Published in final edited form as:

Math Biosci. 2007 August ; 208(2): 644–657.

Identifiability of the Joint Distribution of Age and Tumor Size at Detection in the Presence of Screening

Leonid Hanin^{a,*} and Andrei Yakovlev^b

^a*Department of Mathematics, Idaho State University, Pocatello, ID 83209-8085, USA*

^b*Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, NY 14642, USA*

Abstract

In recent years, a stochastic model of cancer development and detection allowing for arbitrary screening schedules has been developed and applied to analysis of screening trials and population-based cancer incidence and mortality data. The model is entirely mechanistic, builds on a minimal set of biologically plausible assumptions, and yields the joint distribution of tumor size and age of a patient at the time of diagnosis. Whether or not parameters of the model can be estimated from data generated by cohort studies depends on model identifiability. The present paper provides a proof of this important property of the model.

Keywords

Cancer detection; Cancer natural history; Cancer screening; Model identifiability; Tumor latency; Tumor size

1. Introduction

The importance of model identifiability is frequently overlooked in mathematical modeling of biomedical processes. In particular, numerous overly detailed simulation models of cancer screening proposed in the literature tend to be overparameterized, and the utility of these models in applications to analysis of real data is severely limited by their nonidentifiability. A mathematical or simulation model whose parameters are not identifiable is practically useless, at least for the purpose of parameter estimation, because there may be many (typically, infinitely many) parameter combinations providing the same fit to the data. Therefore, prior to estimating model parameters, a theoretical proof of model identifiability has to be provided. For an extensive discussion of the property of model identifiability and its various applications, the reader is referred to [1].

Generally speaking, model nonidentifiability can always be overcome by a proper reparameterization. However, this task may be technically very difficult and, more importantly, the original model parameters cannot be recovered from the new (typically smaller) set of identifiable parameters. The most common way of gaining model identifiability without

* Corresponding author. Tel.: +1 208 282 3293; fax: +1 208 282 2636, *E-mail address:* hanin@isu.edu (L. Hanin).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

collapsing model parameters into a smaller set of their combinations is to augment the set of model outputs and invoke additional sources of information associated with these outputs.

For example, the traditional approach to modeling cancer screening tends to describe the process of tumor development in only one dimension, that is, the *time natural history* of cancer [2-27]. A broader methodological idea is to construct a stochastic model of cancer development and detection that yields the multivariate distribution of observable variables at the time of diagnosis. Some authors [5,20,21] have long realized the advantages of multivariate data analysis in screening studies, but specific modeling and inferential techniques require a much higher level of sophistication than that in the earlier attempts at multivariate methodology. By invoking multivariate observations we seek to ensure model identifiability and improve estimation of unobservable parameters of cancer latency.

The most essential element of the traditional model of time natural history of cancer is the sojourn time in the preclinical stage of cancer development. The common practice is to assume an exponential distribution for this random variable. If this assumption is adopted the model is identifiable in a wide class of functions describing the rate of entry into the preclinical stage [28]. However, no identifiability results of such generality are known for any other parametric family of the sojourn time distributions.

A new model of cancer screening has been recently developed and applied to analyze the Canadian randomized breast screening trials (CNBSS) and the U.S. national trends of breast cancer incidence and mortality [28-32]. This model includes two components: a mechanistic two-stage model of carcinogenesis intended to describe the latent period in breast cancer development and the so-called quantal response model [29] designed to describe a relationship between the event of detection and tumor size; the latter relationship applies equally to both spontaneously detected (incident) and screen-detected (prevalent) cases. Proceeding from the proposed model it is possible to derive an analytic expression for the joint distribution of tumor size and age of a patient at diagnosis, thereby providing the necessary groundwork for maximum likelihood inference [28]. The model and associated mathematical framework are introduced in Section 2.

Several identifiability results related to the model have been reported up to date [1,29]. However, all of them are concerned with the model of cancer detection in the absence of screening. In the present paper, we address a much more challenging problem of identifiability of the joint distribution of age and tumor size at detection in the presence of an arbitrary screening schedule. The main result is presented in Section 3.

2. The Model

The latent period of tumor development can be broken down into the following three stages:

- formation of initiated cells
- promotion of initiated cells resulting in the appearance of the first malignant clonogenic cell termed the onset of the disease
- subsequent tumor growth and progression until the event of detection occurs

We first address the case when no screening is in place. Let T be the age of an individual at tumor onset, and W the time of spontaneous (clinical) tumor detection measured from the onset of the disease. Clearly, the random variable (r.v.) T represents the total duration of the first two stages of carcinogenesis. Throughout the paper we will denote by F_X , $\bar{F}_X$ and p_X the cumulative distribution function (c.d.f.), the survival function and the probability density function (p.d.f.) of any r.v. (or random vector) X , respectively.

One of the most widely used stochastic models of carcinogenesis is a two-stage clone expansion model proposed by Moolgavkar et al. [33-35]. A stationary version of this biologically motivated model yields the following proper survival function

$$\bar{F}_T(t) := Pr(T > t) = \left[\frac{(A+B)e^{At}}{B+ Ae^{(A+B)t}} \right]^\rho, t \geq 0,$$

from which the density $p_T(t)$ can be derived. Here $A, B, \rho > 0$ are identifiable parameters of the model [1,36-38]. Under this model, the process of initiation is assumed to be Poisson, while the clonal expansion of initiated cells and the event of malignant transformation follow the postulates of a linear birth and death process with two types of cells. The parameter ρ is the ratio of the initiation rate and the rate of proliferation of initiated cells. The parameters A and B can be interpreted as parameters of the promotion time distribution [29]. This model has proven to provide a good fit to diverse data on animal and human carcinogenesis [39,40].

Now we need to incorporate a mechanism of cancer detection into the model. Introduce a r.v. S to represent tumor size (the number of cells in a tumor) at spontaneous detection. Suppose that tumor growth can be modeled as a deterministic function $f: [0, \infty) \rightarrow [1, \infty)$ with $f(0) = 1$, so that $S = f(W)$. We hypothesize in addition that

- (1) the r.v.'s T and W are absolutely continuous and independent;
- (2) the function f is differentiable and $f' > 0$;
- (3) the hazard rate h_W for spontaneous (clinical) tumor detection is proportional to the current tumor size with coefficient $\alpha > 0$:

$$h_W(w) = \alpha f(w), w \geq 0. \quad (1)$$

From the above assumptions it follows that

$$\bar{F}_W(w) = e^{-\int_0^w h_W(v) dv} = e^{-\alpha \int_0^w f(v) dv}, p_W(w) = \alpha f(w) e^{-\alpha \int_0^w f(v) dv}, w \geq 0. \quad (2)$$

Observe that the distribution of r.v. W is proper. Also, it follows from (2) that

$$p_S(s) = \alpha s g'(s) e^{-\alpha \int_0^{g(s)} f(v) dv}, s \geq 1,$$

where g stands for the inverse function for f : $g = f^{-1}$. In the case of deterministic exponential tumor growth with rate $\lambda > 0$ ($f(w) = e^{\lambda w}$) we have

$$p_W(w) = \alpha e^{\lambda w - \frac{\alpha}{\lambda}(e^{\lambda w} - 1)}, p_S(s) = \frac{\alpha}{\lambda} e^{-\frac{\alpha}{\lambda}(s-1)}, s \geq 1, w \geq 0.$$

Thus, in this particular case tumor size at detection S follows a translated exponential distribution with parameter α/λ , while the distribution of age at detection measured from the disease onset is a Gompertz distribution. It should be noted that the random variable W has the same meaning as the preclinical stage duration within the traditional structure of the time natural history of cancer.

To obtain an expression for the joint distribution of age of a patient and tumor size at the time of diagnosis, consider the random vector $\mathbf{X} := (T + W, S)$ with the two components representing age and tumor size at diagnosis, respectively. The p.d.f. of $\mathbf{X}$ is given by

$$p_X(u, s) = p_T(u - g(s)) p_S(s), u \geq g(s), s \geq 1, \quad (3)$$

[29]. In the particular case of exponential tumor growth we obtain

$$p_x(u, s) = \frac{\alpha}{\lambda} e^{-\frac{\alpha}{\lambda}(s-1)} p_T(u - \frac{\ln s}{\lambda}), u \geq 0, 1 \leq s \leq e^{\lambda u}. \quad (4)$$

The distribution (3) has been proven to be identifiable [1].

In practice, it is not the number of tumor cells S that is observable but the volume V in appropriate units, and one needs to change variables using the equality $S = V/\gamma$, where γ is the volume of a single tumor cell ($\gamma \approx 10^{-9} \text{ cm}^3$). The parameter γ has only a scaling effect on the distribution of tumor volume.

One can derive different generalizations of the model by assuming that some of its parameters are random. In particular, suppose the quantity $1/\lambda$ (proportional to the tumor doubling time) is gamma distributed with parameters a and b . Note that randomness of λ is reflective of the individual variability of tumor growth rate. Then integrating (4) with respect to the distribution of $1/\lambda$ we have

$$p_x(u, s) = \frac{\alpha b^a}{(\ln s)^{a+1} \Gamma(a)} \int_0^u (u-t)^a \exp\left\{-\frac{b+\alpha(s-1)}{\ln s}(u-t)\right\} p_T(t) dt, \quad (5)$$

for $u \geq 0, s \geq 1$. In this case, the marginal distribution of tumor size is a Pareto distribution [29]. Alternatively, the parameter λ itself can be assumed to follow a gamma distribution.

To complete the model development, we need to describe the impact of screening on the natural history of cancer. Let $\mathcal{T} = \{\tau_1, \tau_2, \dots, \tau_n\}$, where $0 < \tau_1 < \tau_2 < \dots < \tau_n$, be a given screening schedule. It is convenient to set $\tau_0 := 0$ and $\tau_{n+1} := \infty$. Let W_0 be the time of spontaneous detection and W_1 be the time of screening based detection, both times being measured from the disease onset. Then for the time W of combined detection we have: $W = \min(W_0, W_1)$. Then $U = T + W$ and $S = f(W)$ are the age and tumor size at combined detection, respectively.

Introduce a discrete analogue of the conditional (given $T = t$) hazard rate for the screening based detection as

$$\mu_t = \sum_{j=i+1}^n r_j(j) \delta_{\tau_j-t}, \quad \tau_i \leq t < \tau_{i+1}, \quad 0 \leq i \leq n, \quad (6)$$

where δ_x stands for the Dirac measure at x , and the sum over the empty set of indices is set, as usual, to be zero. By definition, the discrete measure μ_t is related to the conditional survival function of r.v. W_1 given $T = t$ through the equation

$$\bar{F}_{w_1|T=t}(w) := \Pr(W_1 > w | T = t) = e^{-\int_0^w d\mu_t}, \quad w \geq 0. \quad (7)$$

Combining (6) and (7) we have

$$\bar{F}_{w_1|T=t}(w) = e^{-\sum_{j=i+1}^k r_j(j)} \quad \tau_i \leq t < \tau_{i+1}, \quad \tau_k \leq t+w < \tau_{k+1}. \quad (8)$$

Equivalently, the process of screen-based detection can be represented through the conditional probabilities $p_t(k)$ of tumor detection at screen k given the disease onset at $T = t$ and provided the previous exams did not detect the tumor. They are related to the conditional discrete hazard rate of r.v. W_1 as follows [30]:

$$p_i(k) = e^{-\sum_{j=i+1}^{k-1} r_i(j)} [1 - e^{-r_i(k)}], \tau_i \leq t < \tau_{i+1}, i+1 \leq k \leq n, 0 \leq i \leq n.$$

The conditional probability $p(t)$ that a tumor was not detected by screening given $T = t$ can be found from (8) by letting $w \rightarrow \infty$:

$$p(t) = e^{-\sum_{j=i+1}^n r_i(j)}. \quad (9)$$

In the spirit of quantal response models [29] one may assume, similarly to spontaneous detection, that the conditional hazard rate of screen-based detection is proportional to the current tumor size:

$$r_i(j) = \alpha_1 S(\tau_j - t), \tau_i \leq t < \tau_{i+1}, 0 \leq i \leq n, i+1 \leq j \leq n, \quad (10)$$

with some constant $\alpha_1 > 0$. Combining (8) and (10) we find that

$$\bar{F}_{w_1|T=t}(w) = e^{-\alpha_1 \sum_{j=i+1}^n f(\tau_j - t)}, \tau_i \leq t < \tau_{i+1}, \tau_k \leq t + w < \tau_{k+1}. \quad (11)$$

It is natural to assume that the r.v.'s W_1 and W_0 are conditionally independent given the onset time T . This is quite plausible in the case of deterministic tumor growth if we hypothesize as above that r.v.'s W_0 and T are independent and adopt the proportionality assumptions (1) and (10).

Proceeding from an arbitrary nonrandom law of tumor growth f as above and assuming independence of r.v.'s W_0 and T as well as conditional independence of r.v.'s W_0 and W_1 given $T = t$ we have for $u \geq 0, s \geq 1$:

$$\begin{aligned} F_{\mathbf{X}}(u, s) &= \Pr(U \leq u, S \leq s) = \Pr(T + W \leq u, f(W) \leq s) \\ &= \int_0^u \bar{F}_{w_1|T=t}(\min\{u - t, g(s)\}) p_T(t) dt = F_T(u) - \int_0^u \bar{F}_{w_1|T=t}(\min\{u - t, g(s)\}) p_T(t) dt \\ &= F_T(u) - \int_0^u \bar{F}_{w_0}(\min\{u - t, g(s)\}) \bar{F}_{w_1|T=t}(\min\{u - t, g(s)\}) p_T(t) dt \\ &= F_T(u) - \bar{F}_{w_0}(g(s)) \int_0^{u-g(s)} \bar{F}_{w_1|T=t}(g(s)) p_T(t) dt \\ &\quad - \int_{u-g(s)}^u \bar{F}_{w_1|T=t}(u - t) \bar{F}_{w_0}(u - t) p_T(t) dt. \end{aligned} \quad (12)$$

Recall that under the proportionality assumptions (1) and (10) one has $\bar{F}_{w_0}(w) = e^{-\alpha \int_0^w f(v) dv}$ and $\bar{F}_{w_1|T=t}$ is given by (11).

A very important fact is theoretical identifiability of the distribution of the random vector $\mathbf{X}$. A proof of this assertion is given in Section 3 with some auxiliary results proven in the Appendix. Unfortunately, no theoretical results on identifiability are available for randomized versions of the model due to prohibitive complexity of the analytic expression for the distribution of random vector $\mathbf{X}$ even in the case of exponential tumor growth with random growth rate, in particular, for the joint distribution given by formula (5). However, our

simulation experiments (and likelihood profiles) have shown that identifiability of the model is preserved if the compounding procedure uses a gamma distribution for λ or its reciprocal.

3. Main Result

Recall that T is the age at tumor onset, and W_0, W_1 denote the times to spontaneous and screen based detection, respectively, measured from onset of the disease. As above, f stands for the law of tumor growth, and $g = f^{-1}$ is the inverse of f . Thus, the r.v. $T + W_1$ takes value τ_k if the tumor was detected by k -th screen, $1 \leq k \leq n$, and $T + W_1 = \infty$ if the tumor was not detected by screening. We keep the notation $p(t) = \Pr(W_1 = \infty | T = t)$, $t \geq 0$, for the conditional probability that the tumor was not detected by screening given the onset at time $T = t$. No specific relationships between the distributions of the r.v.'s W_0, W_1 and the law of tumor growth f are assumed. Functions f, p_T, p_{w_0} and $\bar{F}_{w_1}$ may depend on some parameter sets that may contain common parameters. Finally, $\mathbf{X} = (U, S)$ is the random vector whose components are age and tumor size at combined detection: $U = T + W$ and $S = f(W)$, where $W = \min(W_0, W_1)$. The following two important statements describe the support and density of the distribution of random vector $\mathbf{X}$; their proofs are given in the Appendix.

Proposition 1

Suppose that

1. The r.v.'s T and W_0 are absolutely continuous, proper and independent;
2. For every $t > 0$, the r.v.'s W_0 and W_1 are conditionally independent given $T = t$;
3. $p_T(t) > 0$ for almost all $t > 0$;
4. $p_{W_0}(w) > 0$ for almost all $w > 0$;
5. $p(t) > 0$ for almost all $t > 0$.

Then the (topological) support of the random vector $\mathbf{X}$ is given by

$$E = \{(u, s) : u \geq 0, 1 \leq s \leq f(u)\}.$$

Proposition 2

Suppose that

1. The r.v.'s T and W_0 are absolutely continuous, proper and independent;
2. For every $t > 0$, the r.v.'s W_0 and W_1 are conditionally independent given $T = t$;
3. $f'(t) > 0$ for all $t > 0$;
4. P.d.f.'s p_T and p_{w_0} are continuous on $[0, \infty)$.
5. The function $(t, w) \mapsto \Pr(W_1 > w | T = t)$ is continuous at every point (t, w) such that $t, w \geq 0$ and $t, t+w \notin \mathcal{T}$.

Then, for all (u, s) such that $u \geq 0$, $1 \leq s \leq f(u)$ and $u, u - g(s) \notin \mathcal{T}$,

$$\frac{\partial^2}{\partial s \partial u} F_x(u, s) = \Pr(W_1 > g(s) | T = u - g(s)) p_{w_0}(g(s)) p_T(u - g(s)) g'(s). \quad (13)$$

Observe that the models of tumor latency, growth, spontaneous detection and screening discussed in Section 2 satisfy all conditions of Propositions 1 and 2.

The following theorem represents the main result of the paper.

Theorem

Suppose that all conditions of Propositions 1 and 2 are satisfied. Then the distribution of the random vector $X = (U, S)$ determines the law of tumor growth f and the distributions of r.v.'s T , W_0 and W_1 uniquely. In particular, if f and the distributions of the r.v.'s T , W_0 and W_1 depend on some identifiable (and possibly overlapping) sets of parameters then all these parameters are jointly identifiable from the distribution of the random vector X .

Proof—Suppose that

$$F_X(u, s) = F_{\tilde{X}}(u, s) \text{ for all } u \geq 0, s \geq 1$$

Comparing the supports of the distributions of random vectors X and $\tilde{X}$ and using Proposition 1 we find that the laws of tumor growth are identical: $f = \tilde{f}$. It follows from condition 3 of Proposition 2 that $g'(s) > 0$ for all $s \geq 1$. Therefore, by Proposition 2 we obtain for all $t, w \geq 0$ such that $t, t+w \notin \mathcal{T}$

$$\Pr(W_1 > w | T=t) p_{W_0}(w) p_T(t) = \Pr(\tilde{W}_1 > w | \tilde{T}=t) p_{\tilde{W}_0}(w) p_{\tilde{T}}(t). \quad (14)$$

Notice that if $t > \tau_n$ then, conditionally on $T=t$, we have $W_1 = \infty$. Therefore,

$$\Pr(W_1 > w | T=t) = \Pr(\tilde{W}_1 > w | \tilde{T}=t) = 1, t > \tau_n, w \geq 0.$$

Then we conclude from (14) that

$$p_{W_0}(w) p_T(t) = p_{\tilde{W}_0}(w) p_{\tilde{T}}(t), t > \tau_n, w \geq 0. \quad (15)$$

Integrating (15) in w and taking into account that $\int_0^\infty p_{W_0}(w) dw = \int_0^\infty p_{\tilde{W}_0}(w) dw = 1$ We obtain that

$$p_T(t) = p_{\tilde{T}}(t), t > \tau_n. \quad (16)$$

Then in view of condition 3 of Proposition 1 eq. (15) implies that $p_{W_0} = p_{\tilde{W}_0}$. Together with condition 4 of Proposition 1 this allows us to rewrite formula (14) for almost all $w \geq 0$ in a simpler form

$$\Pr(W_1 > w | T=t) p_T(t) = \Pr(\tilde{W}_1 > w | \tilde{T}=t) p_{\tilde{T}}(t), t \geq 0, t, t+w \notin \mathcal{T}. \quad (17)$$

Furthermore, it follows from condition 5 of Proposition 2 that (17) holds for all $t, w \geq 0$ such that $t, t+w \notin \mathcal{T}$. In particular, setting in (17) $w = 0$ and observing that $\Pr(W_1 > 0 | T=t) = \Pr(\tilde{W}_1 > 0 | \tilde{T}=t) = 1$ for all $t \geq 0, t \notin \mathcal{T}$, we find that

$$p_T(t) = p_{\tilde{T}}(t), 0 \leq t < \tau_n, t \notin \mathcal{T}.$$

Along with (16) this implies that $p_T(t) = p_{\tilde{T}}(t), t \notin \mathcal{T}$, and hence by continuity $p_T = p_{\tilde{T}}$. Therefore, integrating (17) in t we find that $\bar{F}_{W_1} = \bar{F}_{\tilde{W}_1}$. The proof of the theorem is complete.

4. Discussion and Concluding Remarks

The progression of cancer can be characterized by parameters of the initiation, promotion and progression stages of tumor development with the structure of the latter stage being dependent on a specific mechanism of cancer detection. Such natural history of the disease is not only of great scientific interest in its own right but also constitutes a foundation for solving a variety of related biomedical problems. Therefore, estimation of parameters of the natural history of cancer is a task of singular importance. In particular, it is an indispensable step in building parametric models designed to predict potential effects of screening in combination with better methods of cancer treatment. As an alternative to large-scale screening trials, stochastic modeling of cancer screening provides a more cost-effective means of evaluating the efficacy of the existing image-based screening technologies as well as of those that may emerge in the future. Optimization of cancer screening schedules is yet another beneficiary of this approach [30,41].

Certain characteristics (covariates) of cancer patients observed at the time of detection are deemed to contain the information necessary to estimate parameters of the natural history of cancer. In the present paper, our focus has been on two such covariates: age of a patient and tumor size at detection. The joint distribution of the said variables can be derived from mechanistic considerations (Section 2). However, the feasibility of this theoretical framework in applications to data analysis depends critically on identifiability of the parameters incorporated into the resultant two-variate distribution. In this paper, we gave a proof of the desired identifiability, thereby providing the necessary theoretical underpinning for associated methods of statistical inference.

The model considered in this paper has proven to be an efficient means of data analysis. It has enjoyed many applications [29,31,32,40] and has been tested in predictive settings [31]. However, as any other model of a complex biological process, it is necessarily simplistic. When individually scrutinized, each element of the model may be in doubt as based on overly simplistic mathematical assumptions. Such assumptions are unavoidable because the biological processes they are referring to are typically extremely complex and not directly observable. The question arises as to whether the quest for more realism and generality is always warranted. The answer is positive as long as a delicate compromise between flexibility and identifiability of the model can be retained. In other words, every model should be made as realistic as possible, subject to stringent constraints on its tractability and identifiability. Caution should be exercised when attempting to improve a model by increasing its structural complexity and introducing more unknown parameters because model identifiability, and thus its usefulness, may have been lost.

The validity of a model does not necessarily imply that all its assumptions are perfectly valid. The practice of mathematical statistics illustrates this point. Another example is provided by statistical mechanics. The classic kinetic theory of gases is of great practical utility despite the fact that molecules bear little resemblance to absolutely elastic balls and their interactions do not obey the laws of classical mechanics. Strictly speaking, the Markovian character of the two-stage model of carcinogenesis is unrealistic just because the mitotic cycle time is not exponentially distributed. Likewise, the Poissonian nature of the process of initiation may be questioned as well. Furthermore, the latter assumption is untestable given the modern state of the art in experimental biology. And yet, even this simplistic model of carcinogenesis fails to meet the requirement of identifiability and needs to be reparameterized when applied to time-to-tumor observations. The situation would have been only exacerbated if one had attempted to use a more flexible two-parameter distribution for the mitotic cycle duration. Alternative or extended variants of a given model are warranted only when it is in conflict with primary endpoints, i.e., the data this model is intended to describe. Its potential conflict with secondary

or unobservable end-points are of much less concern and cannot serve as a sufficient reason for its disqualification. The main criteria that still remain are a good fit of the model to the data of primary interest and its predictive power.

Yet another relevant example is worth discussing in the context of our model of the natural history of cancer. It is tempting to generalize the model by making the tumor growth rate λ dependent on the age of a patient at the time of tumor onset and thus making r.v.'s T and W_0 dependent. There are indications in the literature [42,43] that primary breast tumors in older women tend to grow at a slower rate than in younger counterparts. We were among those who studied the possibility to extend the model by making the mean value of the growth rate λ dependent on the age of a patient at the time of tumor onset. This relationship was estimated nonparametrically by modulating (scaling) the baseline mean growth rate by a piecewise constant function and fitting the extended model to age-specific incidence in several birth cohorts identified through the the NCI Surveillance, Epidemiology and End Results (SEER) database. Our analysis suggested that the mean growth rate attains its maximum when the age at tumor onset is about 50 years and then drops sharply. At the same time, it is important to note that the scaling procedure had virtually no effect on the other parameters incorporated into the model, thereby suggesting that the observed effect is too weak to be of significance for parameter estimation. This is consistent with a recent cohort study [40] showing no tangible birth cohort effect on the nonparametrically estimated distribution of tumor size at detection. If the tumor growth rate were strongly dependent on the age at tumor onset, one would expect the birth cohort effect to manifest itself in the distribution of tumor size at diagnosis. Therefore, this extension of the model appears to be unnecessary when the primary focus is on the estimation of model parameters from the joint distribution of age and tumor size at the time of detection. The same argument applies equally to the sensitivity parameter α_1 . Some authors [44] suggest that the sensitivity of mammography increases with age. However, the tendency does not appear to be strong enough to manifest itself in the type of data we have dealt with in various applications [31,40]. For a model to be useful, it should be parsimonious in that its complexity must be adequate for the information contained in the data to be analyzed.

Threshold models of cancer detection represent the most natural alternative to quantal response models. Such models typically assume a random threshold of tumor mass attaining which a tumor becomes detectable. When fitting the number of prevalent and screen-detected cases together with tumor size measurements obtained from the CNBSS data, we noticed that the postulated dependence of the discrete hazard rate for the events of screen-based detection on tumor size is a little too rigid, hampering our attempts to improve the fit to all the endpoints simultaneously. The challenge is to make the model more flexible without introducing many additional parameters so that its identifiability can be preserved. It seems likely that this can be accomplished through incorporating a threshold mechanism of tumor detection for the screening component of the model, while retaining the quantal response mechanism for clinical detection. We studied the class of randomized threshold models in the context of clinical detection [45-47]. Accommodating arbitrary screening schedules in this class of models would be an interesting avenue for future research. Some preliminary results [28] lead us to expect the problem to be approachable.

We believe that improved models of cancer screening can be developed in the future through including more components (in addition to age and tumor size) in the vector of clinical covariates accessible to measurement at the time of diagnosis; see [28] for the general idea and associated analytical techniques. However, proving identifiability of such models would represent a very serious challenge.

Appendix

1. Proof of Proposition 1

Since the r.v.'s T and W_0 are independent, proper and absolutely continuous, the r.v. $U_0 = T + W_0$ (age at spontaneous detection) is proper and absolutely continuous. Also, for the r.v. $U_1 = T + W_1$ (age at screen based detection) we have $U_1 \in \mathcal{T} \cup \{\infty\}$. Therefore, $\Pr(W_1 = W_0) = \Pr(T + W_0 = T + W_1) \leq \Pr(U_0 \in \mathcal{T} \cup \{\infty\}) = 0$. This implies

$$\Pr(W_1 = W_0) = 0. \quad (\text{A1})$$

From $U = T + W \geq 0$ and $1 \leq S = f(W) \leq f(U)$ it follows that $\Pr(X \in E) = 1$. Therefore, since the set E is closed, the support of the random vector X is contained in E . To show the reverse inclusion, it suffices to prove that, for every open rectangle R contained in E , we have $\Pr(X \in R) > 0$.

Let $R = (a, b) \times (c, d)$, where $0 \leq a < b$ and $1 \leq c < d \leq f(a)$. Denote $\alpha = g(c)$ and $\beta = g(d)$, then $0 \leq \alpha < \beta \leq a$. By taking a smaller subinterval of (a, b) we may (and will) assume that $T \cap (a, b) = \emptyset$. Then $\Pr(T + W_1 \in (a, b)) = 0$, and hence also

$$\Pr(T + W \in (a, b), W = W_1) = 0. \quad (\text{A2})$$

In view of (A1) and (A2) we have

$$\begin{aligned} \Pr(X \in R) &= \Pr(U \in (a, b), S \in (c, d)) = \Pr(T + W \in (a, b), W \in (\alpha, \beta)) \\ &= \Pr(T + W \in (a, b), W \in (\alpha, \beta), W = W_0) + \Pr(T + W \in (a, b), W \in (\alpha, \beta), W = W_1) \\ &= \Pr(T + W \in (a, b), W \in (\alpha, \beta), W = W_0) = \Pr(T + W_0 \in (a, b), W_0 \in (\alpha, \beta), W_0 < W_1) \\ &\geq \Pr(T + W_0 \in (a, b), W_0 \in (\alpha, \beta), W_1 = \infty). \end{aligned}$$

By conditioning on the r.v. T and using conditions 1 and 2 of Proposition 1 we continue the above chain of relations as follows:

$$\begin{aligned} \Pr(X \in R) &\geq \int_0^\infty \Pr(W_0 \in (a - t, b - t) \cap (\alpha, \beta)) p(t) p_T(t) dt \\ &= \int_{a-\beta}^{b-\alpha} \Pr(W_0 \in (a - t, b - t) \cap (\alpha, \beta)) p(t) p_T(t) dt, \end{aligned}$$

where we used the fact that $(a - t, b - t) \cap (\alpha, \beta) \neq \emptyset$ if and only if $t \in (a - \beta, b - \alpha)$. Finally, taking into account conditions 3-5 of Proposition 1 we conclude that $\Pr(X \in R) > 0$.

2. Proof of Proposition 2

According to formula (12), that was derived on the basis of assumptions 1 and 2 of Proposition 2, we have

$$\begin{aligned} F_X(u, s) &= F_T(u) - \bar{F}_{W_0}(g(s)) \int_0^{u-g(s)} \Pr(W_1 > g(s) | T=t) p_T(t) dt \\ &\quad - \int_{u-g(s)}^u \Pr(W_1 > u - t | T=t) \bar{F}_{W_0}(u - t) p_T(t) dt. \end{aligned} \quad (\text{A3})$$

Recall the formula

$$\frac{d}{du} \int_{\varphi(u)}^{\psi(u)} \Phi(t, u) dt = \psi'(u) \Phi(\psi(u), u) - \varphi'(u) \Phi(\varphi(u), u) + \int_{\varphi(u)}^{\psi(u)} \frac{\partial}{\partial u} \Phi(t, u) dt, \quad (\text{A4})$$

where, for a given point u of interest, functions ϕ and ψ are assumed to be differentiable at the point u , and function Φ is assumed (1) to be integrable in t on $(\phi(v), \psi(v))$, where v belongs to some neighborhood of u ; (2) to be jointly continuous at points $(\phi(u), u)$ and $(\psi(u), u)$; and (3) to have, for almost all $t \in (\phi(u), \psi(u))$ and for all v in some neighborhood of u , a uniformly bounded partial derivative $(\partial\Phi/\partial u)(t, v)$.

Since $W_1 + T \in \mathcal{T}$, the admissible values of W_1 , conditional on $T = t$, $\tau_i < t < \tau_{i+1}$, $0 \leq i \leq n$, are $\tau_{i+1} - t, \dots, \tau_n - t, \infty$ only. Hence, for each $t \notin \mathcal{T}$, the function

$$G(t, u) := \Pr(W_1 > u - t | T = t), u \geq t \geq 0,$$

is constant in a neighborhood of every point $u \notin \mathcal{T}$. Therefore,

$$\frac{\partial G}{\partial u}(t, v) = 0, t, v \notin \mathcal{T}.$$

Then, differentiating (A3) in variable u and observing that, due to conditions 4 and 5 of Proposition 2, formula (A4) is applicable to both integrals in (A3) we obtain after some cancellations that

$$\frac{\partial F_x}{\partial u}(u, s) = \int_{u-g(s)}^u \Pr(W_1 > u - t | T = t) p_{w_0}(u - t) p_T(t) dt, u, u - g(s) \notin \mathcal{T}.$$

Finally, differentiating this equality in variable s with conditions 4 and 5 of Proposition 2 taken into account leads to the required formula (13).

Acknowledgements

This research was supported by NIH/NCI grant U01 CA88177 and by Collaborative Linkage grant PST CLG 979045 of the NATO Science Programme. We are grateful to the anonymous reviewers for their helpful comments.

References

1. Hanin L. Identification problem for stochastic models with application to carcinogenesis, cancer detection and radiation biology. *Discrete Dyn Nat Soc* 2002;7:177.
2. Albert A, Gertman PM, Louis TA. Screening for the early detection of cancer. I. The temporal history of a progressive disease state. *Math Biosci* 1978;40:1.
3. Albert A, Gertman PM, Louis TA, Liu SI. Screening for the early detection of cancer. II. The impact of the screening on the natural history of the disease. *Math Biosci* 1978;40:61.
4. Blumenson LE. Detection of disease with periodic screening: transient analysis and application to mammography examination. *Math Biosci* 1977;33:73.
5. Blumenson LE, Bross IDJ. A mathematical analysis of the growth and spread of breast cancer. *Biometrics* 1969;22:95. [PubMed: 5786312]
6. Brookmeyer R, Day N. Two-stage model for the analysis of cancer screening data. *Biometrics* 1987;43:657. [PubMed: 3663822]
7. Day NE, Walter SD. Simplified models of screening for chronic disease. Estimation procedures from mass screening programs. *Biometrics* 1984;40:1. [PubMed: 6733223]
8. Duffy SW, Chen HH, Tabar L, Day NE. 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. *Stat Med* 1995;14:1531. [PubMed: 7481190]

9. Flehinger, BJ.; Kimmel, M. Screening for cancer in relation to the natural history of the disease. In: Axelrod, DE.; Kimmel, M., editors. *Mathematical Population Dynamics*. Marcel Dekker; New York: 1991. p. 383
10. Hu P, Zelen M. Planning clinical trials to evaluate early detection programmes. *Biometrika* 1997;84:817.
11. Lee SJ, Zelen M. Scheduling periodic examinations for the early detection of disease: applications to breast cancer. *J Am Stat Assoc* 1998;93:1271.
12. Louis T, Albert A, Heghinian S. Screening for the early detection of cancer. III. Estimation of disease natural history. *Math Biosci* 1978;40:111.
13. Parmigiani G. On optimal screening ages. *J Am Stat Ass* 1993;88:622.
14. Parmigiani G. Timing medical examinations via intensity functions. *Biometrika* 1997;84:803.
15. Pinsky PF. Estimation and prediction for cancer screening models using deconvolution and smoothing. *Biometrics* 2001;57:389. [PubMed: 11414561]
16. Prorok PC. The theory of periodic screening I: Lead time and proportion detected. *Adv Appl Probab* 1976;8:127.
17. Prorok PC. The theory of periodic screening II: Doubly bounded recurrence times and mean lead time and detection probability estimation. *Adv Appl Probab* 1976;8:460.
18. Shen Y, Zelen M. Parametric estimation procedures for screening programmes: stable and nonstable disease models for multimodality case finding. *Biometrika* 1999;86:503.
19. Shen Y, Zelen M. Screening sensitivity and sojourn time from breast cancer early detection clinical trials: mammograms and physical examinations. *J Clin Oncol* 2001;19:3490. [PubMed: 11481355]
20. Schwartz M. An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease. *Cancer* 1978;41:1550. [PubMed: 639011]
21. Schwartz M. A mathematical model used to analyse breast cancer screening strategies. *Oper Res* 1978;26:937.
22. Tsodikov AD, Yakovlev AY. On the optimal policies of cancer screening. *Math Biosci* 1991;107:21. [PubMed: 1806106]
23. Tsodikov AD, Asselain B, Fourquet A, Hoang T, Yakovlev AY. Discrete strategies of cancer post-treatment surveillance. Estimation and optimization problems. *Biometrics* 1995;51:437. [PubMed: 7662836]
24. Walter SD, Day NE. Estimation of the duration of a preclinical state using screening data. *Am J Epidemiol* 1983;118:865. [PubMed: 6650488]
25. Yakovlev, AY.; Tsodikov, AD. *Stochastic Models of Tumor Latency and their Biostatistical Applications*. World Scientific Publ; Singapore: 1996.
26. Zelen M. Optimal scheduling of examinations for the early detection of disease. *Biometrika* 1993;80:279.
27. Zelen M, Feinleib M. On the theory of screening for chronic disease. *Biometrika* 1969;56:601.
28. Hanin LG, Yakovlev AY. Multivariate distributions of clinical covariates at the time of cancer detection. *Stat Methods Med Res* 2004;13:457. [PubMed: 15587434]
29. Bartoszyński R, Edler L, Hanin L, Kopp-Schneider A, Pavlova L, Tsodikov A, Zorin A, Yakovlev A. Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis. *Math Biosci* 2001;171:113. [PubMed: 11395047]
30. Hanin LG, Tsodikov AD, Yakovlev AY. Optimal schedules of cancer surveillance and tumor size at detection. *Math Comp Modell* 2001;33:1419.
31. Hanin, L.; Miller, A.; Zorin, A.; Yakovlev, A. The University of Rochester model of breast cancer detection and survival; The National Cancer Institute Monograph Series; to appear
32. Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, Mandelblatt J, Yakovlev AY, Habbema JDF, Feuer EJ. Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine* 2005;353:1784. [PubMed: 16251534]
33. Moolgavkar SH, Venzon DJ. Two event model for carcinogenesis: Incidence curves for childhood and adult tumors. *Math Biosci* 1979;47:55.
34. Moolgavkar SH, Knudson AG. Mutation and cancer: a model for human carcinogenesis. *J Natl Cancer Inst* 1981;66:1037. [PubMed: 6941039]

35. Moolgavkar SH, Luebeck EG. Two-event model for carcinogenesis: Biological, mathematical and statistical considerations. *Risk Anal* 1990;10:323. [PubMed: 2195604]
36. Heidenreich WF. On the parameters of the clonal expansion model. *Radiat Environ Biophys* 1996;35:127. [PubMed: 8792461]
37. Hanin LG, Yakovlev AY. A nonidentifiability aspect of the two-stage model of carcinogenesis. *Risk Anal* 1996;16:711. [PubMed: 8962520]
38. Heidenreich WF, Luebeck EG, Moolgavkar SH. Some properties of the hazard function of the two-mutation clonal expansion model. *Risk Anal* 1997;17:391. [PubMed: 9232020]
39. Gregori G, Hanin L, Luebeck G, Moolgavkar S, Yakovlev A. Testing goodness of fit with stochastic models of carcinogenesis. *Math Biosci* 2001;175:13. [PubMed: 11779625]
40. Zorin, AV.; Edler, L.; Hanin, LG.; Yakovlev, AY. Estimating the natural history of breast cancer from bivariate data on age and tumor size at diagnosis. In: Edler, L.; Kitsos, CP., editors. *Quantitative Methods for Cancer and Human Health Risk Assessment*. Wiley; New York: p. 317-327.
41. Hanin, LG.; Pavlova, LV. Optimal regimens of cancer screening. In: Edler, L.; Kitsos, CP., editors. *Quantitative Methods for Cancer and Human Health Risk Assessment*. Wiley; New York: p. 177-191.
42. van Oortmarssen GJ, Habbema JD, Lubbe JT, van der Maas PJ. A model-based analysis of the HIP project for breast cancer screening. *Int J Cancer* 1990;46:207. [PubMed: 2384271]
43. Peer PG, van Dijck JA, Hendriks JH, Holland R, Verbeek AL. Age-dependent growth rate of primary breast cancer. *Cancer* 1993;71:3547. [PubMed: 8490903]
44. Mandelblatt J, Saha S, Teutsch S, Hoerger T, Siu AL, Atkins D, Klein J, Helfand M. A systematic review for the U.S. Preventive Services Task Force: The cost-effectiveness of screening mammography beyond age 65. *Ann Intern Med* 2003;139:835. [PubMed: 14623621]
45. Yakovlev AY, Hanin LG, Rachev ST, Tsodikov AD. A distribution of tumor size at detection and its limiting form. *Proc Nat Acad Sci USA* 1996;93:6671. [PubMed: 8692876]
46. Hanin LG, Rachev ST, Tsodikov AD, Yakovlev AY. A stochastic model of carcinogenesis and tumor size at detection. *Adv Appl Prob* 1997;29:607.
47. Tsodikov AD, Asselain B, Yakovlev AY. A distribution of tumor size at detection: an application to breast cancer data. *Biometrics* 1997;53:1495. [PubMed: 9423264]