratt et al. was that the maximum number of cell doublings should be 40 corresponding to a tumour with diameter 128 mm. In the range of tumour sizes found in the early clinical period they found that the growth curve is well described by a specific logistic function. The least impressive model was the exponential. To model the individual growth rate per time unit (one year) the lognormal distribution was found to be a good assumption. The model they proposed is:

$$V(t) = \frac{V_{max}}{\left[1 + \left(\left(\frac{V_{max}}{V_{cell}}\right)^{1/c} - 1\right)e^{-\frac{1}{c}\kappa t}\right]^c}, \tag{7}$$

where κ is lognormally distributed with logmean α_1 and logvariance α_2, i.e.

$$f_\kappa(x) = \frac{1}{x\sqrt{2\pi\alpha_2}}e^{-\frac{(logx - \alpha_1)^2}{2\alpha_2}},\ x > 0. \tag{8}$$

The constant $c = 4$ was found to give the best fit. This model found by Spratt and colleagues is used by Weedon-Fekjær et al. [30] [31] and in the study made by Forastero et al. [11].

Hart et al. [14] found that the best model is the power law growth. The exponential function is in this family of functions. Those functions were found to be better than the sigmoidal functions like the logisitic and Gompertz functions.

Let's now go further to see how one can estimate tumour growth depending on what type of data that is available.

2.3.1 Approaches in the absence of screening

Plevritis et al. [21] have proposed an approach based on using a population of women not attending screening. Tumour growth can be assessed in terms of various characteristics. In this project we primarily focus our attention on tumour size, but in the approach presented by Plevritis et al. [21] tumour stages, divided into local stage, regional stage and distant stage, are also considered. For their model with respect to tumour size Plevritis et al. [21], as Bartoszyński et al. [2], assume that the tumour volume grows exponentially from a starting volume V_{cell}, from a sphere with a diameter of two millimeters. Note that the value of V_{cell} here is different than in the model of Bartoszyński et al. [2] and that the natural history of the tumour is not modeled prior to V_{cell}. The volume at time t follows equation (5) in which r is an outcome from the gamma distributed random variable R, see equation (6) for the density function. One more assumption proposed by Bartoszyński et al. [2] has been adopted by Plevritis et al. [21]. It is the assumption that the time to clinical detection from the time the tumour has the volume V_{cell}, represented by the random variable T_{det}, depends on the size of the tumour which is assumed to be spherical. Plevritis et al. assume that

$$P(T_{det} \in [t, t + dt) | T_{det} > t) = \gamma V(t)dt + o(dt). \tag{9}$$

This is also called the hazard function when dt goes to 0; see the following definitions from [16].

Definition 5. *The survival function is the probability that a random variable X is bigger than a value x, i.e. $S_X(x) = 1 - F_X(x) = P(X > x)$.*

Definition 6. *The hazard rate is defined by*

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x}. \tag{10}$$

If the hazard function is known it is possible to calculate the survival function using

$$S_X(x) = exp(-\int_0^x h(t)dt); \tag{11}$$

see [16].

Plevritis et al. [21] also propose a model for a random variable T_{reg} which represents the time between when the tumour is of size V_{cell} and when the tumour transitions to the regional stage. Following is the hazard function:

$$P(T_{reg} \in [t, t+dt) | T_{reg} > t) = \eta V(t) dt + o(dt). \tag{12}$$

Further let T_{dist} be the random variable representing time between when the tumour is of size V_{cell} and when the tumour transitions to the distant stage. Here follows the hazard function for T_{dist}:

$$P(T_{dist} \in [t, t+dt) | T_{dist} > t, T_{reg} = t_{reg}) = \omega V(t) dt + o(dt),\ t > t_{reg}. \tag{13}$$

When $t \leq t_{reg}$ the probability is 0. From the assumptions above Plevritis et al. show how it is possible to derive expressions for three different volume distributions: at clinical detection, at the transition from local to regional stages and at the transition from regional to distant stages. Here it is shown how the first distribution can be derived.

Let V be the random variable for the volume at clinical detection and let the inverse growth rate R be gamma distributed, with the density function (6). The density function for the volume at clinical detection is found by

$$f_V(v) = \int_0^\infty f_{V,R}(v,r)dr = \int_0^\infty f_V(v|R=r) \cdot f_R(r)dr = \int_0^\infty \frac{d}{dv} F_V(v|R=r) \cdot f_R(r)dr.$$

Using the relation between the survival function and the hazard function shown in equation (11) the conditional distribution function can be written as

$$\begin{aligned} F_V(v|R=r) &= P(V(T_{det}) < v | R=r) = P(T_{det} < V^{-1}(v) | R=r) \\ &= P(T_{det} < R \cdot log\tfrac{v}{V_{cell}} | R=r) = P(T_{det} < r \cdot log\tfrac{v}{V_{cell}}) \\ &= 1 - S_{T_{det}}(r \cdot log\tfrac{v}{V_{cell}}) = 1 - exp(-\textstyle\int_0^{V^{-1}(v)} h(t)dt) \\ &= 1 - exp(-\textstyle\int_0^{V^{-1}(v)} \gamma V(t)dt) = 1 - exp(-\gamma r(v - V_{cell})). \end{aligned} \tag{14}$$

Furthermore, we have

$$f_V(v|R=r) = \frac{d}{dv} F_V(v|R=r) = \gamma r \cdot exp(-\gamma r(v - V_{cell})),$$

$$\begin{aligned} f_{V,R}(v,r) &= f_V(v|R=r) \cdot f_R(r) = \gamma r \cdot exp(-\gamma r(v - V_{cell})) \cdot \tfrac{\tau_2^{\tau_1}}{\Gamma(\tau_1)} r^{\tau_1 - 1} exp(-\tau_2 r) \\ &= \tfrac{\gamma \tau_2^{\tau_1}}{\Gamma(\tau_1)} r^{\tau_1} exp(-r(\tau_2 + \gamma(v - V_{cell}))), \end{aligned} \tag{15}$$

$$f_V(v) = \int_0^\infty f_{V,R}(v,r)dr = \gamma \tau_2^{\tau_1} \tau_1 (\tau_2 + \gamma(v - V_{cell}))^{-(\tau_1+1)}, \tag{16}$$

$$F_V(v) = \int_{V_{cell}}^v f_V(x)dx = 1 - \left(\frac{\tau_2}{\tau_2 + \gamma(v - V_{cell})} \right)^{\tau_1},\ v \geq V_{cell}. \tag{17}$$

The volume distributions at the transition from local to regional stage and at the transition from regional to distant stage can be derived in similar ways; see [21] for details. By using the three volume distributions and the hazard rates (9) (12) (13) a multinomial distribution is assumed and derived for the joint distribution of tumour size and stage at clinical detection, see [21] for an expression. Using this distribution

a likelihood function is calculated and parameter estimates are found by maximizing this function given some observed data. For identifiability it is assumed that $\tau_1 = \tau_2$ in the gamma distribution (6) which means that the expected value of the inverse growth rate is always one. Bartoszyński et al. [2] describe another parametrization which can be used in order to estimate both τ_1 and τ_2 without assuming equality. By using that γR is gamma distributed instead of only R the identification problem is eliminated.

Chia et al. [5] proposed an alternative model. Instead of assuming an exponential growth the growth is modeled as a geometric Brownian motion. The following definitions come from [28].

Definition 7. *A random process, $B(t)$, is a standard Brownian motion if*

- $B(0) = 0$,
- $B(t)$ *is continuous for* $t \geq 0$,
- $B(t)$ *has independent increments and* $B(t+s) - B(s)$ *has the Normal distribution* $N(0, t)$, *for all* $s, t \geq 0$.

A standard Brownian motion can also be called a standard Wiener process.

Definition 8. *If B(t) is a standard Brownian motion and*

$$Y(t) = e^{\mu t + \sigma B(t)},$$

then Y(t) is a geometric Brownian motion.

Chia et al. use that the volume at time t is

$$V(t) = V_{cell} e^{\mu t + \sigma B(t)}. \tag{18}$$

Here V_{cell} is used as a starting value for the geometric Brownian motion instead of 1. Different algorithms and techniques are used to estimate the unknown parameters. Essential differences between Plevritis et al. and Chia et al. are:

- Fixed trajectories with variation between individuals (Plevritis et al.) versus non-fix trajectories with variation in time within individuals (Chia et al.).
- A tumour can't decrease in size (Plevritis et al.) versus a tumour have the possibility to regress (Chia et al.).

2.3.2 Approaches in the presence of screening

A population attending screening (only one occassion is assumed here) can be divided into four different subgroups according to timing of tumour onset and detection (timing and mode). Only the women with detected tumours are regarded. The groups can be seen in Figure 3. In a) the onset and detection occur before screening. In b) the group of screening-detected women can be found. The interval cases can be of two different types, either the onset is before the screening as in c), or the onset is after the screening as in d).

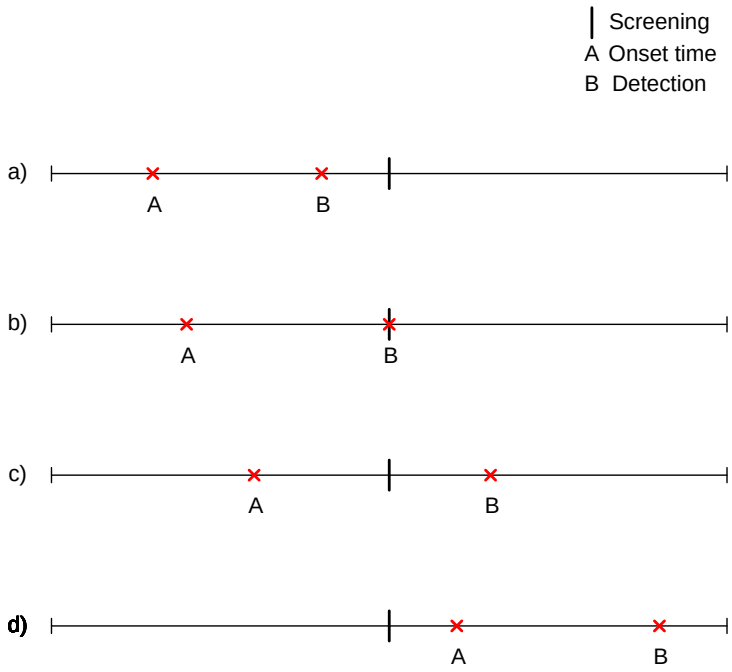


Figure 3: Possible courses of events which can occur among the women with detected tumours in a screening population.

Weedon-Fekjær et al.'s approach

In the presence of screening Weedon-Fekjær et al. [30] [31] have proposed an approach to estimate tumour growth, although it relies on the availability of an external data set collected from an earlier population not attending screening. Not only the tumour growth is modeled but also the screening test sensitivity. Based on estimates of parameters in their models it is possible to estimate a mean sojourn time. New definitions have to be made since there are no states assumed as in the Markov models. Since no real definitions have been proposed by Weedon-Fekjær et al. for the continuous model, we introduce reasonable definitions here.

Definition 9. *The STS is the probability that an existing tumour is found at screening.*

Definition 10. *The MST is the mean time a tumour exists before it is clinically detetcted.*

In [30] the authors rely on information on tumour sizes at detection from both screening detected and interval detected cancers while in [31] the authors also rely

on information on tumour sizes at repeated screening occasions. As opposed to Plevritis et al. [21] Weedon-Fekjær et al.'s approach is not based on specifying hazard functions, neither on the exponential growth function. Instead Weedon-Fekjær et al. propose that the tumour growth model found by Spratt et al. [26] [27] should be assumed, i.e. they use growth function (7) and growth rate density function (8). Tumour diameters, rather than volumes, are typically measured. If all tumours are assumed to be spherical the volume can be calculated as a function of the diameter, d, of a tumour through

$$V(t) = \frac{4}{3}\pi\left[\frac{d(t)}{2}\right]^3. \tag{19}$$

An advantage of the growth model is that an earlier volume can be written as a function of a later one. Assume that we have two time points t_1 and t_2 where $t_1 < t_2$. For both time points the growth formula can be used.

$$V(t_1) = \frac{V_{max}}{\left[1 + \left(\left(\frac{V_{max}}{V_{cell}}\right)^{0.25} - 1\right)e^{-0.25\kappa t_1}\right]^4},$$

$$V(t_2) = \frac{V_{max}}{\left[1 + \left(\left(\frac{V_{max}}{V_{cell}}\right)^{0.25} - 1\right)e^{-0.25\kappa t_2}\right]^4}.$$

Combining the two formulas above

$$\Longrightarrow \left(\left(\frac{V_{max}}{V(t_1)}\right)^{0.25} - 1\right)e^{0.25\kappa t_1} = \left(\left(\frac{V_{max}}{V(t_2)}\right)^{0.25} - 1\right)e^{0.25\kappa t_2}$$

$$\Longleftrightarrow V(t_1) = \frac{V_{max}}{\left[1 + \left(\left(\frac{V_{max}}{V(t_2)}\right)^{0.25} - 1\right)e^{0.25\kappa t}\right]^4}, \tag{20}$$

which now is an expression without the value of V_{cell}. If one knows the volume at the later time point and the growth rate, it is possible to find the volume at an earlier time point by this back calculation. From here it is also possible to get an expression for the growth rate κ, given the tumour volumes,

$$\kappa = \frac{4}{t_2 - t_1} log\left(\frac{\left(\frac{V_{max}}{V(t_1)}\right)^{0.25} - 1}{\left(\frac{V_{max}}{V(t_2)}\right)^{0.25} - 1}\right). \tag{21}$$

The STS can be modeled as an increasing function of the tumour diameter (in mm) since bigger tumours are more likely to be detected at screening. A tumour can either be found or not at the screening, this gives us a binary random variable

depending on the covariate tumour size. It is also possible to let the STS depend on more than one covariate, the first value to use in a covariate vector is often 1. A logistic regression model is a model that can be used to find the probability of interest given a specific set of covariates, $\boldsymbol{x}$. Let $\mu = S(x)$ be the expected value for the binary variable given the covariates, it can also be seen as the probability for a tumour to be found given the specific set of covariates. The logistic regression model,

$$log\big(\frac{\mu}{1-\mu}\big) = \boldsymbol{\beta}\mathbf{x}, \tag{22}$$

is used so that the probabilities will lie between 0 and 1 [10]. $\boldsymbol{\beta}$ contains the corresponding coefficients to the covariates in the vector $\mathbf{x}$. This can be rewritten so that

$$S(\mathbf{x}) = \frac{\exp(\boldsymbol{\beta}\mathbf{x})}{1+\exp(\boldsymbol{\beta}\mathbf{x})}. \tag{23}$$

Weedon-Fekjær et al. [30] [31] use this model, with tumour size as a single covariate. The STS is modeled as

$$S(d) = \frac{\exp(\beta_1+\beta_2 d)}{1+\exp(\beta_1+\beta_2 d)}, \tag{24}$$

where the covariate d is the tumour diameter and $\boldsymbol{\beta} = [\beta_1, \beta_2]^T$ are the coefficients.

To be able to estimate the unknown coefficients $\alpha_1, \alpha_2, \beta_1$ and β_2 in formulas (7) and (24) a likelihood function $L = L_1 L_2$ is proposed by Weedon-Fekjær et al. [30]. The first likelihood is for the sizes of the tumours detected at screening and the second is for the incidence of interval cases. The likelihood function involves lots of integrals if time and tumour sizes are treated as continuous variables, so to ease calculations Weedon-Fekjær et al. discretize time and tumour sizes and use sums and small intervals instead. The likelihood function proposed by Weedon-Fekjær et al. resembles the likelihood function proposed by Prevost et al. [22], used for the Markov discrete growth model.

Likelihood 1 in Weedon-Fekjær's approach

For the number of screening detected tumours in different diameter size intervals, used in the first likelihood, a multinomial model is proposed. Here is the likelihood function:

$$L_1(o_{11}, o_{12}, ..., o_{1n}|\alpha_1, \alpha_2, \beta_1, \beta_2) = \frac{n!}{\prod_i o_{1i}!}\prod_i {p_i}^{o_{1i}}. \tag{25}$$

Here n is the number of screening cases and i is an index for the size interval. Further o_{1i} is the number of observed tumours in interval i and p_i is the probability for a tumour, given detection, to be in that interval. In the model it must hold that

$$\sum_i p_i = 1. \tag{26}$$

The probabilities, p_i, need to be modeled and this is done by using back calculating from the size distribution of the clinical cases and using the screening test sensitivity. In [30] Weedon-Fekjær et al. assume the size distribution of clinical cases is available as external information. To derive p_i we introduce some events:

- C_i = a tumour is in size interval i at screening.
- B = a tumour was found at screening.
- D_f = in the absence of screening, a woman will have a clinically detected tumour between $f-1$ to f months after the screening.
- E_g = a tumour detected clinically is in size interval g at detection.

Then

$$p_i = P(C_i|B). \tag{27}$$

The size diameter intervals are of equal length. Further by using Bayes theorem, see [1],

$$P(C_i|B) = \frac{P(B|C_i)P(C_i)}{P(B)}. \tag{28}$$

Also

$$\sum_i P(C_i) = 1, \tag{29}$$

shall be fulfilled. The smallest size interval is for the non-existing tumours, i.e. tumours with diameter 0 mm. Remember that by the law of total probability, see [1],

$$P(B) = \sum_i P(B|C_i)P(C_i). \tag{30}$$

Thus, needed to model are the probabilities $P(B|C_i)$ and $P(C_i)$. The first probability can be rewritten using the screening test sensitivity:

$$P(B|C_i) = S(d_i|\beta_1, \beta_2), \tag{31}$$

where d_i is the diameter in size interval i. Weedon-Fekjaær et al. suggest estimating $P(C_i)$ by back calculation using the distribution of tumour sizes in a population of women not attending screening. By using the law of total probability

$$P(C_i) = \sum_f P(C_i|D_f)P(D_f). \tag{32}$$

$P(D_f)$ can be estimated as the number of clinically detected tumours in time interval f, divided by the number of persons at risk at the time point when the screening should have taken place. $P(D_f)$ is assumed to equal a constant r for all time intervals, i.e.

$$P(D_f) = r, \ \forall f. \tag{33}$$

Regarding f it is not clear how many time intervals one shall use, but rather too many than too few. Further, $P(C_i|D_f)$ can be derived by using the law of total probability once more:

$$P(C_i|D_f) = \sum_g P(C_i|D_f \cap E_g)P(E_g|D_f). \tag{34}$$

In the absence of screening the size distribution of clinical cases should be equal in each time interval, that is $P(E_g|D_f) = P(E_g)$. Let us state the probabilities used in (34) in words:

- $P(C_i|D_f) =$ Probability that a clinical cancer is in size group i f months before detection.
- $P(C_i|D_f \cap E_g) =$ Probability that a clinical cancer of size g is in size group i f months before detection.
- $P(E_g|D_f) = P(E_g) =$ Probability that a tumour is in size group g at detection, in the absence of screening.

To make the equations a bit shorter, let $p_{igf} = P(C_i|D_f \cap E_g)$ and $q_g = P(E_g)$. The following formula is then used:

$$p_i = \frac{S(d_i) \sum_f r \sum_g q_g p_{igf}}{P(B)}. \tag{35}$$

To derive the value of q_g a population not attending screening is used. The relative proportion of clinically detected tumours in size interval g is assumed to be equal to q_g. To calculate p_{igf} the lognormal distribution of the growth rate is used by Weedon-Fekjær et al. This corresponds to assuming that the growth rate is independent of the size at clinical detection:

$$p_{igf} = P(C_i|D_f \cap E_g) = P(k_1 < \kappa < k_2|D_f \cap E_g) = P(k_1 < \kappa < k_2). \tag{36}$$

The boundaries k_1 and k_2 need to be derived. Let f be the time interval $[t_2 - t_1 - \delta, t_2 - t_1 + \delta)$ for some small δ. The relation

$$\kappa = \frac{4}{t_2 - t_1} log\left(\frac{\left(\frac{V_{max}}{V(t_1)}\right)^{0.25} - 1}{\left(\frac{V_{max}}{V(t_2)}\right)^{0.25} - 1} \right), \tag{37}$$

is used, by letting the middle point in interval f be the time in years, taking the middle point of the diameter interval g and transforming it to volume (instead of $V(t_2)$) and also transforming the two end points of the diameter interval i to volumes (instead of $V(t_1)$). An alternative way of calculating p_{igf} will be shown in the next section, which dosn't assume that the growth rate κ is independent of the size at clinical detection. To make sure that $\sum_i p_{igf} = 1$ a normalization is made:

$$\frac{p_{igf}}{\sum_i p_{igf}}. \tag{38}$$

Instead of maximizing the likelihood function an equal operation is to maximize the log likelihood function. Since constants don't add any information in the maximizing procedure these can be kept away. Doing this the log likelihood function l_1 is obtained as

$$l_1 = \sum_i o_{1i} log(p_i) = \sum_i o_{1i} log(S(d_i) \sum_f \sum_g q_g p_{igf}), \tag{39}$$

where the $S(d_i)$ is dependent on the parameter values β_1 and β_2 and p_{igf} is dependent on the parameters α_1 and α_2.

Likelihood 2 in Weedon-Fekjær's approach

The second likelihood is built on information about the interval cases, which here are defined as the tumours found clinically up to 2 years after the screening. The number of cancers in each time interval, j, is assumed to follow a Poisson distribution, with different intensities in different time intervals. A Poisson distribution is reasonable since the number of new interval cases in an interval is assumed to be quite low in comparison to the number of persons at risk. Here is the second likelihood function, consisting of 24 different assumed independent Poisson distributions, one for each interval:

$$L_2(o_{21}, o_{22}, ..., o_{2\,24}|\alpha_1, \alpha_2, \beta_1, \beta_2) = \prod_{j=1}^{24} \frac{e^{e_j} e_j^{o_{2j}}}{o_{2j}!}. \tag{40}$$

In the formula e_j is the expected number of interval cancers in interval j while o_{2j} is the observed number of interval cancers in interval j. Let PYR_j be the observed number of person years at risk in time interval j. To derive e_j the events introduced in the section describing Likelihood 1 are used once more.

- C_i = a tumour is in size interval i at screening.
- B = a tumour was found at screening.
- D_j = in the absence of screening, a woman will have a clinically detected tumour between $j-1$ to j months after the screening.
- E_g = a tumour detected clinically is in size interval g at detection.

If B^c is the complement to the event B, then e_j can be derived as:

$$e_j = PYR_j P(D_j|B^c) = PYR_j \frac{P(B^c|D_j)P(D_j)}{P(B^c)}, \tag{41}$$

by using Bayes theorem. As before $P(B^c)$ can be seen as a constant and $P(D_j) = r$. By the law of total probability applied two times

$$\begin{aligned} P(B^c|D_j) &= 1 - P(B|D_j) = 1 - \textstyle\sum_g P(B|D_j \cap E_g)P(E_g|D_j) \\ &= 1 - \textstyle\sum_g q_g \sum_i P(B|D_j \cap E_g \cap C_i)P(C_i|D_j \cap E_g) \\ &= 1 - \textstyle\sum_g q_g \sum_i S(d_i)p_{igj}. \end{aligned}$$

$$\Longrightarrow e_j = \frac{PYR_j r(1 - \sum_g q_g \sum_i S(d_i)p_{igj})}{P(B^c)}. \tag{42}$$

Assume now that r is unknown. If we assume that the sum of the expected cases is approximately equal to the sum of the observed number of cases, then the formula

$$e_j = \frac{PYR_j(1 - \sum_g q_g \sum_i S(d_i)p_{igj})}{\sum_j PYR_j(1 - \sum_g q_g \sum_i S(d_i)p_{igj})} \sum_j o_{2j}, \tag{43}$$

can be derived. The log likelihood will now be

$$l_2 = \sum_j (e_j + o_{2j} log(e_j) - log(o_{2j}!)). \tag{44}$$

Both likelihood functions in Weedon-Fekjær's approach
Having two likelihood functions makes all parameters identifiable in the approach proposed by Weedon-Fekjær et al. [30]. The joint log likelihood is

$$l = l_1 + l_2 = \sum_i o_{1i} log(p_i) + \sum_j (e_j + o_{2j} log(e_j) - log(o_{2j}!)). \tag{45}$$

Remember that $S(d_i)$ is dependent on the parameter values β_1 and β_2 and that p_{igf} and p_{igj} are dependent on the parameters α_1 and α_2.

Published estimates of tumour growth rates and STS using the Plevritis & Weedon-Fekjær methods
Weedon-Fekjær et al. used data from the Norwegian Breast Cancer Screening Program (NBCSP) between 1995-2002 and external data from Haukeland University hospital between 1985-1994 in their study [30] to estimate tumour growth rates and STS with the logistic growth model. Plevritis et al. used data from the Surveillance, Epidemiology and End Results (SEER) program between 1975-1981 in their study [21] to estimate tumour growth rates with the exponential growth model. The estimated tumour growth functions are shown in Figure 4 and Figure 5 and the estimated STS from Weedon-Fekjær et al.'s approach can be seen in Figure 6.

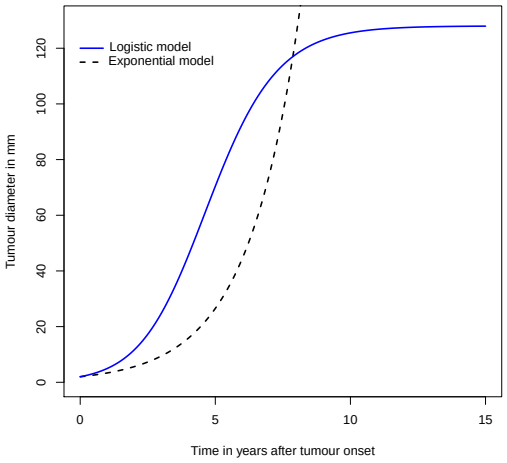


Figure 4: Estimated tumour growth functions for median tumour growth rates for two different models.

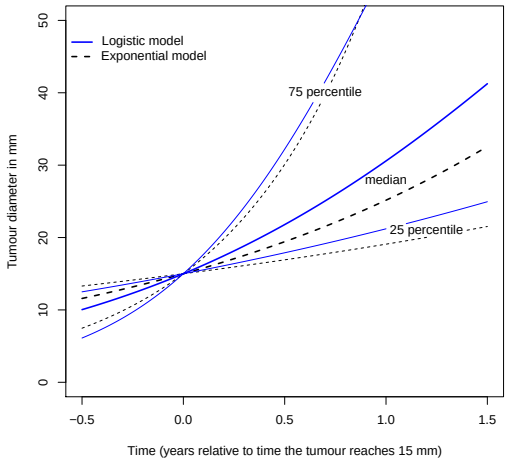


Figure 5: Estimated tumour growth functions represented as size from 15 mm for two different models.

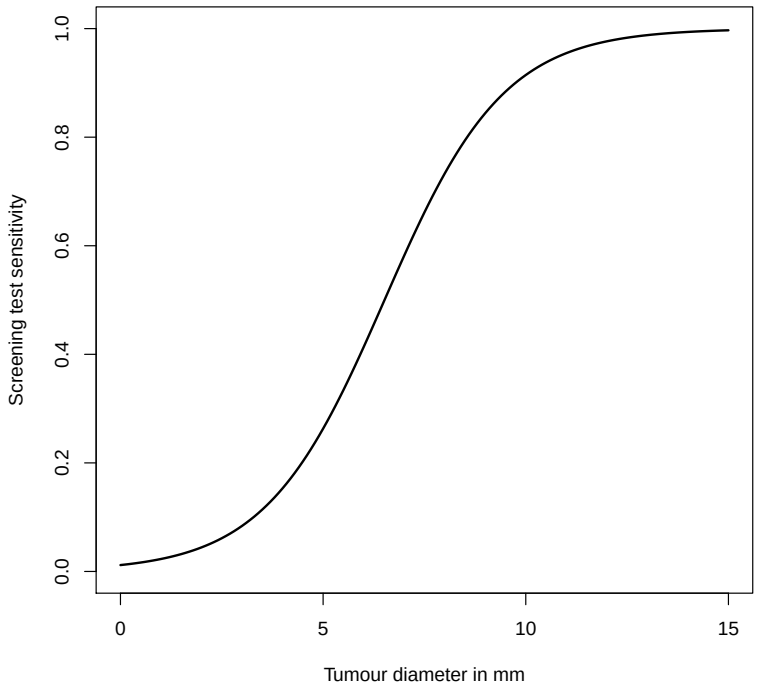


Figure 6: The estimated STS from Weedon-Fekjær's approach.

Other approaches

Weedon-Fekjær et al.'s approach above was published in 2008 [30] and subsequently, in 2010, the authors published an extension to the approach [31]. In their first article [30], Weedon-Fekjær et al. include the incidence of interval cases and the tumour size distribution for screening cases at the initial screening occasion. In their second article [31] Weedon-Fekjær et al. also include the tumour size distribution for screening cases from a population who attended previous screening occasions. The extended approach is also possible to use without information about the interval cases.

Hanin and Yakovlev [12] use a more complex model in their multivariate distribution on both age and tumour size at detection. They model the random variable for age at detection, T, as the sum of the years before the onset of the tumour and the years the tumour exist but isn't detected. In their models some survival analysis is used and in their estimation procedures both populations attending and not attending screening can be used.

3 Simulation study and results for Weedon-Fekjær's approach

In this section Weedon-Fekjær et al.'s procedure from [30] is used in a simulation study and also a correction of the procedure is proposed and tested.

3.1 Correction of the approach

In their description of the evaluation of p_{igf} (see (36)) Weedon-Fekjær et al. [30] do not account for a dependence between the variables size at clinical detection and growth rate, thus they assume that such a dependence does not exist. This seems unreasonable. A tumour which has a slow growth rate has the chance to be detected at a small size for a longer time period than what a faster growing tumour has. This being the case, a clinically detected large tumour is more likely to have a high growth rate than a clinically detected small tumour. This will be further discussed in the later simulation study.

To assume a dependence, changes need to be made in the derivation of p_{igf} and p_{igj} in the likelihood function; see (39) and (43). We use

$$p_{igf} = P(C_i|D_f \cap E_g) = P(k_1 < \kappa_i < k_2|D_f \cap E_g) = P(k_1 < \kappa_i < k_2|E_g), \quad (46)$$

i.e. the growth rate depends on the size at clinical detection but not on the time point for the detection. The distribution for the growth rate given the size at clinical detection needs to be calculated. This is mathematically difficult with the logistic-lognormal model and we therefore decided to use the tumour growth model proposed by Bartoszyński et al. [2] with an exponential tumour growth and an inverse growth rate which is gamma distributed, see (5) and (6). We also assume that the volume-dependent hazard function (9) holds. Now, let V be the random variable for volume at clinical detection and R the inverse growth rate. We therefore require $F_R(r|V = v)$, which can be derived as

$$F_R(r|V = v) = \int_0^r f_R(x|V = v)dx = \int_0^r \frac{f_{V,R}(v, x)}{f_V(v)}dx. \quad (47)$$

Since it is assumed that the volume-dependent hazard function holds, the density functions $f_{V,R}(v, r)$ and $f_V(v)$ are known from the calculations in Plevritis et al. [21], see (15) and (16). Using these density functions in the derivation of the conditional distribution function (47), it can be shown that

$$F_R(r|V = v) = \frac{\gamma(\tau_1 + 1, r(\tau_2 + \gamma(v - V_{cell})))}{\Gamma(\tau_1 + 1)}. \quad (48)$$

This is a distribution function for a gamma distributed random variable with shape parameter $\tau_1 + 1$ and inverse scale parameter $\tau_2 + \gamma(v - V_{cell})$.

The conditional distribution that we propose can be very different to the unconditional one used by Weedon et al. We show this in an example, with values in Appendix B used for the parameters τ_1, τ_2 and γ. In Figure 7 it is shown that a

larger tumour size at clinical detection is associated with a smaller inverse growth rate. The conditional density functions differs much from the unconditional one when the size at clinical detection is either very large or very small.

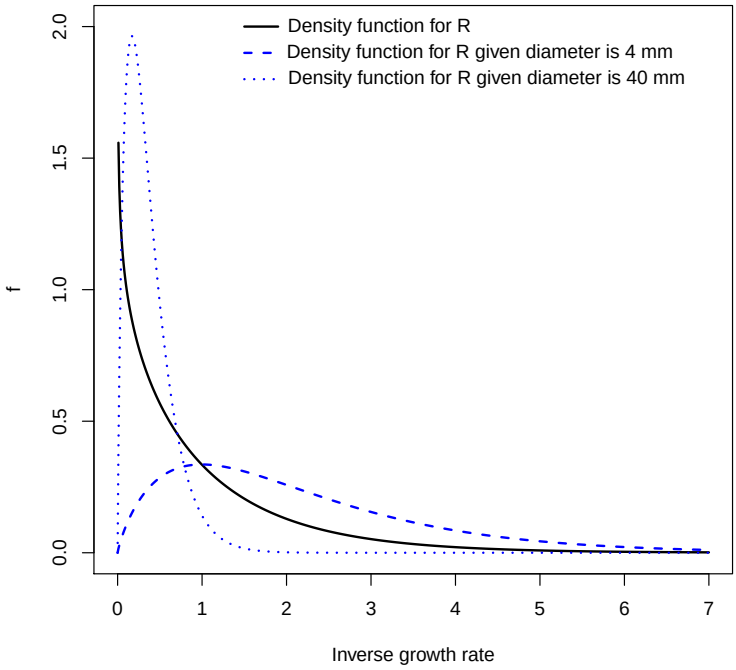


Figure 7: Differences between the conditional and unconditional density functions.

In both likelihood functions we now use

$$p_{igf} = P(C_i|D_f \cap E_g) = P(r_1 < R < r_2|D_f \cap E_g) = P(r_1 < R < r_2|E_g), \tag{49}$$

where we can calculate the boundaries r_1 and r_2 from the relation:

$$r = -\frac{t_2 - t_1}{log\left(\frac{V(t_1)}{V(t_2)}\right)}. \tag{50}$$

This function was derived in an equivalent way to (21). A calculation of the probability (49) can be made by using the distribution function in (48). The corrected procedure has one more unknown parameter than Weedon-Fekjær's original approach, namely γ from function (48). Due to identification issues we assume that this parameter is known and as in Plevritis et al. [21] we also assume that $\tau_1 = \tau_2$.

3.2 Simulation study

In this thesis no real data is used, but simulation studies have been carried out. For all unknown parameters we chose to use some values estimated from real data in the studies of Weedon-Fekjær et al. [30] and Plevritis et al. [21]. These values are presented in Appendix B.

3.2.1 The simulation procedure

The simulation of data was made in order to test Weedon-Fekjær et al.'s approach and to compare it to the corrected one. We simulated a cohort consisting of one million women in such a way that all women would get a tumour at some time point. We could have used a more sophisticated approach for simulation, such as in [11], but this would have been more complicated and our simulation was sufficient and valid for our purposes. Firstly we simulated a time at clinical detection and secondly a time at, a possible, screening detection. Whichever event occured first determined the mode and time of detection. As in study [21] we assume that

$$P(T_{det} \in [t, t+dt) | T_{det} > t) = \gamma V(t)dt + o(dt),$$

where T_{det} is the time at clinical detection, hence the hazard function of T_{det} is $\gamma V(t)$.

In theory tumours may grow very slowly. It is therefore important to have the screening occasion after quite a long time so that the size distribution of existing tumours has stabilized. Due to the same reason it is also important to have an even spread of the tumour onset time points over a long period of time. Therefore we chose to model the onset times uniformly over 200 years and all tumour diameters started from 2 mm as proposed in Plevritis et al. [21]. The screening occasion was imposed after 100 years for all women.

It is assumed that in the absence of screening all tumours will be clinically detected at some point in time. To simulate those time points we firstly generated one million independent values from the gamma distribution to determine the inverse growth rates for all tumours, secondly we generated values from the conditional distribution function $F_V(v|R = r)$, see (14), to obtain the volumes at clinical detection. The inverse function

$$F^{-1}(U) = -\frac{log(1-U)}{\gamma r} + V_{cell}, \tag{51}$$

and the following lemma is used to get random numbers from the conditional distribution.

Lemma 1. *If $U \in U[0,1]$ and F is a continuous distribution function with the inverse function F^{-1}, then the random variable $F^{-1}(U)$ has the distribution F.*

See Devroye [7] for a proof. Once the volumes at clinical detection are known, together with the time points for onset and the exponential tumour growth formula, see (5), the time points for clinical detection can be obtained. Further the STS was used to model whether a tumour is found at screening or not. Then all information about mode and time of detection was known for all tumours. Also a population of women not attending screening was simulated. The same method was used, but for these women there was no screening event.

The simulation program written in R-code [23] is attached as Appendix C. To calculate the maximum of the log likelihood the function *optim* in *R* is used.

Regarding the intervals in the likelihood function we have chosen to divide the time intervals into months. In the first likelihood 400 time intervals are used which

is approximately 33 years and in the second likelihood 24 time intervals are used. The tumour diameter intervals are (with two exceptions) 4 mm each, ranging from 2 to 128 mm. The exceptions are the last interval which only has a length of 2 mm and the first interval which consists of the non-existing tumours.

3.2.2 Presentation of the simulated cohort

In the simulated data set 532,000 women attended the screening, the other women had clinically detected tumours before the screening event. The number of screening-detected tumours was 16,327 (3 %) and the number of interval cases was 2,600. In this simulation the interval cases are women with tumours detected clinically up to two years after a negative screening.

The distribution for the tumour sizes at detection depends on the mode of detection. In Figure 8 three different estimated size distributions can be seen. The estimation has been made using the function *density* in R [23]. Keep in mind that those values are simulated and might not match real distributions. In the screening detected population the tumours tend to be smaller than in the other populations. For persons not attending screening with clinically detected tumours the sizes at detection are larger. For the interval cases the size distribution is more complex. It is a mixture of two distributions since the population is a mixture of women having their onset time before the screening and women having their onset time after the screening. This can be seen in Figure 9. The large tumours in the size distribution for the interval cases are often tumours with an onset time after the screening, tumours which are growing fast. The distribution also depends on the time since screening. Interval cases occurring during the first year after screening and during the second year after screening are compared in Figure 10.

In the likelihood function proposed by Weedon-Fekjær et al. r is assumed to be constant. This will hold if, in the absence of screening, the number of clinical cases are constant in all of the time intervals used in the likelihood. See Figure 11 to see that the assumption is correct in the simulation. In Figure 12 one can see the number of clinical cases after the time point for the screening.

Another assumption made in the likelihood function is that the number of interval cases in a time interval is Poisson distributed. This seems reasonable since the number of cases compared to the number of person years at risk are very low. In Figure 13 the number of person years used in likelihood 2 can be seen.

In this simulated cohort there is one big difference in comparison to real data. In the simulation the tumours will be clinically detected at some point in time which will not happen in reality due to mortality. In two examples we show how this difference affects some of the calculations. Let us firstly regard the size distribution for the clinical cases. This distribution depends on the time a tumour is allowed to exist. In Figure 14 three different distributions are shown. In the first distribution all tumours are detected and in the other two the tumours detected after a maximum of ten or twenty years are shown. When using real data this difference has to be considered.

Secondly, in the absence of screening, let us regard the probability for a tumour to be of size i given that the tumour is of size g, f months later. This probability

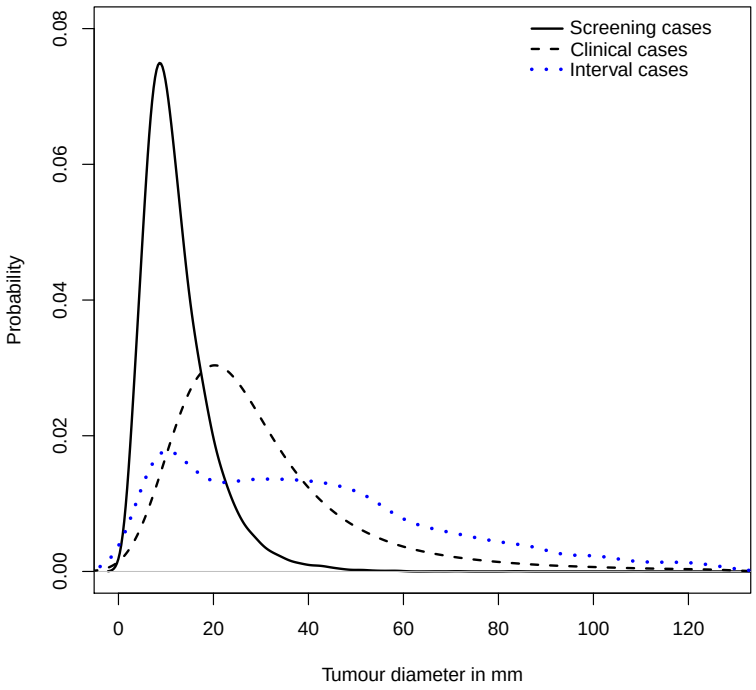


Figure 8: Size distributions for the different cases.

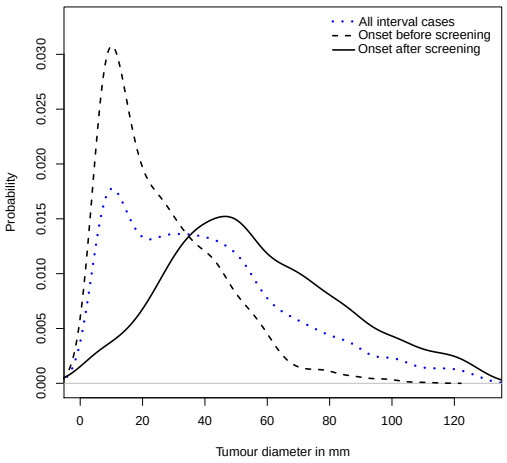


Figure 9: Size distributions for the different interval cases.

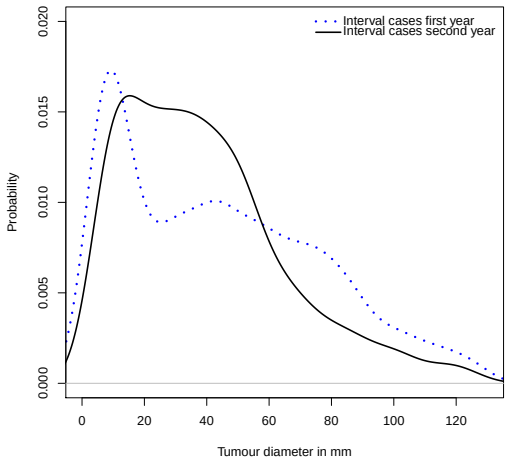


Figure 10: Size distributions for the interval cases at different time intervals.

might be different in the simulated data and in the real data. In the simulated data all tumours will, in time, be of size g if they are not found clinically before then. This is not true for real data. For example, a woman can get a tumour when she is old that might not grow for a long time period and might not ever reach the size

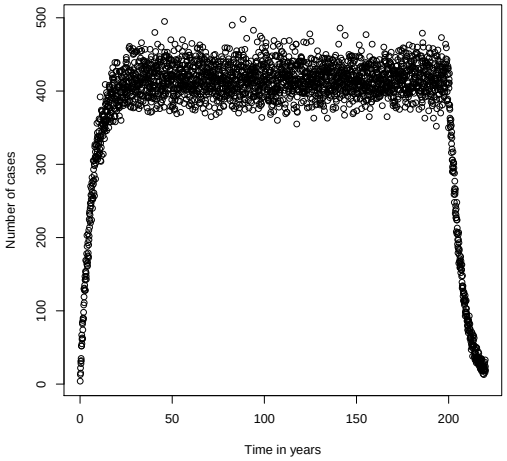


Figure 11: Number of cases in the whole simulation in absence of screening.

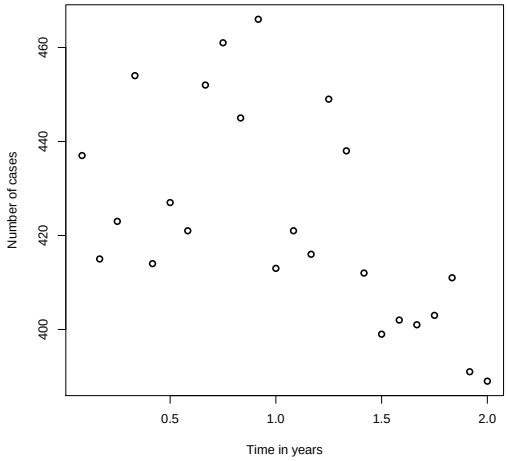


Figure 12: Number of cases after 100 years in absence of screening.

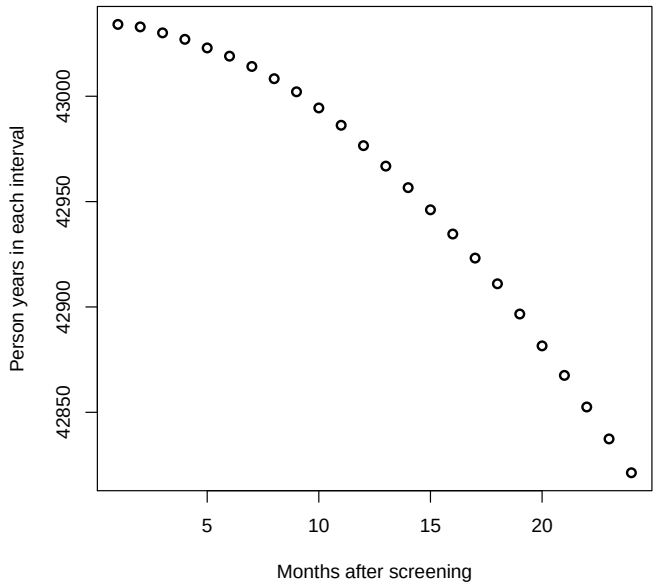


Figure 13: The number of person years at risk after the screening occasion.

g. Let us now assume that $f = 24$ months, g is the interval for sizes between 31-32 mm and that we wish to calculate the probability to be of size i. We simulated data under two different conditions and compared the simulated values to the expected. In the simulations 1000 tumours were followed for 5 years (Figure 15) and 167 years (Figure 16). Once the tumours reached 31 mm we extracted their sizes two years earlier. When the tumours were followed for 5 years, only 35 % of them reached 31 mm, while all the tumours followed for 167 years reached that size.

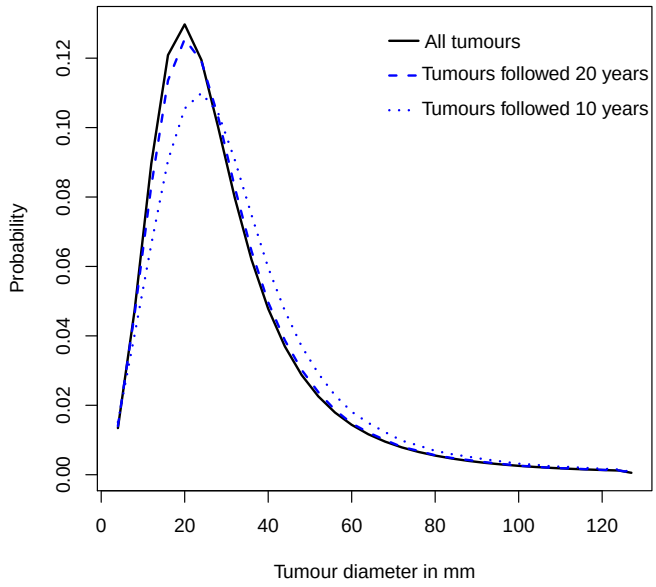


Figure 14: Size distributions for clinical cases for different follow-up times.

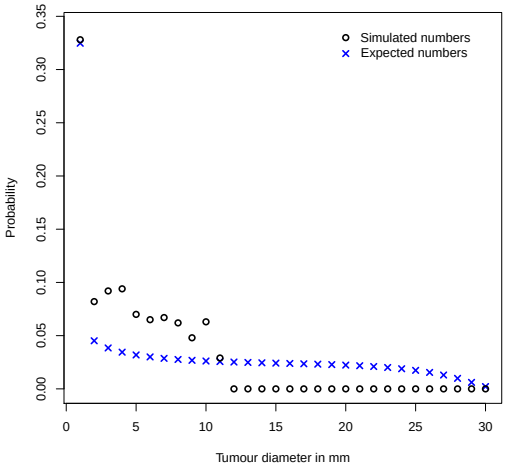


Figure 15: Longitudinal distribution of tumour size 2 years ago, given current size of 31-32 mm. Here the tumours are followed for five years.

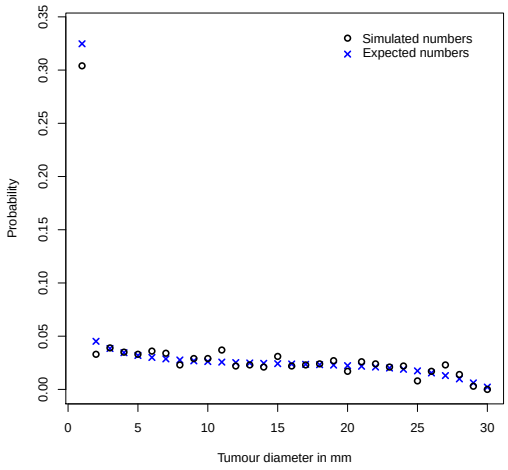


Figure 16: Longitudinal distribution of tumour size 2 years ago, given current size of 31-32 mm.

3.3 Comparison of the corrected and original models

In Table 1 we present the derived likelihood estimates from both Weedon-Fekjær's approach and our corrected approach. These values are compared to the correct parameter values, used in the simulation.

Although our corrected approach performs better with this data set it is impor-

Table 1: Estimated values from the corrected and original models.

Parameter	Correct value	W-F's approach	Corrected approach
τ_1, τ_2	0.85	0.93	0.82
β_1	-4.43	-15.55	-4.14
β_2	0.68	3.60	0.60

tant to keep in mind that the parameter γ is assumed to be known in our corrected approach. Due to lack of time only one estimation procedure has been made for each approach, so unfortunately there are no measures of certainty available. Bootstrap confidence intervals could have been calculated as in the study made by Weedon-Fekjær et al. [30]. Another possibility to see how certain the estimates are would be to make more data simulations.

Let us now look at the two distributions used in the likelihood function: the size at screening detection and the number of interval cases. In Figure 17 the observed numbers from the simulation together with the expected and estimated numbers from both approaches are plotted. When comparing the curves for the expected numbers it is obvious that the corrected approach works better. In Figure 18 the number of interval cases are shown. If one regards the expected numbers one might think that Weedon-Fekjær's approach has a better fit for the last time intervals in the plot. But if Figure 12 is considered once more, one can see that due to random variation the number of clinical cases, in the absence of screening, are low in these time intervals.

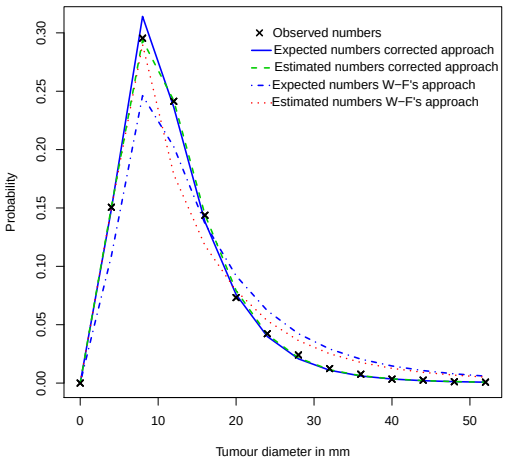


Figure 17: Observed (simulated), expected and estimated numbers of screening detected tumours in different size intervals, for W-F's and the corrected approach.

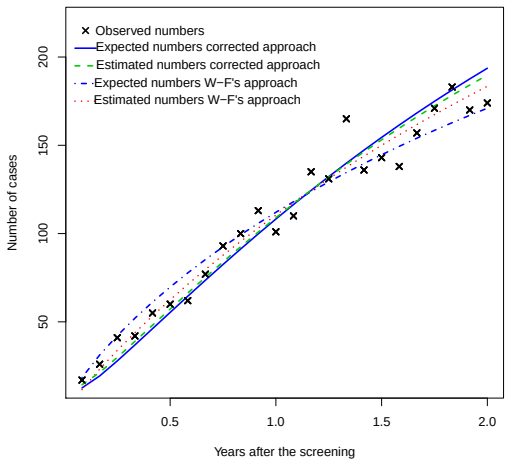


Figure 18: Observed (simulated), expected and estimated numbers of interval cases in different time intervals, for W-F's and the corrected approach.

It is also interesting to see how well the estimated growth function and STS fit

with the actual functions used in the simulation. See Figures 19 and 20 for results. It is clear that the estimates from Weedon-Fekjær's approach are inadequate.

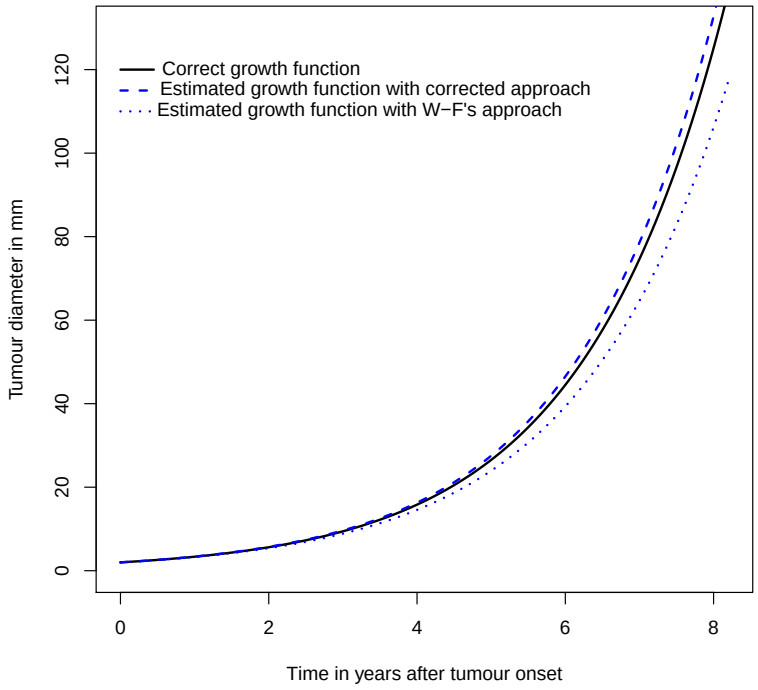


Figure 19: Correct and estimated tumour growth functions for median tumour growth rates, for W-F's and the corrected approach.

4 An extension of Weedon-Fekjær's model

In their approach to estimate tumour growth Weedon-Fekjær et al. [30] rely on an external data set from 1985-1994 of women not involved in a screening program, while their primary data set (in the presence of a screening program) was from 1995-2002. This might lead to time bias. Nowadays screening programs are widely spread which makes it hard to use up to date external data sets. For the exponential growth model, we propose that no external data shall be used and instead q_g in the formulas (39) and (43) is calculated by the distribution function for tumour volume at clinical detection, derived by Plevritis et al. [21]:

$$F_V(v) = 1 - \left(\frac{\tau_2}{\tau_2 + \gamma(v - V_{cell})} \right)^{\tau_1}, \; v \geq V_{cell}.$$

Let g be the size interval with diameters between g_a and g_b mm, corresponding to the volumes v_a and v_b. Then

$$q_g = P(v_a < V \leq v_b) = F_V(v_b) - F_V(v_a). \tag{52}$$

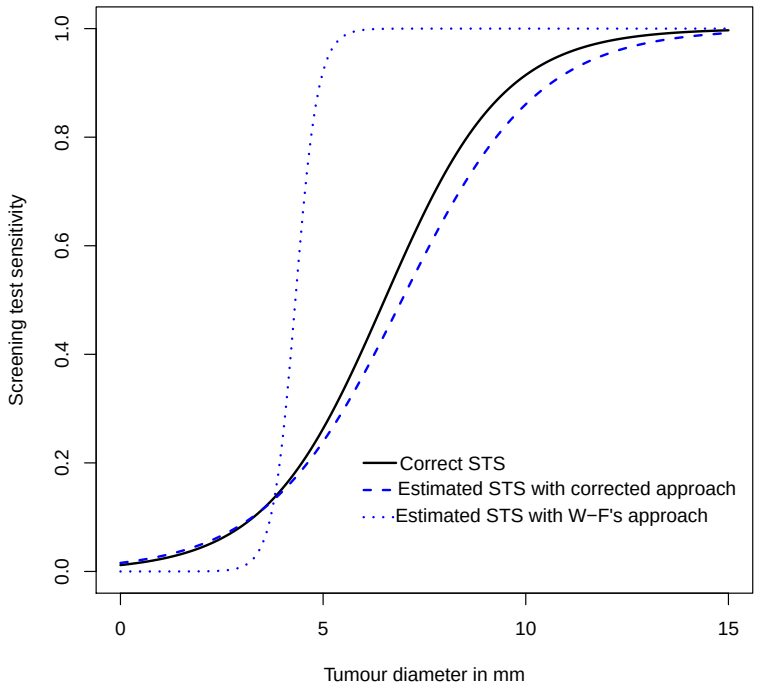


Figure 20: Correct and estimated STS, for W-F's and the corrected approach.

In the external data all tumours are not captured mainly due to slow growth rates but this will not be a problem in our approach. Our extended approach and the corrected Weedon-Fekjær's approach will be compared under the assumption that the exponential growth model holds. However the assumption might not be correct for real data and therefore the approaches will be compared for simulated data under both the exponential and the logistic growth model.

4.1 Results for the exponential growth model

The estimation procedures with and without external data, using the exponential growth model, are compared (see Table 2 for the estimates). Once more let's look

Table 2: Estimated values for the corrected approach under the exponential growth model in the absence of external data.

Parameter	Correct value	With external data	Without external data
τ_1, τ_2	0.85	0.82	0.71
β_1	-4.43	-4.14	-4.13
β_2	0.68	0.60	0.59

at the size distribution for the screening detected tumours and the distribution for the number of interval cases; see Figures 21 and 22. If we regard the estimated

number of interval cases, it might be a bit unstable without the external data, but the differences do not appear to be big.

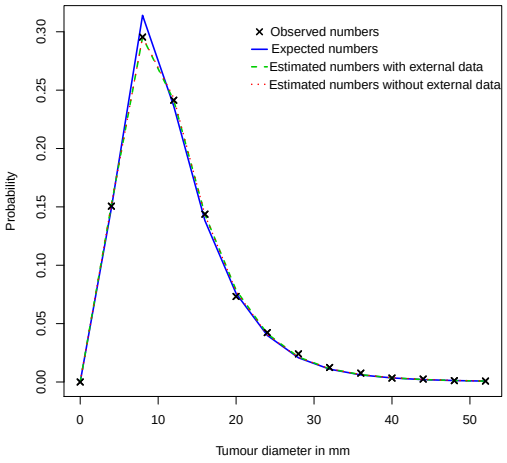


Figure 21: Observed (simulated), expected and estimated numbers of screening detected tumours in different size intervals under the exponential growth model in the absence of external data.

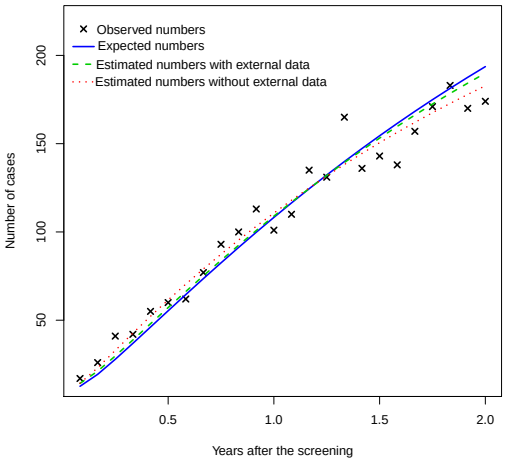


Figure 22: Observed (simulated), expected and estimated numbers of interval cases in different time intervals under the exponential growth model in the absence of external data.

It is interesting to study how the growth function and STS are estimated with and without the external data; see Figures 23 and 24. For STS the results are similar. For the growth function, the result is a bit better with the external data. From this small study it seems like both approaches are satisfactory.

When the external data is not available it is now possible to estimate it using the distribution function for tumour volumes at clinical detection (17) with the parameter estimates from Table 2. To find the actual size distribution the parameter values from the simulation (see Appendix B) are used. See Figure 25 for the result.

4.2 Results for the logistic growth model

We next examined what happens if the assumption about an exponential growth function with a gamma distributed inverse growth rate is incorrect. We carried out a new simulation, where the tumours follow the logistic growth function with a lognormal growth rate, as assumed by Weedon-Fekjær et al. [30]. Though the same likelihood function as before is used. The parameter values used in the simulation, α_1, α_2, β_1 and β_2, are shown in Appendix B. The simulated cohort consists of 100,000 women and the onset time is randomly chosen over 120 years. The screening occasion appears after 60 years. 54,354 women attended the screening and of these, 2,153 women (4 %) were screening cases. The number of interval cases was 491.

In Table 3 the estimated parameter values are shown. It is important to keep in mind that the tumour growth functions and the distributions for the growth rates

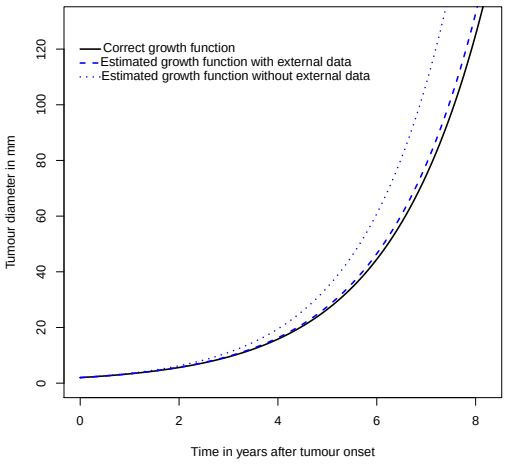


Figure 23: Correct and estimated tumour growth functions for median tumour growth rates under the exponential growth model in the absence of external data.

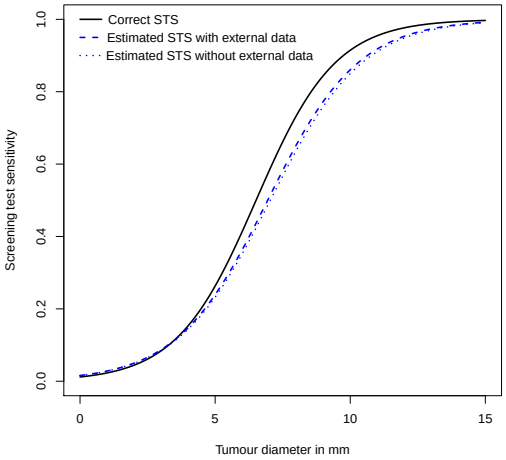


Figure 24: Correct and estimated STS under the exponential growth model in the absence of external data.

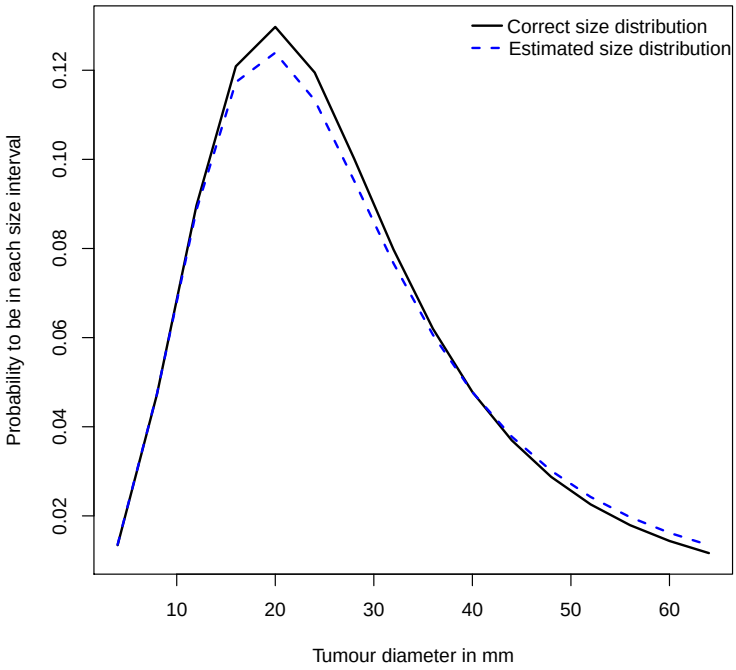


Figure 25: Correct and estimated size distribution for clinically detected tumours under the exponential growth model in the absence of external data.

are different in the simulation and the likelihood function; there is therefore no correct value for the parameter τ_1. For the logistic growth model we have not derived

Table 3: Estimated values for the corrected approach under the logistic growth model in the absence of external data.

Parameter	Correct value	With external data	Without external data
τ_1, τ_2	–	1.27	1.28
β_1	−4.43	-4.33	-4.26
β_2	0.68	0.64	0.59

explicit formulas for the distribution of tumour sizes at screening detection. Therefore 'expected' numbers are not available for the distributions of size at screening detection, nor for the number of interval cases. In Figure 26 and Figure 27 the observed (simulated) and estimated (fitted) numbers are however shown.

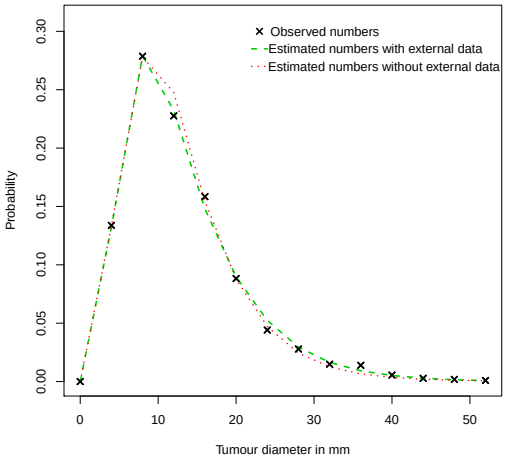


Figure 26: Observed (simulated) and estimated numbers of screening detected tumours in different size intervals under the logistic growth model in the absence of external data.

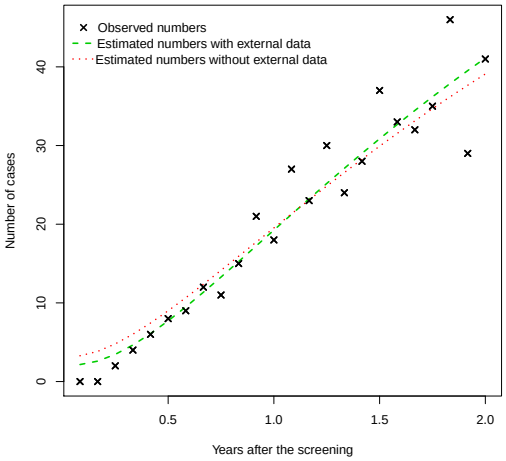


Figure 27: Observed (simulated) and estimated numbers of interval cases in different time intervals under the logistic growth model in the absence of external data.

It is not relevant to compare the estimated and correct growth models since the assumption of an exponential growth function made in the fitted model was wrong. It is more interesting to examine how well the STS function was estimated; see Figure 28. Both with and without external data the estimated function looks satisfactory.

5 Discussion

In this thesis we have explored some continuous models used to estimate tumour growth and STS. The decision to concentrate on continuous models rather than

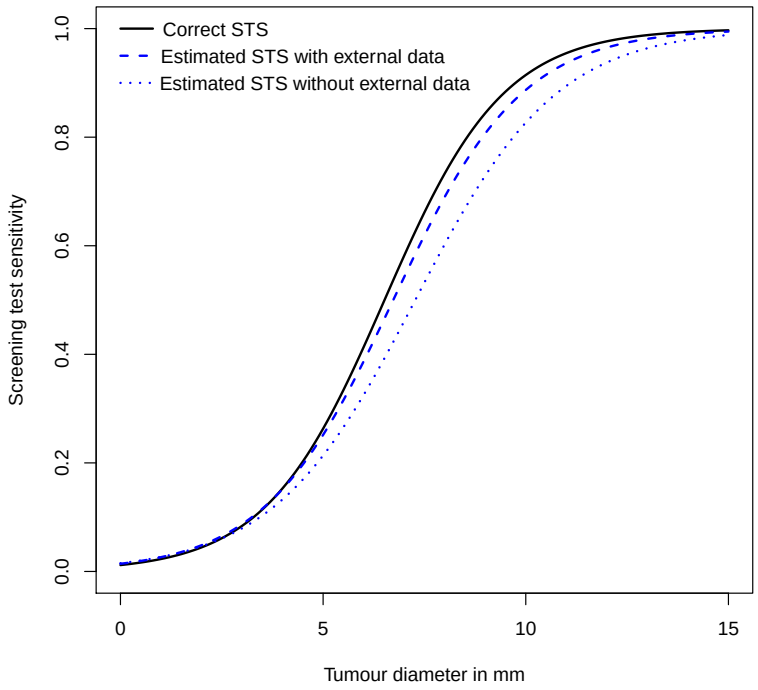


Figure 28: Correct and estimated STS under the logistic growth model in the absence of external data.

on Markov models is that the former are more appealing in many respects. One important reason for the choice is that the individual variation is not captured in the Markov models at all. One can also question whether there are any biological explanations for the choice of states. In reality there does not exist a specific time point when a tumour goes from "no disease" to a "preclinical but detectable tumour" state. The assumed Markov property can also be doubted, since it is reasonable to believe that the probability for a tumour to be clinical depends on whether the tumour has been in the preclinical state for a long or short time. Regarding the model proposed by Prevost et al. [22], called model 2, some drawbacks exist. Firstly, the intensity, λ_1, to go from "no disease" to "preclinical but detectable tumour" is estimated as the incidence of breast cancers in the absence of screening. This can not be entirely correct since there exist tumours which are never found. Secondly, the estimates for STS and MST have been shown by Weedon et al. [31] to have a strong negative correlation. It is also not reasonable to estimate STS as a constant, equal for all tumours in the preclinical state. Furthermore one should remember that the states in the Markov models depend on the type of screening method and that the STS, in both types of models, also depends on the screening method. To summarize, the continuous models seem to be more reasonable and also allow for variability.

In this thesis we have focused on tumour growth models in the presence of screening. It is arguably easier to estimate tumour growth from data in the absence of

screening. The STS is however very important to know in order to evaluate screening programs and it can only be estimated from data in the presence of screening. We have focused in particular on the methods described by Weedon-Fekjær et al. in [30]. The approach suffers from incorrectly assuming that the growth rate is conditionally independent of the tumour size at clinical detection. In this thesis large samples are used, it would be interesting to use different population sizes in the simulation to see how the size affects the variability of the estimates. We have also shown that the size distribution for clinical tumours depends on whether the slow growing tumours are found or not. When using the external data (from reality) this fact could possibly affect the parameter estimates. If we instead use the assumed size distribution for the clinical tumours, as in our extension of Weedon's model, this should not be a problem. Another drawback with relying on external data is that it is older data, before the introduction of screening, which might not be accurate and there may have been changes, in time in, for example, awareness of breast cancer symptoms.

In future studies we will investigate the importance of the choice of value of γ for making inference on other model parameters and explore the use of sensitivity analysis.

In this paper we have not described in any detail how the number of time intervals in the first likelihood function was chosen. Different time lengths were however tried and by using a sufficient time period the parameter estimates were found to converge.

There are several possible future extensions for the methods evaluated in this thesis. As mentioned earlier mammographic density, HRT and genetic factors affect the tumour growth and STS [4], [13], [17]. In the function for STS it would be very interesting to add a covariate for mammographic density. More covariates could also be added for the tumour growth rate, for example mammographic density, HRT and genetic factors. To incorporate these factors the method based on external data could not be used. New likelihoods could also be derived, for example a likelihood function describing the size distribution for the interval cases. It would also be interesting to explore how a tumour growth model described by a geometric Brownian motion could be used.

In this thesis we have concentrated on understanding one particular approach for tumour growth modeling and correcting it, as well as proposing an extension. Time limitations restricted us from carrying out more extensive simulations to more thoroughly evaluate the properties of the models and for studying how they would perform on real data. These aspects will be explored in future work.

References

[1] Alm S A, Britton T. *Stokastik: Sannolikhetsteori och statistikteori med tillämpningar.* Liber, Stockholm 2008; ISBN 9147053518.

[2] Bartoszyński R, Edler L, Hanin L, Kopp-Schneider A, Pavlova L, Tsodikov A, Zorin A, Yakovlev A Y. Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis. *Mathematical Biosciences* 2001; **171**: 113-142.

[3] Berry D A, Cronin K A, Plevritis S K, Fryback D G, Clarke L, Zelen M, Mandelblatt J S, Yakovlev A Y, Habbema J D F, Feuer E J. Effect of screening and adjuvant therapy on mortality from breast cancer. *The New England Journal of Medicine* 2005; **353**: 1784-1792.

[4] Biesheuvel C, Czene K, Orgéas C C, Hall P. The role of mammography screening attendance and detection mode in predicting breast cancer survival-is there added prognostic value? *Cancer Epidemiology* 2011; **35(6)**: 545-550.

[5] Chia Y L, Salzman P, Plevritis S K, Glynn P W. Simulation-based parameter estimation for complex models: a breast cancer natural history modelling illustration. *Statistical Methods in Medical Research* 2004; **13**: 507-524.

[6] Day N E, Walter S D. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. *Biometrics* 1984; **40**: 1-14.

[7] Devroye L. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York 1986; ISBN 0387963057.

[8] Duffy S W, Chen H H, Tabar L, Day N E. 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. *Statistics in Medicine* 1995; **14**: 1531-1543.

[9] Duffy S W, Chen H H, Prevost T C, Tabar L. Markov chain models of breast tumour progression and its arrest by screening. In Duffy S W, Hill C, Estève J, eds. *Quantitative Methods for the Evaluation of Cancer Screening.* Arnold, London 2001; ISBN 0340741252.

[10] Fitzmaurice G M, Laird N M, Ware J H. *Applied Longitudinal Analysis.* John Wiley & Sons, New Jersey 2004; ISBN 0471214876.

[11] Forastero C, Zamora L I, Guirado D, Lallena A M. A Monte Carlo tool to simulate breast cancer screening programmes. *Physics in Medicine and Biology* 2010; **55**: 5213-5229.

[12] Hanin L G, Yakovlev A Y. Multivariate distributions of clinical covariates at the time of cancer detection. *Statistical Methods in Medical Research* 2004; **13**: 457-489.

[13] Hankinson S, Tamimi R, Hunter D. Breast cancer. In Adami H-O, Hunter D, Trichopoulos D, eds. *A Textbook of Cancer Epidemiology,* 2nd edition, OUP USA, New York, 2008; ISBN 9780195311174.

[14] Hart D, Shochat E, Agur Z. The growth law of primary breast cancer as inferred from mammography screening trials data. *British Journal of Cancer* 1998; **78**: 382-387.

[15] Karma studien [www.karmastudien.se] (accessed 2012 February 6)

[16] Klein J P, Moeschberger M L. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, New York 1997; ISBN 0387948295.

[17] Li J, Szekely L, Eriksson L, Heddson B, Sundbom A, Czene K, Hall P, Humphreys K. High-throughput mammographic density measurement: A tool for risk predicition of breast cancer. Submitted.

[18] Li J, Seibold P, Chang-Claude J, Flesh-Janys D, Liu J, Czene K, Humphreys K, Hall P. Coffee consumption modifies risk of estrogen-receptor negative breast cancer. *Breast Cancer Research* 2011; **13**: R49.

[19] Mandelson M T, Oestreicher N, Porter P L, White D, Finder C A, Taplin S H, White E. Breast density as a predictor of mammographic detection: Comparison of interval- and screen-detected cancers. *Journal of the National Cancer Institute* 2000; **92(13)**: 1081-1087.

[20] van Oortmarssen G J, Habbema J D F, van der Maas P J, de Koning H J, Collette H J A, Verbeek A L M, Geerts A T, Lubbe K T N. A model for breast cancer screening. *Cancer* 1990; **66**: 1601-1612.

[21] Plevritis S K, Salzman P, Sigal B M, Glynn P W. A natural history model of stage progression applied to breast cancer. *Statistics in Medicine* 2007; **26**: 581-595.

[22] Prevost T C, Launoy G, Duffy S W, Chen H H. Estimating sensitivity and sojourn time in screening for colorectal cancer - A comparison of statistical approaches. *American Journal of Epidemiology* 1998; **148(6)**: 609-619.

[23] R [http://www.r-project.org]

[24] Santen R J. Breast Pain/Lumps in Women [Internet]. Version 4. Knol. 2008 Jun 10. Available at: URL: http://knol.google.com/k/richard-j-santen/breast-pain-lumps-in-women/C7SqxBoM/a68vpw (accessed 2011 December 16).

[25] Socialstyrelsen, Cancerfonden. *Populärvetenskapliga fakta om cancer - Cancer i siffror 2009.* Socialstyrelsen, Cancerfonden 2009; ISBN 978-91-89446-36-6

[26] Spratt J A, von Fournier D, Spratt J S, Weber E E. Decelerating growth and human breast cancer. *Cancer* 1993; **71**: 2013-2019.

[27] Spratt J A, von Fournier D, Spratt J S, Weber E E. Mammographic assessment of human breast cancer growth and duration. *Cancer* 1993; **71**: 2020-2026.

[28] Stirzaker D. *Stochastic Processes & Models.* Oxford University Press, Oxford 2005; ISBN 9780198568148.

[29] Walter S D, Day N E. Estimation of the duration of a pre-clinical disease state using screening data. *American Journal of Epidemiology* 1983; **118(6)**: 865-886.

[30] Weedon-Fekjær H, Lindqvist B, Vatten L J, Aalen O O, Tretli S. Breast cancer tumor growth estimated through mammography screening data. *Breast Cancer Research* 2008; **10**: R41.

[31] Weedon-Fekjær H, Aalen O O, Tretli S. Estimating screening test sensitivity and tumour progression using tumour size and time since previous screening. *Statistical Methods in Medical Research* 2010; **00**: 1-21.

Appendix A

Here are the calculations of $\mathbf{P}$ by the *forward equations*, $\mathbf{P}' = \mathbf{PQ}$, derived from Theorem 1. The nine differential equations are represented by

$$p'_{j1}(t) = -\lambda_1 p_{j1}(t),\ j \in S$$

$$p'_{j2}(t) = \lambda_1 p_{j1}(t) - \lambda_2 p_{j2}(t),\ j \in S$$

$$p'_{j3}(t) = \lambda_2 p_{j2}(t),\ j \in S$$

To solve these equations it is also used that

$$p_{jj}(0) = 1,\ j \in S,$$

$$p_{jk}(0) = 0,\ j \neq k,\ j, k \in S,$$

$$\sum_k p_{jk}(t) = 1,\ j, k \in S.$$

Solutions:

$$p_{11}(t) = e^{-\lambda_1 t}$$

$$p_{12}(t) = \frac{\lambda_1(e^{-\lambda_2 t} - e^{-\lambda_1 t})}{(\lambda_1 - \lambda_2)}$$

$$p_{13}(t) = 1 + \frac{\lambda_2 e^{-\lambda_1 t}}{(\lambda_1 - \lambda_2)} + \frac{-\lambda_1 e^{-\lambda_2 t}}{(\lambda_1 - \lambda_2)}$$

$$p_{22}(t) = e^{-\lambda_2 t}$$

$$p_{23}(t) = 1 - e^{-\lambda_2 t}$$

$$p_{21}(t) = p_{31}(t) = p_{32}(t) = 0$$

$$p_{32}(t) = 1$$

Appendix B

The parameter values in the simulations are presented below. The value of gamma is only an approximation, so that it won't be to small.

Table B-1: Parameter values in the simulations.

Parameter	Value	Reference
γ	e^{-9}	[21]
τ_1, τ_2	$e^{-0.165}$	[21]
α_1	1.07	[30]
α_2	1.31	[30]
β_1	-4.43	[30]
β_2	0.68	[30]

Appendix C

```
###screening simulation of tumour growth with STS
###under the exponential model

#number of women
no<-1000000

tau<-exp(-0.165)
gamma<-exp(-9)
beta1<--4.43
beta2<-0.68

##individual inverse growth rate
rscreen<-rgamma(no,shape=tau,scale=1/tau)

#starting time point of tumour
onsettime<-200*runif(no,0,1)

c0<-(2^3)*pi/6

vdetscreen<-matrix(NA,no,1)
vdetintervalscreen<-matrix(NA,no,1)
tdetscreen<-matrix(NA,no,1)
tdetscreentmp<-matrix(NA,no,1)
rtdetscreentmp<-matrix(NA,no,1)
vrtdetscreentmp<-matrix(NA,no,1)
tdettmp<-matrix(NA,no,1)
vdettmp<-matrix(NA,no,1)
tintervaldet<-matrix(NA,no,1)
intervaltimeatrisk<-matrix(NA,no,1)

for(a in 1:no)
{
##time point for possible screening
rtdetscreentmp[a]<-100

if(rtdetscreentmp[a]<onsettime[a]){vrtdetscreentmp[a]<-0
}else{timediff<-rtdetscreentmp[a]-onsettime[a]
vrtdetscreentmp[a]<-c0*exp((timediff)/rscreen[a])}

d<-(6*vrtdetscreentmp[a]/pi)^(1/3)

if(d==0){STS<-0
}else if(beta1+beta2*d>709){STS<-1}else{
STS<-(exp(beta1 + beta2*d))/(1+exp(beta1 + beta2*d))}
```

```
tmp<-runif(1,0,1)

if(STS<tmp) ##that is not detected (to small tumour)
{tdetscreentmp[a]<-NA
}else{tdetscreentmp[a]<-rtdetscreentmp[a]}

##when it will be clinically detected
q<-runif(1,0,1)
vdettmp[a]<-(-log(1-q)/(gamma*rscreen[a])) +c0
tdettmp[a]<-rscreen[a]*log(vdettmp[a]/c0)+onsettime[a]

##tumours not found clinically
if(is.na(tdettmp[a])==TRUE){tdetscreen[a]<-tdetscreentmp[a]
vdetscreen[a]<-vrtdetscreentmp[a]
##tumours not found at screening due to STS
}else if(is.na(tdetscreentmp[a])==TRUE){tdetscreen[a]<-NA
vdetscreen[a]<-NA
##tumours found at screening
}else if(tdetscreentmp[a]<=tdettmp[a]){tdetscreen[a]<-tdetscreentmp[a]
vdetscreen[a]<-vrtdetscreentmp[a]
##tumours found clinically before the screening
}else{tdetscreen[a]<-NA
vdetscreen[a]<-NA}

#####tumours found at screening are not interval cases
if(is.na(tdetscreen[a])==FALSE){
tintervaldet[a]<-NA
intervaltimeatrisk[a]<-NA
vdetintervalscreen[a]<-NA
###interval detected after the failured screening
}else if(tdettmp[a]>rtdetscreentmp[a]){
tintervaldet[a]<-tdettmp[a]-rtdetscreentmp[a]
intervaltimeatrisk[a]<-tdettmp[a]-rtdetscreentmp[a]
vdetintervalscreen[a]<-vdettmp[a]
}else{tintervaldet[a]<-NA}
}

length(na.omit(tdetscreen))
length(na.omit(tintervaldet))
length(na.omit(intervaltimeatrisk))

##diameter at detection
ddetscreen<-matrix(NA,no,1)

##diameter at detection interval
```

```
ddetintervalscreen<-matrix(NA,no,1)
ddet<-matrix(NA,no,1)

for(a in 1:no)
{
ddetscreen[a]<-(6*vdetscreen[a]/pi)^(1/3)
ddetintervalscreen[a]<-(6*vdetintervalscreen[a]/pi)^(1/3)
ddet[a]<-(6*vdettmp[a]/pi)^(1/3)
}
```