Methodology and Computing in Applied Probability (2021) 23:241–255
https://doi.org/10.1007/s11009-020-09771-4

Gamma Process-Based Models for Disease Progression

Ayman Hijazy[1,2] · **András Zempléni**[1,2]

Received: 15 March 2019 / Revised: 18 October 2019 / Accepted: 1 January 2020 /
Published online: 29 January 2020

Abstract
Classic chronic diseases progression models are built by gauging the movement from the disease free state, to the preclinical (asymptomatic) one, in which the disease is there but has not manifested itself through clinical symptoms, after spending an amount of time the case then progresses to the symptomatic state. The progression is modelled by assuming that the time spent in the disease free and the asymptomatic states are random variables following specified distributions. Estimating the parameters of these random variables leads to better planning of screening programs as well as allowing the correction of the lead time bias (apparent increase in survival observed purely due to early detection). However, as classical approaches have shown to be sensitive to the chosen distributions and the underlying assumptions, we propose a new approach in which we model disease progression as a gamma degradation process with random starting point (onset). We derive the probabilities of cases getting detected by screens and minimize the distance between observed and calculated distributions to get estimates of the parameters of the gamma process, screening sensitivity, sojourn time and lead time. We investigate the properties of the proposed model by simulations.

Keywords Disease progression models · Gamma process · Sojourn time · Lead time bias · Sensitivity

Mathematics Subject Classification (2010) 60K10 · 62P10 · 62B10

1 Introduction

The natural progression model proposed by Zelen and Feinleib (1969) is a three state model. Progression starts from being disease free (S_f), then one moves into the preclinical state

✉ Ayman Hijazy
aymanh@cs.elte.hu

András Zempléni
zempleni@caesar.elte.hu

[1] Department of Probability Theory and Statistics, Eötvös Loránd University, Budapest, Hungary

[2] Faculty of Informatics, University of Debrecen, Debrecen, Hungary

(S_p), in which one has the disease but it has not yet manifested itself through clinical symptoms. The progression ends from our point of view when symptoms appear and one reaches the clinical state (S_c) (Zelen and Feinleib 1969).

Screening programs are organized aiming for early detection of diseases in hopes of improving survival. However, early detection automatically means that the survival of cases that were diagnosed by screens is longer than the survival of cases that were diagnosed by clinical symptoms (S_c). In Fig. 1 one can see case A (in black) of which the disease was detected early, and case B (in grey) which was detected after showing symptoms. Although both cases become onset and are deceased at the same time, case A will appear to have survived longer simply because its survival is recorded from the first date of diagnosis. This apparent increase in survival which is observed purely due to early detection is called lead time bias (Gordis 2008).

The estimation of the sojourn time, which is the amount of time spent in the preclinical state (S_p) allows correcting the lead time bias as well as the evaluation of existing programs and optimizing future ones. Sojourn time thus governs the movement between S_p and S_c. For gauging the movement between S_f and S_p, we define the preclinical intensity as the probability of moving from S_f into S_p during $(t + dt)$.

The classical approach for modelling the process (Wu et al. 2005) is by assuming that both sojourn time and the preclinical intensity are random variables of which the parameters are to be estimated. The sensitivity is given a functional form (e.g. logistic form) with some parameters. Thus one can determine the probability of a case to be detected by a screen or during an interval between screens by symptoms in terms of the parameters. Subsequently a likelihood function can be formulated and maximized in order to get parameter estimates, see e.g. Hijazy and Zempléni (2020).

However, there is a lot of discrepancy between results in the literature, for instance an entry-exit model into and out of S_p is used by Duffy et al. (1995), who applied their model to the Swedish two-county study of breast cancer, the resulting estimate for the mean sojourn time is 2.3 years. On the other hand, Weedon-Fekjaer et al. (2005) obtained their estimates through weighted non-linear least-square regression using a three step Markov chain model, applying their method to the Norwegian Breast Cancer Screening Program (NBCSP), they estimated the mean sojourn time to be 6.1 years for those aged between [50,59] and 7.9 years for those between [60,69]. These differences are most likely caused by the underlying assumptions while building the model, as these can have a crucial effect on the estimates. To deal with this discrepancy, more information has to be incorporated into such models. Namely, we chose to include a measure of sickness or degradation. The measure of sickness on diagnosis should provide some information about the preclinical time, e.g. in breast

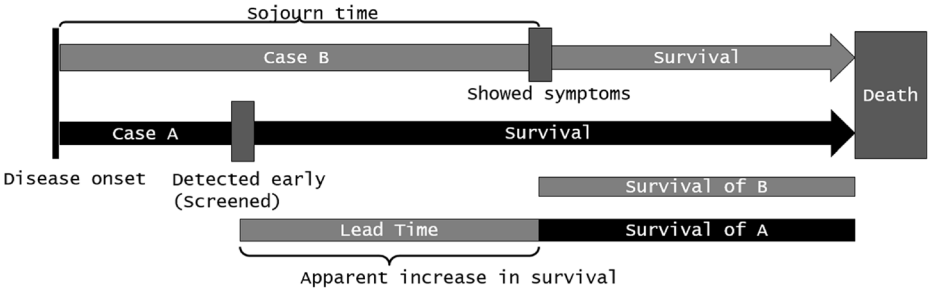

Fig. 1 Apparent increase in survival due to early detection

cancer, a case detected with a large tumor is likely to have been preclinical for a long time. In other words, integrating the tumor size at diagnosis into the model forms an additional constraint and deals with the discrepancies between the results. Nonetheless, in order to incorporate tumor size into the model, one has to specify the rate of growth.

Very early results by Collins et al. (1956) and Schwartz (1961) proposed simple linear exponential growth, introducing the notion of doubling time, that is the time needed for a tumor to double its size. Using an exponential growth is still by far the most popular in the literature. However, Laird (1964) found that exponential growth of tumors is realistic only in cases where tumors were observed for relatively brief periods. Moreover, when tumor growth was followed for a sufficiently extensive period of time, the results showed that nearly all tumors grow more and more slowly as the tumor got larger, opposing the constant specific growth rate that would be expected for simple exponential growth. Norton (1988) suggested a Gompertz function in which successive doublings occur at increasingly longer intervals. Figure 2 shows the decelerating rate of Gompertzian growth beside an exponentially growing tumor of constant rate.

Speer et al. (1984) proposed modelling by a more generalized approach using the Gompertzian kinetics. They included the noisiness in the growth process, i.e. they assumed that from time to time, in a random fashion, a spontaneous change occurs in the growth rate.

Although it seems that there is no general consensus about the shape of tumor growth, it looks like a logistic or a Gompertz growth shape is more likely than an exponential one. Besides, it seems that the growth process has some randomness and the growth rate is not constant. These properties led us to use the gamma process, since the expected value of the gamma process is given by the product of the shape and the scale. We used the Gompertz function as the shape of a gamma process as it was proposed by Norton (1988). As a result, in our model, tumors grow in gamma distributed increments with an expected Gompertz growth (see Fig. 3).

Since its introduction by Abdel-Hameed (1975) the gamma process has been extensively used in the literature to describe the stochastic and monotone degradation accumulating over time. The gamma process is very tractable and has nice properties, besides, the process is flexible in the sense that multiple functional forms of the shape can be easily adapted. This enabled us to establish the distributions of detected tumor sizes on screen, which in turn allows the use of classical estimation methods.

The paper is organized as follows: in Section 2 we lay the setup of the model, derive the distributions of the sojourn time, lead time and the distributions on screens. In Section 3 we apply the model on simulated data and show the results. In Section 4 we state some concluding remarks, possible extensions of the model as well as its limitations.

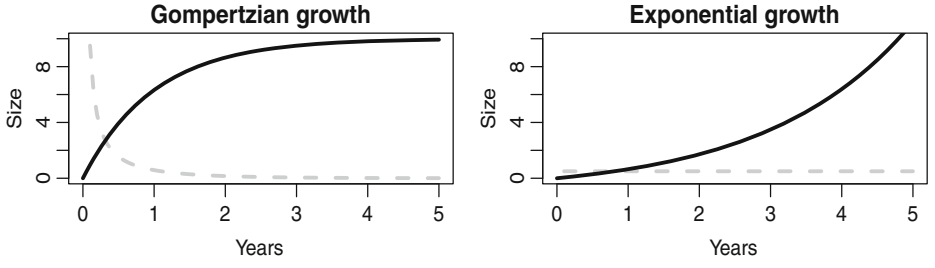

Fig. 2 Gompertzian and exponential growths and rates of growth (grey)

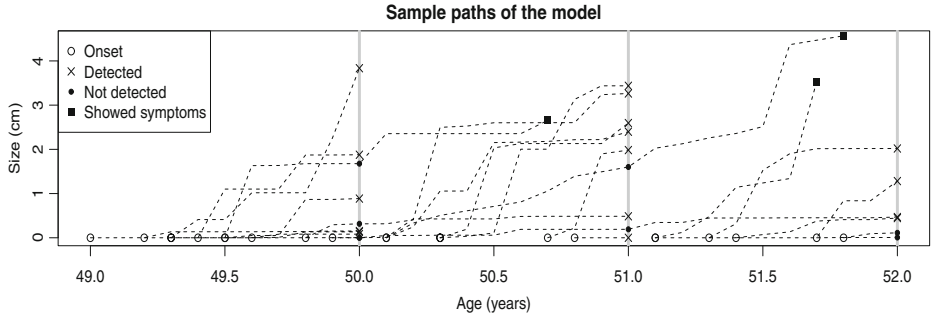

Fig. 3 Overview of the model for $m_1 = 5$, $m_2 = 0.2$, $\beta = 2$, $\mu = 3.86$ $s = 0.293$, $\lambda = 8$, $\xi = 0.25$, $b_0 = -2.5$ and $b_1 = 3.5$ with 3 screens, the increments of the gamma process were simulated on monthly intervals. The random threshold as well as the size dependent sensitivity can be observed

2 Model

Gamma processes have been identified as the main way to model degradation phenomena. A non-stationary gamma process can model degradation when there is some temporal variability in the degradation phenomenon. For a shape parameter $\eta(t)$ (time dependent) and a scale parameter β, the marginal distribution of a non-stationary gamma process Y_t at time t assuming that the process starts from 0 at $t = 0$ is the gamma distribution, namely:

$$f_{Y_t}(y) = \frac{1}{\Gamma(\eta(t))\beta^{\eta(t)}} y^{\eta(t)-1} \exp(-y/\beta) \ , \ y > 0.$$

where $\eta(t)$ is a non-negative, monotone increasing function for $t \geq 0$ and $\eta(0) = 0$. Recall that one says that $(Y_t)_{t\geq 0}$ is a non-stationary gamma process if:

- The increments of the gamma process in the interval $(t, t + h)$ denoted by $\Delta Y(t, h) = Y_{t+h} - Y_t$, $t > 0$, $h > 0$ are independent random variables over disjoint time intervals.
- Each increment $\Delta Y(t, h)$ follows a gamma distribution with constant scale parameter and time-varying shape parameter $\Delta\eta(t, h) = \eta(t + h) - \eta(t)$ for all t and h. The density of the increments is given by:

$$f_{\Delta Y(t,h)}(y) = \frac{1}{\Gamma(\Delta\eta(t, h))\beta^{\Delta\eta(t,h)}} y^{\Delta\eta(t,h)-1} \exp(-y/\beta), \ y > 0. \tag{1}$$

Now suppose that the tumor growth is a gamma process Y_t with a time dependent shape parameter $\eta(t) = m_1(1 - \exp(-m_2 t))$ (Norton 1988) and a constant scale parameter β. For a tumor becoming onset at age $t_p > 0$ with size 0, denote by $X(t_p, t_p, t_1) = Y_{t_1-t_p}$, $t_1 \geq t_p$ the tumor size at age t_1. Similarly, let $X(t_p, t_1, t_2) = Y_{t_2-t_p} - Y_{t_1-t_p}$, $t_2 \geq t_1 \geq t_p$ denote the increments from age t_1 till t_2 for a tumor becoming onset at t_p. Then, for a given t_p the density of $X(t_p, t_1, t_2)$ is given by $f_{X(t_p,t_1,t_2)}(x) = f_{\Delta Y(t_1-t_p,t_2-t_1)}(x)$.

As the exact onset of the disease is unknown, the preclinical intensity is assumed to be a random variable independent of the tumor growth. The distribution is chosen to be lognormal $LN(\mu, s^2)$ sub-density, meaning that it is the density multiplied by the lifetime risk, as not everyone in the population is affected by the disease. The choice of the lognormal distribution is based on the results of Lee and Zelen (1998) who found that the transition probability of breast cancer to the preclinical state is right skewed with a heavy tail, the lognormal distribution has similar properties. Denote by t_p the age in which the case moves

into S_p, and by r the life time risk of breast cancer, the density of the preclinical intensity is then given by:

$$w_{T_p}(t_p) = \frac{r}{t_p s\sqrt{2\pi}} \exp\left(-\frac{\ln(t_p - \mu)^2}{2s^2}\right),\ t_p > 0.$$

Now consider a breast cancer screening program consisting of K screens with fixed inter-screening time (time between two consecutive screens) Δ, suppose that the first screen takes place when a patient is aged t_0, Denote by $\tau_i = t_0 + (i-1)\Delta$ the time of the i^{th} screen. Moreover, assume that the sensitivity of the screen has a logistic form depending on the tumor size. This is motivated by the results of Michaelson et al. (2003) who determined estimates of the sizes at which breast cancers become detectable on mammographic and clinical grounds, showing that smaller tumors are very hard to detect, other factors include the density of the breast, age and others. Keeping our approach simple, we will only use the tumor size, namely, let us define the sensitivity as:

$$\Lambda(x) = \frac{1}{1 + \exp(-b_0 - b_1 x)},\ x \geq 0.$$

where b_0 and b_1 are parameters to be estimated and x is the tumor size. Since the logistic function takes values between 0 and 1 and is monotonically increasing in x, it is suitable for modelling sensitivity. The parameter b_0 determines the location of the curve while b_1 is the growth rate or the steepness of the curve.

Furthermore, suppose that progression into the clinical state S_c happens when the tumor size reaches a critical size denoted by C independent from the growth process, in other words, the patient's tumor size has reached a level in which it is noticed or causes symptoms. Assume that C is a gamma distributed random variable with shape λ and scale ξ, where λ and ξ are parameters to be estimated. Figure 3 shows some paths of the gamma process for the parameter values which will be used throughout the paper (Scenario 1). A tumor becomes onset at a random time t_p triggering the growth of a tumor, which then grows as a gamma process till it is either detected by a screen or by reaching its critical threshold (showing symptoms). The aim is to establish the distribution of the size of detected tumors on screens and then use suitable estimation methods for the parameter governing the process.

2.1 Sojourn Time

As we defined the sojourn time as the amount of time spent in the preclinical state S_p, it is then the amount of time before hitting the critical tumor size C. Denote by X_t the marginal distribution of the tumor size t time after the onset ($X(t_p, t_p, t + t_p)$) and consider the stopping time $T_C = \inf\{t \geq 0;\ X_t \geq C\}$, which is just the sojourn time with cdf F_{T_C}. Note that for some parametrizations, there could be a positive probability that T_C is infinite. In these cases, the distribution of T_C will not be a proper one and consequently the expected value and the variance of T_c are undefined. Nonetheless, after the parameters are estimated, one may truncate the distribution to get finite estimates of the mean and variance of cases with finite sojourn time as shown in Section 2.3. The cdf of the sojourn time may be derived using the law of total probability, namely:

$$F_{T_C}(t) = 1 - P(T_C \geq t) = 1 - \int_0^\infty P(T_C \geq t | C = x) \cdot f_C(x) dx.$$

and as the process is increasing, $P(T_c \geq t|C = x) = P(X_t < x)$. When C is gamma distributed, a closed form for the cdf was derived by Paroissin and Salami (2014):

$$F_{T_C}(t) = \frac{\Gamma(\eta_t + \lambda)}{\Gamma(\lambda+1)\Gamma(\eta_t)} \left(1 + \frac{\beta}{\xi}\right)^{-\eta_t} \left(1 + \frac{\xi}{\beta}\right)^{-\lambda} {}_2F_1\left(1, \eta_t + \lambda, \lambda + 1, \frac{\beta}{\xi + \beta}\right).$$

where ${}_2F_1$ is the generalized hypergeometric function of order (2,1) (Gradshteyn and Ryzhik 1965), note that ${}_pF_q$ is given by:

$${}_pF_q(a_1, \cdots, a_p; b_1, \cdots, b_q; z) = \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_q)_k} \frac{z^k}{k!}.$$

where $(a)_k$ is the Pochammer symbol defined as $(a)_k = \Gamma(a + k)/\Gamma(k)$. Assuming that η is differentiable, then the derivative of F_{T_C} is given by:

$$\begin{aligned} f_{T_C}(t) = \eta_t' \frac{\Gamma(\eta_t + \lambda)}{\Gamma(\lambda+1)\Gamma(\eta_t)} \left(\frac{\beta}{\beta+\xi}\right)^{\lambda} \left(\frac{\xi}{\beta+\xi}\right)^{\eta_t} \Bigg[(\psi(\eta_t + \lambda) - \psi(\eta_t) \\ - \log\left(\frac{\beta+\xi}{\xi}\right) {}_2F_1\left(1, \eta + \lambda, \lambda + 1, \frac{\beta}{\xi+\beta}\right) \\ + \frac{\beta}{\xi+\beta} \frac{1}{\lambda+1} F^{2:2,1}_{2:1,0}\left(\begin{matrix} \eta_t+\lambda+1, 2:1, \eta_t; 1 \\ 2, \lambda+2: \eta_t+1; - ; \end{matrix} \frac{\beta}{\beta+\xi}, \frac{\beta}{\beta+\xi}\right)\Bigg]. \end{aligned}$$

where ψ is the digamma function and $F^{2:2,1}_{2:1,0}$ is the Kampé de Fériet function (Ancarani and Gasaneo 2009).

2.2 Distribution of the Tumor Sizes on Screens

Next, the distribution of screened cases has to be established. Note that we will be dealing with subdensities, since not all cases will move to the preclincal state and not all preclinical cases are detected. In other words, we will be establishing the subdensity of the sizes of detected tumors on screens. That being said, we also need to consider that an individual participating in screen i means that the individual has not shown symptoms yet. This means that the subdensity is derived under the condition that the individual did not hit the critical threshold yet. Starting with the first screen, the subdensity is built up from those who have progressed to the preclinical state S_p before τ_1, did not move into the clinical state before τ_1, and were screened positively.

Since the random threshold, the onset, and the process are assumed independent, for a given preclinical age t_p the conditional subdensity of screened tumor sizes on the first screen is obtained by the density of the increments in (t_p, τ) weighted by the sensitivity $\Lambda(x)$ as it was screened positively and by the probability of not hitting the threshold before τ_1. Consequently, the conditional subdensity is then given by:

$$f_{\tau_1|t_p}(x) = f_{X(t_p,t_p,\tau_1)}(x) \cdot \Lambda(x) \cdot P(x < C). \tag{2}$$

The full subdensity is derived by applying the law of total probability to Eq. 2, however, as the participants in screen τ_1 have not yet become clinical, the subdensity needs to be adjusted by the probability of a case not becoming clinical before τ_1, therefore we have:

$$f_{\tau_1}(x) = \Lambda(x) \cdot P(x < C) \cdot \frac{\int_0^{\tau_1} w(t_p) \cdot f_{X(t_p,t_p,\tau_1)}(x)\, dt_p}{1 - \int_0^{\tau_1} w(t_p) F_{T_C}(\tau_1 - t_p)\, dt_p}. \tag{3}$$

The subdensity of screened tumor sizes on the second screen is derived in a similar fashion, though we need to consider two parts of the subdensity separately. The first part corresponds

to those who have moved into S_p during (τ_1, τ_2) and the second is for those who moved to S_p before τ_1. Denote by $f_{\tau_j,(\tau_i,\tau_{i+1})}$ the contribution of cases developing between (τ_i, τ_{i+1}) to the subdensity on τ_j. Starting with the first part of the the subdensity $f_{\tau_2,(\tau_1,\tau_2)}$, that is built from cases becoming preclinical between τ_1 and τ_2, not becoming clinical before τ_2 and then screened positively:

$$f_{\tau_2,(\tau_1,\tau_2)}(x) = \Lambda(x) \cdot P(x < C) \cdot \frac{\int_{\tau_1}^{\tau_2} w(t_p) \cdot f_{X(t_p,t_p,\tau_2)}(x)\, dt_p}{1 - \int_{\tau_1}^{\tau_2} w(t_p) F_{T_C}(\tau_2 - t_p)\, dt_p}.$$

The second part $f_{\tau_2,(0,\tau_1)}$ corresponds to those who have moved to S_p before τ_1, their disease was not detected by the first screen and stayed in the preclinical state till τ_2 when they were finally screened positively. Therefore:

$$\begin{aligned} f_{\tau_2,(0,\tau_1)}(x) \;=\; & P(x < C) \cdot \Lambda(x) \cdot \int_0^{\tau_1} \int_0^{x} w(t_p) \cdot f_{X(t_p,t_p,\tau_1)}(x_1) \cdot (1 - \Lambda(x_1)) \\ & f_{X(t_p,\tau_1,\tau_2)}(x - x_1) \cdot\; dx_1\, dt_p \;\cdot \frac{1}{1 - \int_0^{\tau_1} w(t_p) F_{T_C}(\tau_2 - t_p)\, dt_p}. \end{aligned}$$

As a result, the subdensity of screened tumor sizes on the second screen is

$$f_{\tau_2}(x) = f_{\tau_2,(0,\tau_1)}(x) + f_{\tau_2,(\tau_1,\tau_2)}(x).$$

In general, the subdensity on screen i can be derived following the same logic, dividing the time-line into disjoint intervals $(0, \tau_1), (\tau_1, \tau_2), \cdots, (\tau_{i-1}, \tau_i)$. Thus, it is given by the sum of the contributions. Namely:

$$f_{\tau_i}(x) = f_{\tau_i,(0,\tau_1)}(x) + \sum_{2 \le j \le i} f_{\tau_i,(\tau_{j-1},\tau_j)}(x),$$

where $f_{\tau_i,(\tau_{j-1},\tau_j)}(x)$ is given by:

$$\begin{aligned} f_{\tau_i,(\tau_{j-1},\tau_j)}(x) \;=\; & \int_{\tau_{j-1}}^{\tau_j} \int_0^{x} \int_{x_j}^{x} \cdots \int_{x_{i-1}}^{x} \big[w(t_p) f_{X(t_p,t_p,\tau_j)}(x_j) \cdot (1 - \Lambda(x_j)) \cdot \\ & f_{X(t_p,\tau_j,\tau_{j+1})}(x_{j+1} - x_j) \cdot (1 - \Lambda(x_{j+1})) \cdots \\ & f_{X(t_p,\tau_{i-1},\tau_i)}(x - x_{i-1}) \cdot \\ & \cdot\, dx_{i-1} \cdots dx_{j+1}\, dx_j\, dt_p \big] \cdot \frac{\Lambda(x) \cdot P(x < C)}{1 - \int_{\tau_{j-1}}^{\tau_j} w(t_p) F_{T_C}(\tau_i - t_p)\, dt_p}. \end{aligned} \tag{4}$$

Note that the distribution of interval cases is derived in a similar manner. Interval cases are defined as those who progress into S_c between two screens, namely, these are ones who reach the critical threshold between screens (see Fig. 3). However, we chose not use them in the current model as data between screens is not always available.

2.3 Estimation of Parameters

Under the current setup, the parameters to be estimated are those for the sensitivity (b_0 and b_1), those for the preclinical intensity (μ and s), the parameters controlling the gamma process (m_1, m_2 and β) and the random threshold parameters λ and ξ.

However, as the integrals in expression (4) do not have a closed form, numerical integration must be used. But as multidimensional numerical integration is slow and calculating the full likelihood means computing these integrals for all the sample elements and parameters, that would not be computationally feasible. Thus, we suggest using the so called minimal divergence estimators instead. These estimators are based on dividing the sample space

into intervals and minimizing the divergence between observed and expected distributions, therefore limiting the number of integrals that need to be computed.

From a theoretical point of view, if we have a set of observations $\mathcal{X}$, grouping them into $\{A_1, \cdots A_M\}$ defines a discrete statistical model in which the expected probabilities of A_i are denoted by $q_i(\theta) = P_\theta(A_i)$ for $i = 1 \cdots M$ and let $\hat{p}_i = n_i/n$ be the relative frequency of A_i, $i = 1 \cdots M$. For fixed $(n_1, \cdots, n_M)$ the likelihood is:

$$P_\theta(N_1 = n_1, \cdots, N_M = n_M) = \frac{n!}{n_1! \cdots n_M!} q_1(\theta)^{n_1} \cdots q_M(\theta)^{n_M}.$$

Therefore, the log-likelihood can be written as:

$$\begin{aligned} \log P_\theta(N_1 = n_1, \cdots, N_M = n_M) &= \log \frac{n!}{n_1! \cdots n_M!} + n \sum_{i=1}^{M} \frac{n_i}{n} \log(q_i) \\ &= \log \frac{n!}{n_1! \cdots n_M!} + n \sum_{i=1}^{M} (\hat{p}_i \log\left(\frac{q_i}{\hat{p}_i}\right) + \hat{p}_i \log(\hat{p}_i)) \\ &= -n D^{\text{K-L}}(\hat{P}, Q(\theta)) + l. \end{aligned} \tag{5}$$

where $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_M)$, $Q(\theta) = (q_1(\theta), \cdots q_M(\theta))$, $D^{\text{K-L}}$ is the Kullback-Leibler divergence (Kullback and Leibler 1951) $D^{\text{Kullback}}(\hat{P}, Q(\theta)) = -\sum_{i=1}^{M} \hat{p}_i \log\left(\frac{q_i}{\hat{p}_i}\right)$ and l is a constant in θ. Then to estimate θ by the discrete model maximum-likelihood estimator is equivalent to minimize the Kullback-Leibler divergence.

However, the Kullback-Leibler divergence is not the unique divergence measure, one can choose $\hat{\theta}$ as the estimator which solves the following:

$$D(\hat{P}, Q(\hat{\theta})) = \inf_{\theta \in \Theta} D(\hat{P}, Q(\theta)).$$

With D being a divergence measure for our discrete model. Some of the other divergence measures include the chi-squared divergence, Hellinger divergence , Burbea– Rao divergence (Burbea and Rao 1982) and several others. Minimum divergence estimators are quite popular in the literature, especially when the likelihood has a complex form. Asymptotic properties of these estimators and comparisons to the maximum likelihood estimator are also studied in the literature, see Broniatowski (2014), Jimenz and Shao (2001), and many others. We decided to use the χ^2 measure as it is well known and easy to interpret. Recall that the χ^2 divergence is defined as:

$$\chi^2(\hat{P}, Q(\theta)) = n \sum_{i=1}^{M} \frac{(\hat{p}_i(\theta) - q_i(\theta))^2}{q_i(\theta)}.$$

Translating this to our setup, binning the observations on screens into intervals leads to a similar discrete model. One way to carry this out is by grouping the observations on each screen into intervals with approximately equal frequencies. Formally, let us introduce some notations: denote by M the chosen number of intervals, denote the resulting intervals on screen j by $Y_{i,j}$ $i = 1 \cdots M$, $j = 1 \cdots K$. Let N_j the number of participants in the j^{th} screen and denote by $\hat{p}_{i,j} = n_{Y_{i,j}}/N_j$ the observed relative frequency of $Y_{i,j}$. Introduce $q_{i,j}(\theta)$ as the probability of having a tumor size in $Y_{i,j}$ on the j^{th} screen i.e. $q_{i,j}(\theta) = \int_{Y_{i,j}} f_{\tau_j}(x)dx$. Furthermore, denote by $Y_{0,j}$ the number of cases who participated in screen

j but no tumor was detected ($q_{0,j}(\theta) = 1 - \sum_{i=1}^{M} q_{i,j}(\theta)$). The divergence to be minimized is then:

$$d^2 = \sum_{j=1}^{K} N_j \sum_{i=0}^{M} \frac{(\hat{p}_{i,j}(\theta) - q_{i,j}(\theta))^2}{q_{i,j}(\theta)}. \tag{6}$$

Note that if the model is to be applied to a data set with different ages at program entry, one can extend (6) by summing over the ages at program start t_0. On the other hand, if data about interval cases is available, one can further extend the model by including the distance between expected and observed interval cases, therefore using all the available data.

Having the parameters estimated, the main aim is to calculate the sojourn time distribution, however, under some parametrizations of the process, there is a positive probability that the tumor will never reach the critical threshold C. In other words, the tumor would never become symptomatic. As a result, the expected value of the sojourn time will be mathematically infinite. In fact, a serious concern around screening programs is that they may cause overdiagnosis. Overdiagnosis is the diagnosis of a medical condition that would never have caused any symptoms or problems. Reported estimates of breast cancer overdiagnosis range from 0% to 54% (Elmore and Fletcher 2012).

Therefore, after the parameters are estimated, one can compute the probability of never showing symptoms by $P(NS) = 1 - F_{T_C}(\infty)$. If this probability is positive, then one has to truncate the distribution of the sojourn time to get estimates for the mean and variance. The truncation is done at the maximum realistic value of the sojourn time (Q). The adjusted distribution is then given by:

$$F'_{T_C}(t) = \begin{cases} \frac{F_{T_C}(t)}{F_{T_C}(Q)} \text{ if } t < Q \\ 1 \text{ if } t \geq Q \end{cases} \tag{7}$$

The expected value and the variance of the sojourn time are then directly obtainable even if η is not differentiable using $E[T_C] = \int_0^\infty (1 - F'_{T_C}(t))dt$ and $E(T_C^2) = 2\int_0^\infty t(1 - F'_{T_C}(t))\,dt$. Note that the truncation is done to get finite expected values and is not incorporated into the model.

2.4 Lead Time

Lead time bias in the current framework is the amount of time that a screened case would have needed to show symptoms. Specifically, it is the unobservable future time needed to exceed C. From a reliability point of view, lead time is the remaining preclinical lifetime till the degradation reaches C. For a given threshold $C = c$, onset t_p and size at detection x_τ. The conditional survivor function of the lead time R is given by the probability that the tumor size will not exceed c in time t_r after the screen τ as:

$$R_\tau(t_r|x_\tau,\ t_p,\ c\ ) = P\{X(t_p, \tau, \tau + t_r) \leq c - x_\tau\}.$$

To release the condition on t_p, we need to get the the conditional density of $t_p|X_\tau = x_\tau$ denoted by $w'(t_p|x_\tau)$ which is obtainable by Bayes law. Namely:

$$w'(t_p|x_\tau) = \frac{w(t_p) f_{X(t_p,t_p,\tau)}(x_\tau)}{\int_0^\tau w(y) f_{X(y,y,\tau)}(x_\tau)dy}\ (0 < t_p < \tau).$$

Note that a detected tumor of size x_τ means that the threshold C must be larger than x_τ. Taking all of that into account gives:

$$R(t_r|x_\tau) = \frac{\int_0^\tau \int_{x_\tau}^\infty w'(t_p|x_\tau) P\{X(t_p, \tau, \tau + t_r) \leq c - x_\tau\} f_C(c) dc dt_p}{P(C > x_\tau)}.$$

As the lead time is directly related to the sojourn time, a similar truncation to (7) is needed to get finite values. In other words, the lead time will be computed given that detected cases are not overdiagnosed. A favorable outcome of our approach is that one is able to estimate the expected value and the variance of the lead time based on the tumor size at detection x_τ. Lead time bias is corrected by deducting the expected lead time from the overall survival of a screened patient.

3 Applications

In order to check the performance of the model, simulations were carried out by discretizing the time-line into h-sized intervals. Movement into the preclinical state is simulated by a Bernoulli random variable with probability $p_h = \int_{t_{h-1}}^{t_h} w(t_p) dt_p$ where t_{h-1} and t_h are the boundaries for the h-sized interval. When the Bernoulli (p_h) gives a success, the gamma process is then simulated starting from $(t_{h-1} + t_h)/2$. A patient in a screen if it did not reach the critical threshold C or if it is still disease free. Note that the threshold is simulated by a gamma random variable and the screens are simulated using a Bernoulli random variable with probability $\Lambda(x)$. The simulator is initiated under the assumption that all cases were disease free B years before the first screen (occurring when cases are 50 years old), where B is the maximum feasible value of the sojourn time. Patients which show symptoms before the first screen are discarded. The simulator is run until the desired number of participants on the first screen N is reached.

The preclinical intensity parameters used in simulations are $u = 3.86$ and $s = 0.293$, these values result in an average preclinical age of 50 and a standard deviation of 15, the risk r is set to one for the sake of getting more preclinical cases. Parameters for the critical threshold are defined as $\lambda = 8$ and $\xi = 0.25$ resulting in an average critical size of 2 cm and a variance of 0.5.

That being said, the first aim is to study the performance of the model for different number of participants combined with different number of intervals (M) on screens. For that purpose, we simulated Scenario 1, the defined process parameters in this scenario are: $m_1 = 5$, $m_2 = 0.2$ and $\beta = 2$ resulting in a mean sojourn time of 2.15 years (truncated at 20 years). The defined values for sensitivity are $b_0 = -2.5$ and $b_1 = 3.5$. We decided to simulate 3 screening programs for a single age group $(t_0 = 50)$ with different number of participants: $N = 10000$, $N = 50000$ and $N = 200000$. 50 datasets were generated for each of the programs, the model was run for $M = 5$, 10, 20, 30 and 40 on each dataset.

The second aim is to check the performance of the model under different setups, for that purpose, we simulated Scenario 2 with $N = 50000$. This scenario was simulated to mimic extremely aggressive growth with $m_1 = 10$, $m_2 = 1$ and $\beta = 2.5$ resulting in an adjusted mean sojourn time of 0.144 years (truncated at 2 years). The defined sensitivity parameters are $b_0 = 1.5$ and $b_1 = 3$ leading to a larger probability of detection for smaller tumors combined with a steeper increase in the sensitivity as tumor size increases. The model is run on each dataset (N=50000) for the different values of M and the results are presented in Table 2. Plots of the sojourn time cdf and screening sensitivity are shown in Fig. 5.

From a practical point of view, the critical threshold parameters λ and ξ are usually known. Estimates of these parameters can be directly obtained by applying maximum likelihood to symptomatic cases who never participated in a screen. Likewise, the parameters controlling the preclinical intensity for breast cancer are also given in terms of age-specific incidence rates (see Lee and Zelen (1998)). Employing this, we decided to fix preclinical intensity and critical threshold parameters, leaving only the parameters controlling the process and the sensitivity to be estimated $\theta = (m_1, m_2, \beta, b_0, b_1)$. We binned the measurements on screens into $M = 5, 10, 20, 30, 40$ intervals with almost equal frequencies. The numerical integration is implemented using the statistical software ***R*** by the function *suave* from the package *cubature*, *suave* implements a Monte Carlo algorithm for multidimensional numerical integration by importance sampling combined with a globally adaptive subdivision strategy (Hahn 2005). The minimization of the distance is done using the function *optim*. The mean of the estimates resulting from the 50 datasets and their standard deviations are displayed in Table 1.

Starting with the first scenario, the tumor size densities on the first and second screens for $N = 10000$ and $N = 200000$ are displayed in Fig. 4. Note that the number of detected cases on the first screen is 8516 in the large program and 417 for the small one. Whereas on the second screen the number of detected cases is 3731 and 186 respectively. It is noticed that the resulting densities have many irregularities, additionally we could observe that the tumor size on the second screen is denser around its peak. This is natural since larger cases either were detected on the first screen or reached the threshold during the inter-screening

Table 1 Results for different sample sizes at first screen (N) and number of intervals (M) for Scenario 1: $m_1 = 5, m_2 = 0.2, \beta = 2, b_0 = -2.5, b_1 = 3.5$

Parameter	Actual value	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
		$M = 5$		$M = 10$		$M = 20$		$M = 30$		$M = 40$	
		N=200000									
m_1	5	4.977	0.476	4.965	0.444	5.025	0.38	5.031	0.29	5.019	0.289
m_2	0.2	0.197	0.014	0.202	0.014	0.199	0.009	0.2	0.011	0.199	0.015
β	2	2.052	0.554	1.997	0.3	1.998	0.228	1.999	0.202	2.014	0.216
b_0	−2.5	−2.347	0.738	−2.481	0.323	−2.49	0.259	−2.515	0.162	−2.484	0.151
b_1	3.5	3.596	0.752	3.514	0.159	3.54	0.181	3.485	0.143	3.463	0.142
		N=50000									
m_1	5	5.136	0.942	5.093	0.616	5.078	0.519	5.089	0.541	5.009	0.49
m_2	0.2	0.21	0.048	0.196	0.044	0.198	0.035	0.199	0.028	0.197	0.027
β	2	2.124	0.683	2.08	0.539	2.093	0.433	2.024	0.486	2.01	0.506
b_0	−2.5	−2.482	0.915	−2.427	0.589	−2.406	0.511	−2.58	0.352	−2.414	0.391
b_1	3.5	3.334	0.601	3.449	0.624	3.366	0.372	3.439	0.271	3.491	0.238
		N=10000									
m_1	5	5.058	1.867	5.08	1.637	4.936	1.18	5.051	0.923	5.019	0.872
m_2	0.2	0.195	0.07	0.206	0.067	0.202	0.055	0.201	0.06	0.212	0.056
β	2	1.811	0.938	1.93	0.698	1.978	0.55	2.062	0.472	1.92	0.513
b_0	−2.5	−2.581	0.985	−2.501	0.835	−2.244	0.685	−2.099	0.741	−2.51	0.744
b_1	3.5	−3.456	0.672	−3.64	0.632	−3.474	0.551	−3.561	0.443	3.436	0.415

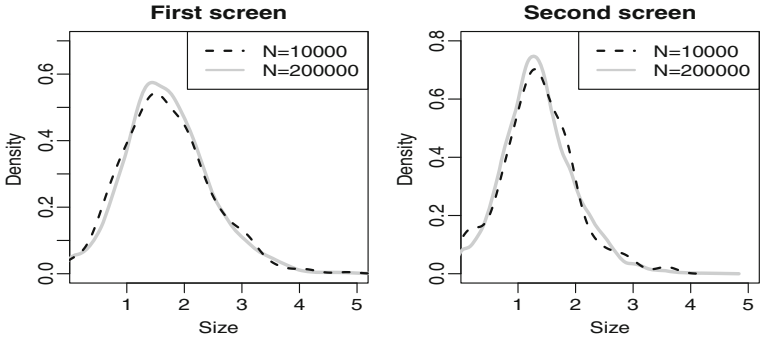

Fig. 4 Density of the tumor size on the first and the second screen for scenario 1: $m_1 = 5, m_2 = 0.2, \beta = 2, b_0 = -2.5, b_1 = 3.5$

interval. Nonetheless, even the small sample seems to be very informative, although some irregularities in the density are noticed.

From the first block of Table 1 where $N = 200000$, it is noticed that the model performs really well and is nearly unbiased even for $M = 5$. We observed a decrease in the standard deviations with the increase of M, however there is no significant decrease beyond M=30. Moving to the second block, the estimates are good although with higher standard deviations. Additionally, we still observe an increase in the precision with the increase of M. In the third block, where $N = 10000$ it is noticed that the standard deviations are much larger. This is caused by the irregularities in the densities for the small sample size, nonetheless, the model still gives acceptable estimates.

Furthermore, we have noticed that there is a strong correlation between the elements of θ, the strongest positive correlation was found between b_0 and m_2 (its value for $N = 200000$ and $M = 40$ was 0.753). Recall that m_2 controls the process rate to reach m_1, and b_0 controls the location of the curve, in other words, b_0 adjusts the weight of the process on a screen. Decreasing b_0 means less weight on smaller tumors, the model adjusts to this by a smaller rate m_2 and vice versa. The strongest negative correlation was found to be between the scale of the process β and m_1 (for $N = 200000$ and $M = 40$ it was -0.724), both of which control the expected value and the variance of the process, so to preserve the balance for decreasing values of β the values for m_1 are increased and vice versa.

Moving on to the results of Scenario 2 displayed in Table 2, it seems that there is a small bias combined with large standard deviations. We observed a large variation in sensitivity estimates (b_0 and b_1). This is likely due to the sharp decrease of the number of screened

Table 2 Scenario 2 results for $N = 50000$ and number of intervals M

		$M = 5$		$M = 10$		$M = 20$		$M = 30$		$M = 40$	
Parameter	Actual value	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
m_1	10	9.77	1.706	10.153	0.775	9.872	0.587	9.978	0.714	9.828	0.725
m_2	1	0.996	0.302	1.029	0.246	1.035	0.24	1.016	0.228	1.011	0.207
β	2.5	2.631	1.015	2.61	0.69	2.533	0.699	2.412	0.586	2.381	0.525
b_0	−1.5	−1.613	1.264	−1.525	1.196	−1.453	1.039	−1.63	0.813	−1.692	0.803
b_1	3	2.847	1.238	2.945	0.921	2.773	0.746	3.108	0.955	2.943	0.979

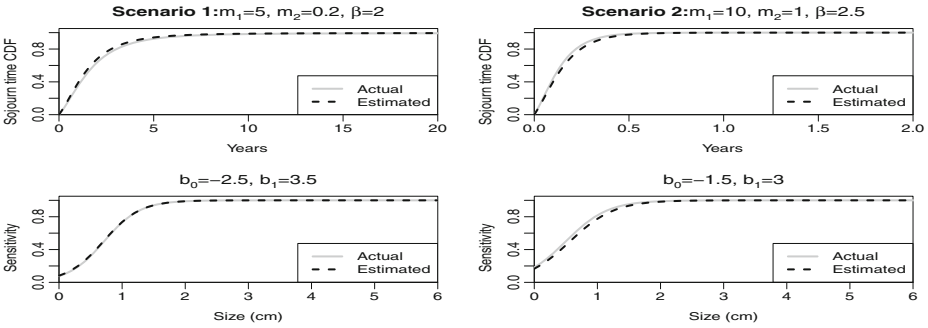

Fig. 5 Plots for the sojourn time (top), sensitivity (bottom) for $N = 50000$ and $M = 20$ for the two scenarios

cases in this scenario. In fact, under a very short sojourn time, the chances for a case to reach a screen before becoming symptomatic are quite slim. As a result, only around 208 cases are detected on the first screen. Nonetheless, the resulting average sojourn time distribution seen in the top right part of Fig. 5 is very close to the actual one. The same is true for the sensitivity, however it is noticed that on average, the estimated sensitivity is slightly higher than the actual one.

After estimating the parameters, we calculated the lead time for cases detected on the first screen (aged 50), and had a tumor of size 1 cm on detection. After truncating the lead time distribution, the resulting expected lead time bias is 1.849 years for scenario 1 (and 0.165 years for scenario 2). This shows the magnitude of the bias caused by early detection, as a screened case with a tumor of size 1 cm would appear to have survived 1.849 years more than a symptomatic with the same onset and date of death. The implications of this are very serious, as any administered treatments after screening might falsely appear to be effective due to the prolonged survival. Furthermore, it is clearly beneficial to link lead time bias with the tumor size on detection, as the tumor size gives an indication for the duration of which the tumor size was asymptomatic, therefore giving information of how much more time it needs to show symptoms.

4 Concluding Remarks

Although the proposed model is somewhat computationally expensive, it proves to be an accurate and a powerful tool to use with degradation processes triggered at a random onset. The model is very flexible, as one is free to choose e.g the form of η. There is definitely room for further research, the extended gamma process can also be used (Guida et al. 2012) if the mean over variance ratio is not constant. Moreover, it is also possible to use the transformed gamma process (Giorgio et al. 2018) if damage accumulates gradually over time in a sequence of tiny increments in which the degradation increments over disjoint time intervals are not independent.

One more advantage of using the gamma process-based approach is that with some modifications, one is able to incorporate a process with random covariates as in (Lawless and Crowder 2004). For instance, if one wishes to investigate the effect of age on tumor growth, adding a covariate corresponding to age in the shape of the process should do the trick.

On the other hand, one main limitation for focusing on tumor size as a degradation measure is that the approach is limited to study solid cancers. Besides, some types of cancer are expressed through multiple tumors and are not restricted to a single primary one. However,

it might be possible to find a proper measure of degradation on diagnosis in case of other degenerative diseases.

Acknowledgments Open access funding provided by Eötvös Loránd University (ELTE). The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.2-16-2017-00015).

References

Abdel-Hameed M (1975) A gamma wear process. IEEE Trans Reliability 24(2):152–153

Ancarani LU, Gasaneo G (2009) Derivatives of any order of the Gaussian hypergeometric function 2F1(a,b,c,z) with respect to the parameters a, b and c. J Phys Math Theor 42(39):1–10

Burbea J, Rao C (1982) On the convexity of some divergence measures based on entropy functions. IEEE Trans Inform Theor 28(3):489–495

Broniatowski M (2014) Minimum divergence estimators, maximum likelihood and exponential families. Stat Probability Lett 93:27–33

Collins VP, Loeffler RK, Tivey H (1956) Observations on growth rates of human tumors. Am J Roentgen 76:988–1000

Duffy S, Chen H, Tabar L, Day N (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. Statistics in Medicine 14:1531–1543

Elmore JG, Fletcher SW (2012) Overdiagnosis in breast cancer screening: time to tackle an underappreciated harm. Annals Internal Med 156(7):536–537

Giorgio M, Guida M, Pulcini G (2018) The transformed gamma process for degradation phenomena in presence of unexplained forms of unit-to-unit variability. Qual Reliab Engng Int 34:543–562

Gordis L (2008) Epidemiology. Saunders, Philadelphia, p 318

Guida M, Postiglione F, Pulcini G (2012) A time-discrete extended gamma process for time-dependent degradation phenomena. Reliability Engineering & System Safety 105:73–79

Gradshteyn IS, Ryzhik IM (1965) Table of integrals, series and products. Academic Press, City

Hahn T (2005) CUBA- a library for multidimensional numerical integration. Comput Phys Commun 168:78–95

Hijazy A, Zempléni A (2020) How well can screening sensitivity and sojourn time be estimated? http://arxiv.org/abs/2001.07469

Jimenz R, Shao Y (2001) On robustness and efficiency of minimum divergence estimators. Test 10:241

Kullback S, Leibler R (1951) On information and sufficiency. Arm Math Statist 22:79–86

Laird AK (1964) Dynamics of tumor growth. Br J Cancer 13:490–502

Lawless J, Crowder M (2004) Covariates and random effects in a gamma process model with application to degradation and failure. Lifetime Data Anal 10(3):213–227

Lee SJ, Zelen M (1998) Scheduling periodic examinations for the early detection of disease: applications to breast cancer. J American Stat Association 93:1271–1281

Michaelson J, Satija S, Moore R, Weber G, Halpern E, Garland A, Kopans D, Hughes K (2003) Estimates of the sizes at which breast cancers become detectable on mammographic and clinical grounds. Journal of Women's Imaging 5(1):3–10

Norton L (1988) A Gompertzian model of human breast cancer growth. Cancer Res 48:7067–7071

Paroissin C, Salami A (2014) Failure time of non homogeneous gamma process. Commun Stat Theor Methods 43(15):3148–3161

Schwartz M (1961) A biomathematical approach to clinical tumor growth. Cancer 14:1272–1294

Speer JF, Petrovsky VE, Retsky MW, Wardwell RH (1984) A stochastic numerical model of breast cancer that simulates clinical data. Cancer Res 44:4124–4130

Wu D, Rosner G, Broemeling L (2005) MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. Biometrics 61(4):1056–1063

Weedon-Fekjaer H, Vatten LJ, Aalen O, Lindqvist B, Tretli S (2005) Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results. J Med Screening 12:172–178

Zelen M, Feinleib M (1969) On the theory of screening for chronic diseases. Biometrika 56(3):601–614