Discrete Dynamics in Nature and Society, 2002 VOL. 7 (3), pp. 177–189

Taylor & Francis
Taylor & Francis Group

Identification Problem for Stochastic Models with Application to Carcinogenesis, Cancer Detection and Radiation Biology

L.G. HANIN

Department of Mathematics, Idaho State University and Huntsman Cancer Institute of the University of Utah, Idaho State University, Pocatello, ID 83209-8085, USA

(Received 16 June 2001)

A general framework for solving identification problem for a broad class of deterministic and stochastic models is discussed. This methodology allows for a unified approach to studying identifiability of various stochastic models arising in biology and medicine including models of spontaneous and induced carcinogenesis, tumor progression and detection, and randomized hit and target models of irradiated cell survival. A variety of known results on parameter identification for stochastic models is reviewed and several new results are presented with an emphasis on rigorous mathematical development.

Keywords: Cancer detection; Carcinogenesis; Difference equation; Hazard function; Hit and target model of irradiated cell survival; Identification problem

INTRODUCTION

In a very general form, the identification problem for a model describing behavior of a physical, chemical, biological or other system can be formulated as follows. Suppose a system is governed by the equation

$$y = f(x, \theta), \tag{1}$$

where x is a scalar or vector variable descriptive of the state of the system (e.g. time, vector of spatial coordinates, etc.), y is an *observable* scalar of vector quantity characterizing the system "output," θ is a (typically unknown and unobservable) parameter or set of parameters labeling a specific system within a class of similar systems, and f is a *known* function that relates the value x of the descriptive variable of the system to its output y. Elements of the parameter set θ are usually represented by real numbers, functions or probability distributions. It will be assumed that the parameter set θ is *minimal* in that none of its elements is a function of other elements.

We assume that the state variable x takes values in a given set $\mathscr{D}$ which is typically independent of θ. For deterministic models, Eq. (1) can be obtained by solving a differential, difference, integral or other equation that governs behavior of the system. For stochastic models, Eq. (1) has the form $Y = f(X, \theta)$, where X and Y are random variables (RVs), random vectors, stochastic processes or random fields. In most cases, observed is the *distribution* of Y rather than its particular realization. The distribution of RV Y is completely characterized by either the *cumulative distribution function* (CDF) $F_Y(y) := \Pr(Y \le y)$ or the corresponding *survival function* $\bar{F}_Y(y) := 1 - F_Y(y) = Pr(Y > y)$. If RV Y is absolutely continuous then its distribution is also uniquely determined by its *probability density function* (PDF) f_Y or (in the case of non-vanishing survival function) by the *hazard function* φ_Y.

In stochastic models described in this paper, the observed output Y is a non-negative RV in which case Eq. (1) takes on the form

$$\bar{F}_Y(t) = f(t, \theta), \quad t \ge 0, \tag{2}$$

where $\bar{F}_Y$ can be replaced by whatever characteristic of the distribution of RV Y (CDF, PDF, hazard function, characteristic function, etc.) that is appropriate and for which Eq. (2) assumes the simplest form. Observe also that survival and hazard functions of such RV Y are related as follows:

$$\bar{F}_Y(t) = \exp\left(-\int_0^t \varphi_Y(u)\mathrm{d}u\right), \quad t \ge 0. \tag{3}$$

DEFINITION 1 The model (1) is said to be *identifiable* if

$$f(x, \theta_1) = f(x, \theta_2) \quad \text{for all} \quad x \in \mathscr{D} \quad \text{implies} \quad \theta_1 = \theta_2. \tag{4}$$

ISSN 1026-0226

Property (4) suggests that the same output (or its distribution if the output is random) cannot arise from two different sets of model parameters. Identifiability of a model leads to a possibility of determining the unique set θ of model parameters from the observed output. However, it does not generally guarantee stability of this procedure with respect to perturbations of the output which may become an important consideration when the output is incomplete, truncated, censored, noised or measured with inherently limited accuracy. A non-identifiable model can still be useful for the purpose of qualitative analysis of the system in question and making general conclusions about its behavior, but its utility for quantitative inference about parameters of the system from its observed output is limited.

The output of stochastic models is typically observed in a form of a finite sample from the distribution of RV (random vector, etc.) Y. The empirical distribution of such sample can be thought of as an approximation to the "true" distribution of Y. Then identifiability of the model makes it plausible that parameters of the model can be estimated from the sample observations using standard statistical methods (maximum likelihood, moment, Bayesian, etc.). It should be emphasized that identifiability is a structural property of the model that has analytic nature. If met, it removes the crudest obstruction to determining or estimating model parameters but cannot automatically ensure any properties of statistical estimators such as consistency or asymptotic normality. On the contrary, non-identifiability suggests that the original model parameters cannot even in principle be estimated from output observations. Practically, lack of identifiability manifests itself in the instability of statistical estimators of model parameters, should these estimators be constructed formally with no regard to model non-identifiability.

Let Θ be the collection of all admissible parameter sets θ for the model (1). The choice of model parameter θ is certainly a matter of convenience. The same model can also be parameterized by any collection Ξ that is a bijective image of Θ, which leads to an equivalent model $y = g(x, \xi)$, $\xi \in \Xi$. Indeed, the latter model and Eq. (1) are identifiable or not simultaneously. However, a convenient choice of model parameters can make a model more tractable and its identifiability properties more transparent.

The present paper deals with identification of several models arising in carcinogenesis, oncology, and radiation biology. Before going into analysis of these complex stochastic models, it seems worthwhile to discuss a simple deterministic physical model in which non-identifiability can be seen quite easily.

Consider vertical oscillations of a body with mass m attached to a spring with spring constant k. Assuming that the mass of the spring is small as compared to m and air resistance is negligible, we describe the motion of the body by means of the differential equation

$$my'' + ky = 0,$$

where $y = y(t)$ is the vertical position of the body at time t relative to the point of static equilibrium. The general solution of this equation is

$$y(t) = C_1 \cos \omega t + C_2 \sin \omega t, \tag{5}$$

where

$$\omega = \sqrt{\frac{k}{m}} \tag{6}$$

and constants C_1, C_2 are uniquely determined by the initial position $y_0 = y(0)$ and initial velocity $v_0 = y'(0)$ of the body (specifically, $C_1 = y_0$ and $C_2 = v_0/\omega$). The "natural" parameter set for the system in question is $\theta = (m, k, y_0, v_0)$. Could this set be identified from the observed motion (Eq. (5))? Clearly, the answer is NO, because it is only y_0, v_0 and the frequency ω that can be determined by the output $y(t)$, where ω is the combination (Eq. (6)) of the model parameters m, k, so that each value ω corresponds to infinitely many parameter sets θ. Thus, the model is not identifiable.

The structure of the paper is as follows. In the second section, we discuss a general methodology for solving the identifiability problem for finite parametric models. This analysis leads us to formulation of main questions of interest with reference to model identifiability and sets up a stage for the ensuing study of identifiability properties of various stochastic models arising in biology and medicine. In the third and fourth sections, we introduce Moolgavkar–Venzon–Knudson and Yakovlev–Polig models of carcinogenesis, respectively, and discuss at length their identifiability. Some general observations related to identification of stochastic models providing the distribution of the total duration of two-stage processes of any nature are presented in the fifth section. Continuing this line of reasoning in the sixth section, we further incorporate into our analysis, in addition to the stage of tumor latency discussed in the third and fourth sections, also the stage of tumor progression and stochastic models of cancer detection. Our main attention there is focused on identifiability of the joint distribution of age and tumor size at spontaneous detection. Finally, the seventh section deals with identification properties of randomized hit and target models of irradiated cell survival.

IDENTIFICATION OF FINITE PARAMETRIC MODELS

In this section, we make a few general observations regarding identification properties of models (1) or (2) depending on a *finite* set of parameters $\theta = (\theta_1, \ldots, \theta_n)$. Studying such properties begins with solving the equation $f(x, \theta) = g(x)$, $x \in \mathscr{D}$, for θ, where g is a given output function, in an attempt to identify all possible independent combinations of parameters $\theta_1, \ldots, \theta_n$ that are determined by the function g. This typically results in equations of the

form

$$\varphi_i(\theta_1, \ldots, \theta_n) = L_i(g), \quad 1 \le i \le m, \tag{7}$$

with some functions $\varphi_i : \Theta \to \mathbb{R}$ and functionals $L_i, 1 \le i \le m$, defined on the set of all admissible outputs of the model. The functionals L_i may involve values or limits of the function g or its derivatives at certain points or at infinity, plain or weighted integrals of function g and its derivatives as well as similar characteristics of Fourier, Laplace or other appropriate transforms of the function g, etc. Independence of Eq. (7) usually implies that $m \le n$, for, should $m > n$, then one of the Eq. (7) would normally follow from other equations and hence could be eliminated. Thus we will assume that $m \le n$. The set of Eq. (7) should be not only *minimal*, but also *complete* so that $f(x, \theta) = f(x, \theta')$ for all $x \in \mathcal{D}$ is equivalent to $\varphi_i(\theta) = \varphi_i(\theta')$, $1 \le i \le m$.

This means that parameters

$$\eta_i := \varphi_i(\theta_1, \ldots, \theta_n), \quad 1 \le i \le m, \tag{8}$$

are identifiable. Furthermore, the model can be expressed in terms of the combinations $\eta_1, \ldots, \eta_m$ of parameters $\theta_1, \ldots, \theta_n$: there exists a function h such that $f(x, \theta) = h(x, \eta)$ for all $x \in \mathcal{D}$, where $\eta = (\eta_1, \ldots, \eta_m)$. If $m = n$ then system of Eq. (7) typically has a unique solution θ which entails identifiability of the model. In the case $m < n$, parameters $\theta_1, \ldots, \theta_n$ are generally not determined uniquely by the output function g rendering the model non-identifiable. The number m, which is usually invariant under the choice of different ways to set up Eq. (7), can be referred to as the *parametric dimension of the model.*

The following problems constitute natural milestones in studying model identifiability.

(1) To find parametric dimension of the model.
(2) To obtain the most natural identifiable combinations (Eq. (8)) of the model parameters, find their joint range, and express the model in terms of these new parameters.
(3) To study relations (8) that provide a mathematical insight into the cause of model non-identifiability; in particular, to express, if necessary, the original set of parameters θ through the new parameter vector η and $n - m$ free parameters.

For example, the parametric dimension of the model given by Eqs. (5) and (6) is three, the most natural identifiable combinations of the model parameters m, k, y_0, v_0 are $\omega = \sqrt{k/m}$, y_0, v_0, so that the rescaling transformation $k \to \lambda k$, $m \to \lambda m$ for any $\lambda > 0$ does not change the relation between the descriptive variable t and the output $y(t)$ of the model.

It should be kept in mind that although dealing with the identifiable form $y = h(x, \eta)$ of a non-identifiable model (1) has a distinct mathematical advantage, parameters η need not necessarily have a straightforward interpretation, by contrast to the original parameters θ that usually have a readily available "mechanistic" connotation.

NON-IDENTIFIABILITY OF THE MOOLGAVKAR–VENZON–KNUDSON TWO-STAGE MODEL OF CARCINOGENESIS

The most widely accepted mechanistic model of carcinogenesis is usually referred to as the Moolgavkar–Venzon–Knudson (MVK) model (Moolgavkar and Venzon, 1979; Moolgavkar and Knudson, 1981). This Markovian model has had a profound impact in carcinogenesis modeling and quantitative analysis of various experimental data, see e.g. Heidenreich *et al.* (1997) and references therein. In this and many other mechanistic models of spontaneous and induced carcinogenesis, formation of malignant cells is viewed as consisting of two stages: (1) induction of primary precancerous lesions in the population of susceptible target normal stem cells (cells bearing primary lesions are called *initiated*); and (2) *promotion* of initiated cells resulting in the transformation of *intermediate* cells (i.e. initiated cells and their offspring) into malignant cells in the course of cell division. The two-stage theory was coined by Armitage and Doll (1957). The MVK model is based on the following commonly accepted assumptions:

(1) The number of first-generation initiated cells follows a (generally, non-homogeneous) Poisson process with intensity $\nu(t)$.
(2) An intermediate cell divides into two intermediate daughter cells with rate $\alpha(t)$ (in the sense of Markov processes), dies or differentiates with rate $\beta(t)$, and divides into one intermediate and one malignant cell with rate $\mu(t)$. The usual independence hypotheses for the birth-and-death branching Markov process are accepted, see e.g. Karlin (1966).
(3) Tumors arise from a single malignant progenitor cell. Once a malignant cell is generated, its subsequent growth is irreversible and leads to appearance of a detectable tumor.

As shown by Hanin and Yakovlev (1996) and Yakovlev and Polig (1996), under assumption (2) and for arbitrary promotion time distribution with CDF F, formation of clonogenic tumor cells is governed by a Poisson process with the integral rate

$$\Lambda(t) = \int_0^t \nu(x)F(t - x)\mathrm{d}x.$$

Then the general form of the survival function $\bar{G}(t)$ of the time to tumor, that is, the probability of no malignant clonogenic cells at time t, is given by

$$\bar{G}(t) = \exp\left\{ -\int_0^t \nu(x)F(t - x)\mathrm{d}x \right\}. \tag{9}$$

In the particular case of constant initiation rate ν, this formula takes on the form

$$\bar{G}(t) = \exp\left\{-\nu\int_0^t F(x)\mathrm{d}x\right\}, \tag{10}$$

compare with (3).

Observe that in the case of spontaneous carcinogenesis, time to tumor (also referred to as time of *tumor latency*) is counted from the moment of birth while for induced carcinogenesis, the time is measured from the initial moment of exposure to a carcinogen.

Model (9) and its particular case (10) was widely used to describe induced and spontaneous carcinogenesis (Klebanov *et al.*, 1993; Yakovlev *et al.*, 1996; Hanin *et al.*, 1997), and hormesis (Yakovlev *et al.*, 1993).

When the rate of initiation (ν) and the rates of cell division (α), death or differentiation (β), and malignant transformation (μ) for initiated cells are all constant, and the number of target normal cells is effectively constant, the following explicit formula for the CDF F of the promotion time was obtained by Kopp-Schneider *et al.* (1994) and Zheng (1994):

$$F(t) = \frac{(\alpha-\beta-\mu+c)(\beta+\mu+c-\alpha)(1-\mathrm{e}^{-ct})}{2\alpha[(\alpha-\beta-\mu+c)\mathrm{e}^{-ct}+(\beta+\mu+c-\alpha)]}, \tag{11}$$

where

$$c = \sqrt{(\alpha+\beta+\mu)^2-4\alpha\beta}. \tag{12}$$

Observe that

$$\lim_{t\to\infty} F(t) = \frac{\alpha-\beta-\mu+c}{2\alpha} < 1.$$

This reflects the fact that the probability of the event that no malignant clonogenic cells are produced is positive. In the case when the parameters ν, α, β and μ are piecewise constant on the same arbitrary time intervals, recursive formulas for computing the hazard function of the time of tumor latency were found by Moolgavkar and Luebeck (1990), see also Heidenreich *et al.* (1997).

It was initially pointed out by Heidenreich (1996) and subsequently by Hanin and Yakovlev (1996) and Heidenreich *et al.* (1997) that the four parameters ν, α, β, μ of the MVK model (which may be constants or, more generally, functions of time) are not jointly identifiable from time-to-tumor data alone. This explained the failure of attempts by several researchers over a substantial period of time to estimate the four parameters of the MVK model from time-to-tumor data. A rigorous treatment of the identifiability problem for the MVK model with constant parameters along the lines indicated in the previous section was first implemented by Hanin and Yakovlev (1996), and is presented below.

As found by Kopp-Schneider *et al.* (1994) and Zheng (1994) (or can be obtained on the basis of formulae (10) and (11)), the survival function of the time of tumor latency in the MVK model with constant parameters ν, α, β, μ is

$$\bar{G}(t) = \left[\frac{2c\mathrm{e}^{-(\alpha-\beta-\mu+c)t/2}}{(\alpha-\beta-\mu+c)\mathrm{e}^{-ct}+(\beta+\mu+c-\alpha)}\right]^{\nu/\alpha}, \tag{13}$$

where c is specified in Eq. (12). To show dependence of the function $\bar{G}$ on the four biologically motivated parameters notationally, we will write $\bar{G}(t) = \bar{G}(t;\nu,\alpha,\beta,\mu)$. The following result was obtained by Hanin and Yakovlev (1996).

THEOREM 1 *The equality* $\bar{G}(t;\nu_1,\alpha_1,\beta_1,\mu_1) = \bar{G}(t;\nu_2,\alpha_2,\beta_2,\mu_2)$, $t \geq 0$, *holds if and only if*

$$\frac{\nu_1}{\alpha_1} = \frac{\nu_2}{\alpha_2},$$

$$\alpha_1-\beta_1-\mu_1 = \alpha_2-\beta_2-\mu_2,$$

and

$$(\alpha_1+\beta_1+\mu_1)^2-4\alpha_1\beta_1 = (\alpha_2+\beta_2+\mu_2)^2-4\alpha_2\beta_2.$$

It follows that parameters

$$\rho := \nu/\alpha, \quad \delta := \alpha-\beta-\mu \quad \text{and} \quad c = \sqrt{(\alpha+\beta+\mu)^2-4\alpha\beta} \tag{14}$$

are identifiable (note that δ can be interpreted as the effective birth rate). Since they are clearly independent, the parametric dimension of the MVK model is three. The range of the parameter vector $\eta = (\rho,\delta,c)$ is given by

$$\mathcal{N} = \{(\rho,\delta,c) : \rho > 0, \quad \delta \in \mathbb{R}, \quad c > |\delta|\}.$$

Conversely, given any $\eta \in \mathcal{N}$, we pick an *arbitrary* $\alpha > (c-\delta)/2$ and set

$$\nu := \alpha\rho, \quad \beta := \frac{(2\alpha-\delta)^2-c^2}{4\alpha}, \quad \mu := \frac{c^2-\delta^2}{4\alpha}$$

to find that ν, α, β, $\mu > 0$ and relations (14) are satisfied. Thus, there is infinitely many parameter sets $\theta = (\nu,\alpha,\beta,\mu)$ corresponding to each parameter vector η and hence to each survival function (13).

Another way to visualize the set of all model parameters θ determined by the function $\bar{G}(t)$ given one of them, $\theta_0 = (\nu_0,\alpha_0,\beta_0,\mu_0)$, is as follows. Setting $\nu = \lambda\nu_0$, $\lambda > 0$, we solve the equations for the two parameter sets given by Theorem 1 to find that

$$\nu = \lambda\nu_0, \quad \alpha = \lambda\alpha_0,$$

$$\beta = (\lambda-1)\alpha_0+\beta_0+\frac{\lambda-1}{\lambda}\mu_0, \quad \mu = \frac{\mu_0}{\lambda}. \tag{15}$$

The claim $\beta > 0$ leads to the restriction $\lambda > \lambda_0$, where

$$\lambda_0 = \frac{[(\alpha_0 - \beta_0 - \mu_0)^2 + 4\alpha_0\mu_0]^{1/2} + (\alpha_0 - \beta_0 - \mu_0)}{2\alpha_0}$$

(note that $0 < \lambda_0 < 1$). Therefore, together with θ_0, the whole curve (Eq. (15)) of parameter vectors θ with $\lambda > \lambda_0$ pertains to the same function $\bar{G}(t)$. This explains numerical findings displayed in Table 1 of Heidenreich (1996).

In terms of parameters ρ, δ, c the function $\bar{G}(t)$ takes on a simpler form

$$\bar{G}(t) = \left[\frac{2ce^{-(c+\delta)t/2}}{(c+\delta)e^{-ct} + c - \delta}\right]^{\rho},$$

compare with Eq. (13). To further simplify it, we introduce a new set of identifiable parameters (a, b, ρ), where $a := (c - \delta)/2$ and $b := (c + \delta)/2$. Clearly, $a, b > 0$, and

$$\bar{G}(t) = \left[\frac{(a+b)e^{at}}{b + ae^{(a+b)t}}\right]^{\rho}, \tag{16}$$

while for the corresponding hazard function we have

$$h(t) = \rho ab \frac{e^{(a+b)t} - 1}{b + ae^{(a+b)t}}. \tag{17}$$

Comparison of Eq. (17) with Eq. (15) by Heidenreich (1996) yields the following relation between parameters X_m, γ, q introduced by Heidenreich (1996) and parameters a, b, ρ : $X_m = \rho ab$, $\gamma = b - a$, $q = a$. Although it was shown by Heidenreich (1996) that the four rates in the MVK model are not jointly identifiable, it is not proven rigorously there that parameters X_m, γ, q are identifiable from the time-to-tumor distribution, i.e. that the parametric dimension of the MVK model with constant parameters is exactly three. Finally, a clear distinction should be drawn (which was not done by Heidenreich, 1996) between *model identifiability* and *goodness of fit* it provides to the observed data (sample of tumor latency times in the case of MVK model). The goodness of fit depends solely on model adequacy and statistical properties of sampling. In contrast to this, the property of model identifiability is intrinsic by nature and bears no *theoretical* relation to the model adequacy. However, from a practical point of view, fitting data requires estimation of model parameters, which is possible only if these parameters are identifiable. The same model expressed through an appropriately chosen non-identifiable set of parameters provides, indeed, the same fit to the data.

IDENTIFIABILITY PROPERTIES OF THE YAKOVLEV–POLIG MODEL OF CARCINOGENESIS

Another two-stage model of carcinogenesis has been proposed by Yakovlev and Polig (1996). It is similar to the MVK model in that the survival function of time to tumor has the same general form (10). However, in contrast to the MVK model, promotion and killing of initiated cells are incorporated in the Yakovlev–Polig (Y–P) model as competing risks.

The Y–P model is based on the following assumptions (Yakovlev and Polig, 1996).

(1) Let h be the time-dependent dose rate of administration of a carcinogen. Precancerous lesions are initiated according to a non-homogeneous Poisson process with the mean number of lesions produced per unit time taken to be $\theta_1 h$. Here θ_1 is a positive constant that records the sensitivity of a cell to the action of the carcinogen.
(2) Lesions responsible for cell death are formed in cells according to a non-homogeneous Poisson process with rate $\theta_2 h$, where the constant $\theta_2 > 0$ refers to the sensitivity of a cell to the killing effect of the carcinogen. Cell death is instantaneous compared to the duration of the promotion stage. Once a cell is killed, it may no longer be promoted and will not form a tumor.
(3) Cells are promoted independently of each other at random times with a CDF F. Once a cell is promoted, the formation of a tumor is irreversible.

Although the Y–P model involves administration of a carcinogen, it includes a simple model of spontaneous carcinogenesis with constant initiation rate as well, by substituting the unit dose-rate function, $h(t) = 1$, and regarding θ_1 as the rate of spontaneous initiation and θ_2 as the rate of spontaneous cell death.

The following classes of dose-rate functions h are of most biological importance.

(a) $h = const$. This is the case when irradiation with constant dose-rate throughout the lifespan is used in animal experiments. Also, as pointed out above, such dose-rate function occurs in the simplest model of spontaneous carcinogenesis.
(b) Function h is monotone decreasing. Such dose-rate functions are characteristic of experiments with incorporated radionuclides, in which case monotonic decrease of dose-rate is due to natural decay of radionuclides and their excretion.
(c) Function h is monotone increasing. This type of dose-rate functions takes into account effects of aging, in particular, the drop in repair efficiency of radiation-induced intracellular lesions. Increasing dose-rate functions are also used in models of spontaneous carcinogenesis, and lead to general models of aging (Yakovlev *et al.*, 1995).
(d) $h = \text{const} > 0$ on an initial interval of time and equals 0 thereafter. This is the simplest model of a single continuous exposure with constant dose-rate.
(e) h is a piecewise constant function on a finite interval and is equal to 0 otherwise. Such dose-rate functions are characteristic of intermittent exposures used in

many animal bioassays with fractionated continuous dose delivery.

It was shown by Yakovlev and Polig (1996) that the hazard rate φ in the Y–P model is given by the following formula:

$$\varphi(t) = \theta_1 \exp^{-\theta_2 \int_0^t h(x)\mathrm{d}x} \int_0^t h(x) f(t-x)\mathrm{d}x. \qquad (18)$$

Here h is a *given* function on $[0,\infty)$ describing variable dose rate, f is the PDF of the time of primary lesion promotion, and θ_1, $\theta_2 > 0$ are constants. It will be assumed throughout that the function h is non-negative, measurable, and positive on a set of positive Lebesgue measure. To emphasize that the model depends on parameters θ_1, θ_2 and f, we will also write $\varphi(t) = \varphi(t; \theta_1, \theta_2, f)$.

Identifiability properties of the Y–P model depend critically on the functional parameter f involved in the parameter set $\theta = (\theta_1, \theta_2, f)$. This motivates the following modification of the general definition (4) of model identifiability.

DEFINITION 2 Let $\mathscr{F}$ be a family of absolutely continuous probability distributions on $[0,\infty)$ with PDF f and CDF F (we will write $f \in \mathscr{F}$ or $F \in \mathscr{F}$). We say that the Y–P model is identifiable *in the family* $\mathscr{F}$ if

$$\varphi(t; \theta_1, \theta_2, f) = \varphi(t; \tilde{\theta}_1, \tilde{\theta}_2, \tilde{f}) \quad \text{for all} \quad t \geq 0; \qquad f, \tilde{f} \in \mathscr{F}, \qquad (19)$$

implies that $\theta_1 = \tilde{\theta}_1$, $\theta_2 = \tilde{\theta}_2$ and $f = \tilde{f}$.

Observe that if the model is identifiable in a family $\mathscr{F}$ then this is also true for every subfamily of $\mathscr{F}$.

We begin our review of identifiability properties of the Y–P model with the simple case when the dose rate h is constant.

PROPOSITION 1 *Suppose that h is constant. Then Y–P model is identifiable.*

For the proof of this and other statements in this section, which proofs are not supplied in the text, the reader is referred to Hanin and Boucher (1999).

Since $\int_0^\infty h(t)\mathrm{d}t$ represents the total dose, it is more realistic to consider the dose-rate functions h subject to the condition

$$\int_0^\infty h(t)\mathrm{d}t < \infty. \qquad (20)$$

Then it follows from the properties of convolution that the function φ is well defined. In practice, there often exists a number $T > 0$ such that $h(t) = 0$ for $t > T$, that is, the function h has compact support. In what follows, it will be assumed that dose rate function h satisfies Eq. (20) and has compact support. As stated above, one of the simplest and most important examples of dose-rate functions of this type is

$$h(x) = a \quad \text{for} \quad 0 \leq x \leq T, \quad \text{and} \quad h(x) = 0 \quad \text{for} \quad x > T, \qquad (21)$$

where a is a positive constant. Another example of dose-rate functions with compact support that will be discussed below is

$$h(x) = a_1 \quad \text{for} \quad 0 \leq x \leq T/2, \quad h(x) = a_2 \quad \text{for} \quad T/2 < x \leq T, \quad \text{and} \quad h(x) = 0 \quad \text{for} \quad x > T, \qquad (22)$$

with constants a_1, $a_2 > 0$, $a_1 \neq a_2$. More generally, we will consider piecewise constant dose-rate functions of the form

$$h(t) = \sum_{i=1}^n a_i \chi_{[\tau_{i-1}, \tau_i)}(t) \quad , \qquad (23)$$

where $a_i > 0$ is the dose rate administered over the time interval $[\tau_{i-1}, \tau_i)$, $i = 1, \ldots, n$, $0 = \tau_0 < \tau_1 < \cdots < \tau_n = T$, and χ_E stands for the indicator function of a set E. It will be assumed without loss of generality that $a_i \neq a_{i+1}$ for $i = 1, \ldots, n-1$.We now formulate a general necessary condition for identifiability of the Y–P model. If it is violated, the model is not identifiable.

THEOREM 2 *Suppose that dose-rate function h satisfies condition (20) and that, for some $T > 0$, $h(t) = 0$ for $t > T$. If Y–P model is identifiable in a family $\mathscr{F}$ then*

$$F(T) > 0 \quad \text{for all} \quad F \in \mathscr{F}. \qquad (24)$$

Theorem 2 states that identifiability of the Y–P model in a family $\mathscr{F}$ of promotion time distributions fails unless the support of the dose-rate function and supports of all distributions in $\mathscr{F}$ overlap. It is quite natural from biological standpoint to assume that there is a time delay between the first moment of exposure to a carcinogen and the moment of appearance of the first clonogenic tumor cell. In these terms, condition (24) requires this time delay to be smaller than the duration of exposure to the carcinogen. In the case of constant dose-rate (in particular, for spontaneous carcinogenesis with constant rate of initiation), this condition is automatically met, in agreement with Proposition 1 claiming that in this case Y–P model is identifiable in *any* family $\mathscr{F}$. In the rest of this section, we will consider families $\mathscr{F}$ for which condition (24) is met and identify circumstances under which this condition is also sufficient for identifiability of the Y–P model.

Let h be a piecewise constant dose-rate function (23). It follows from Eq. (18) that, for $t \geq T$,

$$\varphi_1(t) = C \sum_{i=1}^n a_i [F(t - \tau_{i-1}) - F(t - \tau_i)], \quad \text{where}$$

$$C = \theta_1 \exp^{-\theta_2 \int_0^T h(x)\mathrm{d}x}.$$

Suppose that Eq. (19) holds. Then

$$C\sum_{i=1}^{n} a_i[F(t-\tau_{i-1}) - F(t-\tau_i)]$$

$$= \tilde{C}\sum_{i=1}^{n} a_i[\tilde{F}(t-\tau_{i-1}) - \tilde{F}(t-\tau_i)], \quad t \geq T, \quad (25)$$

where $\tilde{C} = \tilde{\theta}_1 \exp(-\tilde{\theta}_2 \int_0^T h(x)\mathrm{d}x)$. Setting $t = T$ in Eq. (25) and observing that in view of Eq. (24)

$$A := \sum_{i=1}^{n} a_i[F(T-\tau_{i-1}) - F(T-\tau_i)] > 0$$

and similarly

$$\tilde{A} := \sum_{i=1}^{n} a_i[\tilde{F}(T-\tau_{i-1}) - \tilde{F}(T-\tau_i)] > 0$$

we conclude that $CA = \tilde{C}\tilde{A}$. Introduce a function

$$\psi(t) := \frac{F(t)}{A} - \frac{\tilde{F}(t)}{\tilde{A}}, \quad t \geq 0. \quad (26)$$

Then relation (25) becomes

$$\sum_{i=1}^{n} a_i[\psi(t-\tau_{i-1}) - \psi(t-\tau_i)] = 0, \quad t \geq T. \quad (27)$$

To further simplify this equation, denote

$$k_1 := 1 - \frac{a_2}{a_1}, \quad k_2 := \frac{a_2 - a_3}{a_1}, \ldots,$$
$$k_{n-1} := \frac{a_{n-1} - a_n}{a_1}, \quad k_n := \frac{a_n}{a_1}. \quad (28)$$

Note that $k_1, \ldots, k_n \neq 0$ and $k_1 + \cdots + k_n = 1$. Then Eq. (27) acquires the form

$$\psi(t) = k_1\psi(t-\tau_1) + \cdots + k_n\psi(t-\tau_n), \quad t \geq T. \quad (29)$$

Thus, identifiability problem for the Y–P model with dose-rate function (23) leads in a natural way to the functional Eq. (29) on the half-line. Observe that every solution ψ of this equation is uniquely determined by its restriction to the interval $[0,T)$, and the latter can be an arbitrary function. Eq. (29) on the whole real line was studied in the literature. In his classical memoir, Schwartz (1947) investigated the structure of complex solutions of Eq. (29) with arbitrary coefficients. Laczkovich (1986) determined the structure of non-negative measurable solutions of Eq. (29) with positive coefficients. The properties of solutions of the Eq. (29) on the half-line, however, are quite different from those in the case of the real line. Two reasons make studying the solution set of Eq. (29) important: first, piecewise constant dose-rate functions are commonly used in experimental studies; second, any dose-rate function that meets condition (20) can be approximated in L^1 norm by piecewise constant functions (Eq. (23)), which implies almost everywhere convergence of the corresponding hazard functions (Eq. (18)).

The following statement pinpoints a property of Eq. (29) that ensures identifiability of the Y–P model.

PROPOSITION 2 *Suppose that every absolutely continuous solution of Eq. (29) with coefficients (Eq. (28)) that has a finite limit at infinity is constant. Then Y–P model with dose-rate function (23) is identifiable.*

Proof Suppose that Eq. (19) holds for two parameter sets $\theta = (\theta_1, \theta_2, f)$ and $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{f})$. Then the function ψ defined in Eq. (26) satisfies Eq. (29). Clearly, ψ is absolutely continuous on $[0,\infty)$ and has a finite limit $\delta := 1/A - 1/\tilde{A}$ at infinity. According to our assumption, $\psi(t) = \delta$ for all $t \geq 0$. Differentiating Eq. (26) we find that $f(t)/A = \tilde{f}(t)/\tilde{A}$ for almost all $t \geq 0$. In view of $\int_0^\infty f(t)\mathrm{d}t = \int_0^\infty \tilde{f}(t)\mathrm{d}t = 1$, this implies that $A = \tilde{A}$, hence $f = \tilde{f}$.

For $t \in [\tau_{n-1}, T]$, we have

$$g(t) := \int_0^t h(x)f(t-x)\mathrm{d}x$$

$$= \sum_{i=1}^{n-1} a_i[F(t-\tau_{i-1}) - F(t-\tau_i)] + a_nF(t-\tau_{n-1}).$$

Since the function g is continuous on $[\tau_{n-1}, T]$ and $g(T) = A > 0$, there exists $\varepsilon \in (0, T-\tau_{n-1})$ such that $g(t) > 0$ for $t \in (T-\varepsilon, T]$. Then by Eqs. (18) and (19)

$$\theta_1 \exp^{-\theta_2 \int_0^t h(x)\mathrm{d}x} = \tilde{\theta}_1 \exp^{-\tilde{\theta}_2 \int_0^t h(x)\mathrm{d}x}, \quad t \in (T-\varepsilon, T]. \quad (30)$$

We differentiate this equality and recall that $a_n > 0$ to find that

$$\theta_1\theta_2 \exp^{-\theta_2 \int_0^t h(x)\mathrm{d}x} = \tilde{\theta}_1\tilde{\theta}_2 \exp^{-\tilde{\theta}_2 \int_0^t h(x)\mathrm{d}x}, \quad t \in (T-\varepsilon, T]. \quad (31)$$

Juxtaposing Eqs. (30) and (31), we see readily that $\theta_2 = \tilde{\theta}_2$. Then Eq. (30) implies $\theta_1 = \tilde{\theta}_1$. Therefore, $\theta = \tilde{\theta}$, which completes the proof.

COROLLARY 1 In the case $n = 1$, that is, for the dose-rate function (21), the functional equation (Eq. (29)) is simply $\psi(t) = \psi(t-T)$, $t \geq T$. Every solution of this equation is a T-periodic function on $[0,\infty)$. If such function has a finite limit at infinity, then it is necessarily constant. Therefore, in accordance with Proposition 2, Y–P model with the dose-rate function (21) is *identifiable*.

COROLLARY 2 Consider the case $n = 2$ with equally spaced switching points ($\tau_0 = 0$, $\tau_1 = T/2$, $\tau_2 = T$), which corresponds to the dose-rate function (22). Then Eq. (29) becomes

$$\psi(t) = (1-k)\psi(t-T/2) + k\psi(t-T), \quad t \geq T, \quad (32)$$

with $k = a_2/a_1$. Note that $k > 0$, $k \neq 1$. It was shown by Hanin and Boucher (1999) that, for $k > 1$, every bounded solution of Eq. (32) is $T/2$-periodic. Therefore, arguing as above and applying Proposition 2 we find that in the case of the dose-rate function (22) with $a_2 > a_1$ the Y–P model is *identifiable*.

Remark For $0 < k < 1$, Eq. (32) does have non-constant solutions with finite limit at infinity which makes Proposition 2 inapplicable. We also note that the proof of identifiability of the Y–P model by Hanin and Boucher (1999) in the case $0 < k < 1$ contains a technical gap thus leaving the problem open.

We now turn to identifiability of the Y–P model for the more general class of arbitrary bounded dose-rate functions which contains, in particular, the class (23) of piecewise constant functions. To make this problem tractable, we will have to restrict $\mathscr{F}$ to the following non-parametrically specified class of families of promotion time distributions.

DEFINITION 3 A family $\mathscr{F}$ of absolutely continuous probability distributions on $[0,\infty)$ is called *graduated* if for every two distinct PDFs $f, \tilde{f} \in \mathscr{F}$ and for every $\varepsilon > 0$, there is a number $A > 0$ (which may depend on $f, \tilde{f}$ and ε) such that either $f(t) \leq \varepsilon \tilde{f}(t)$ for all $t \geq A$ or $\tilde{f}(t) \leq \varepsilon f(t)$ for all $t \geq A$.

If $\mathscr{F}$ consists of non-vanishing PDFs then this definition is equivalent to claiming that the ratio of any two distinct PDFs from $\mathscr{F}$ must tend to zero or infinity at infinity. Indeed, every subfamily of a graduated family is graduated. As an example, the family of gamma distributions with PDF given by

$$f(t) = f_{\alpha,\beta}(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}, \quad t > 0, \qquad (33)$$

and $f(t) = 0$ for $t \leq 0$, where α, $\beta > 0$ and Γ is the gamma function, is graduated. On the contrary, the family (16) of latency time distributions arising in the MVK model is not graduated. For any PDF from this family with parameters a, b, ρ behaves at infinity like $C\exp\{-b\rho t\}$, where C is a positive constant depending on a, b and ρ, so that the ratio of two distinct PDFs with parameters a_1, b_1, ρ_1 and a_2, b_2, ρ_2 satisfying $b_1\rho_1 = b_2\rho_2$ tends at infinity to a constant different from 0 and infinity. More generally, the same is true for any family of distributions (10) (where it can always be assumed that F is a CDF of a *proper* probability distribution, i.e. that $\lim_{t\to\infty} F(t) = 1$) under the condition that the distribution with CDF F has finite first moment.

Let h be a given measurable non-negative bounded function on $[0,\infty)$ supported on an interval $[0,T]$. Denote $S_h := \{x \in [0,T] : h(x) \neq 0\}$. For a non-negative function $f \in L^1([0,\infty))$, we set

$$E_f := \{t \in [0,T] : f \quad \text{does not vanish a.e. on } t - S_h \cap [0,t]\}.$$

Significance of the set E_f stems from the fact that, for $0 \leq t \leq T$,

$$\int_0^t h(x)f(t-x)\mathrm{d}x = \int_{S_h\cap[0,t]} h(x)f(t-x)\mathrm{d}x = \int_{t-S_h\cap[0,t]} h(t-u)f(u)\mathrm{d}u,$$

so that $t \in E_f$ is equivalent to

$$\int_0^t h(x)f(t-x)\mathrm{d}x > 0.$$

THEOREM 3 *Suppose that dose-rate function h is bounded and for some $T > 0$ we have $h(t) = 0$ for all $t > T$. Let $\mathscr{F}$ be a graduated family.*

1. *If Y–P model is identifiable in the family $\mathscr{F}$ then* $\mathrm{mes}\, E_f > 0$ *for every function* $f \in \mathscr{F}$.
2. *If* $\mathrm{mes}(S_h \cap E_f) > 0$ *for all* $f \in \mathscr{F}$ *then Model 1 is identifiable.*

As a corollary of Theorem 3, we obtain the following verifiable criterion of identifiability of the Y–P model that extends Theorem 2 to the case of graduated families of promotion time distributions.

THEOREM 4 *Suppose that dose-rate function h is bounded, supported on $[0,T]$ for some $T > 0$, and positive almost everywhere on $[0,T]$. Then Y–P model is identifiable in a graduated family $\mathscr{F}$ if and only if $F(T) > 0$ for all $F \in \mathscr{F}$.*

Yakovlev *et al.* (1977), Yakovlev and Polig (1996), Boucher and Yakovlev (1997), Boucher *et al.* (1998) and Tsodikov and Müller (1998) used Y–P model for analysis of real time-to-tumor data that resulted from animal experiments. In particular, Y–P model has explained successfully the inverse dose-rate effect in radiation carcinogenesis (Yakovlev *et al.*, 1977; Yakovlev and Polig, 1996) and allowed estimation of the proportion of initiated cells, which are killed by urethane in mice (Boucher and Yakovlev, 1997; Boucher *et al.*, 1998). The promotion time distribution was assumed to belong to the gamma family with PDF given by Eq. (33). As follows from Theorem 4, in this case the model is identifiable for all practically important dose-rate functions h. This explains why in the papers cited above parameters of the Y–P model were successfully estimated using the method of maximum likehood.

Possible lack of identifiability of the Y–P model leads to the problem of computing its parametric dimension and describing identifiable combinations of parameters θ_1, θ_2, f.

IDENTIFIABILITY PROBLEM FOR TWO-STAGE MODELS

In the two previous sections, we discussed mechanistic models of tumor latency providing biologically motivated explicit analytic expressions for the distribution of time T to the appearance of the first clonogenic tumor cell measured from the birth of individual for spontaneously arising tumors and from the start of exposure to carcinogen in the case of induced tumorigenesis. These formulas contain several parameters which, under certain conditions discussed above, are identifiable from the distribution of RV T. In many instances, the distribution of time to tumor is assumed to belong to the gamma family or any other flexible multiparametric identifiable family for that matter. It should be emphasized that duration of tumor latency is *unobservable* which renders direct inference about its distribution including parameter estimation impossible. This difficulty can be circumvented by resorting to observable endpoints, such as age at tumor detection. This necessitates involvement of the progression stage of tumor development (Yakovlev *et al.*, 1996; Hanin *et al.*, 1997; Bartoszyński *et al.*, 2001) that was neglected in most of the previous works, and modeling the process of tumor detection, see next section for details. A distinctive feature that makes this approach appealing is availability of additional clinical information on tumor size at detection. Consider, in particular, the simplest case of deterministic exponential tumor growth. In this case $f(t) = e^{\lambda t}$, where t is time from the appearance of the first clonogenic tumor cell (that is, from tumor onset), $f(t)$ is the tumor size (the number of cells) at time t, and $\lambda > 0$ is a constant growth rate. As shown in the next section, under natural assumptions about the detection process, the time W of spontaneous tumor detection measured from tumor onset has the distribution with PDF

$$f_W(w) = \alpha e^{\lambda w} e^{-(\alpha/\lambda)(e^{\lambda w}-1)}, \quad w \geq 0, \tag{34}$$

where α is a positive constant interpreted as the rate of spontaneous tumor detection. It is easy to see that parameters α and λ are jointly identifiable from the distribution of RV W.

In practice, observed is a sample from the distribution of RV $T + W$ (age at spontaneous tumor detection). Since biological mechanisms governing duration of tumor latency and those of tumor progression and detection are quite different and have no direct bearing on each other, RVs T and W can be viewed as independent. The question now becomes: *Is the entire set of parameters involved in the distributions of RVs T and W identifiable from the distribution of RV* $T + W$? This leads to the following general problem.

PROBLEM *Let $\mathcal{P}$ and $\mathcal{Q}$ be two families of probability distributions on* $[0,\infty)$. *Is it true that the family of convolutions* $P * Q$, *where* $P \in \mathcal{P}$ *and* $Q \in \mathcal{Q}$, *is identifiable? In other words, does*

$$P_1 * Q_1 = P_2 * Q_2,$$

where $P_1, P_2 \in \mathcal{P}$ *and* $Q_1, Q_2 \in \mathcal{Q}$, *imply that* $P_1 = P_2$ *and* $Q_1 = Q_2$?

Taking both families to be the set of degenerate distributions δ_a, $a \geq 0$, and observing that $\delta_a * \delta_b = \delta_{a+b}$, we conclude that in general the answer to this question is negative. Yet another counterexample is given by gamma distributions $\Gamma(a_1,b)$ and $\Gamma(a_2,b)$ with the convolution being equal to $\Gamma(a_1 + a_2, b)$.

The following theorem provides sufficient conditions for the positive solution of the problem. For the proof of this result, see Bartoszyński *et al.* (2001).

THEOREM 5 *Let $\mathcal{P}$ and $\mathcal{Q}$ be two families of absolutely continuous probability distributions on* $[0,\infty)$ *with PDFs* $p \in \mathcal{P}$ *and* $q \in \mathcal{Q}$. *Suppose that*

(1) *family $\mathcal{P}$ is graduated;*
(2) *for every* $p \in \mathcal{P}$, *there is* $M = M(p) > 0$ *such that* $p(t) > 0$ *for all* $t > M$;
(3) *for every* $p \in \mathcal{P}$ *and for each* $s > 0$, *there exist a finite limit*

$$h_p(s) := \lim_{t\to\infty} \frac{p(t-s)}{p(t)};$$

(4) *for each* $q \in \mathcal{Q}$,

$$\lim_{t\to\infty} \int_0^t \frac{p(t-s)}{p(t)} q(s)\mathrm{d}s = \int_0^\infty h_p(s)q(s)\mathrm{d}s;$$

(5) *for all* $p \in \mathcal{P}$ *and* $q \in \mathcal{Q}$,

$$0 < \int_0^\infty h_p(s)q(s)\mathrm{d}s < \infty.$$

Then the family of convolutions $P * Q$, *where* $P \in \mathcal{P}$ *and* $Q \in \mathcal{Q}$, *is identifiable.*

Remark 1 For concrete families $\mathcal{P}$ and $\mathcal{Q}$, condition (4) of the theorem usually follows from standard results about passage to limit in the Lebesgue integral. In particular, Theorem 5 can be applied in the case when $\mathcal{P}$ is the family of gamma distributions $\Gamma(a,b)$ with shape parameter $a \geq 1$ and $\mathcal{Q}$ is the family (34). First, conditions (1) and (2) of Theorem 5 are clearly satisfied. Next, for $p \in \Gamma(a,b)$ we have for all $s \geq 0$

$$\frac{p(t-s)}{p(t)} = \left(\frac{t-s}{t}\right)^{a-1} e^{bs} \to e^{bs} \quad \text{as} \quad t \to \infty.$$

Hence, condition (3) is met with $h_p(s) = e^{bs}$. Further, assuming that $a \geq 1$ we obtain, for any PDF q from the

family (34),

$$\int_0^t \frac{p(t-s)}{p(t)} q(s)\mathrm{d}s = \int_0^\infty \frac{p(t-s)}{p(t)} \chi_{[0,t]}(s) q(s)\mathrm{d}s \to \int_0^\infty \mathrm{e}^{bs} q(s)\mathrm{d}s < \infty$$

by the Lebesgue's Theorem on Dominated Convergence, which is condition (4). Finally, condition (5) also holds. This leads us to a conclusion that if promotion time T follows a gamma distribution $\Gamma(a,b)$ with $a \geq 1$ and tumor growth is exponential with rate λ then parameters a, b, α and λ are jointly identifiable from the observed distribution of the age at spontaneous tumor detection.

Remark 2 Since the MVK family (16) is not graduated and for the family (34) condition (3) of Theorem 5 is missed, the convolution of the PDF corresponding to the MVK model with Eq. (34) is not covered by Theorem 5. The same is true for the Y–P model. However, this does not mean necessarily that those convolutions are non-identifiable, but rather that more powerful analytic methods are required to clarify their properties associated with the notion of identifiability.

IDENTIFIABILITY OF THE JOINT DISTRIBUTION OF AGE AND TUMOR SIZE AT DETECTION

Age U and tumor size S at spontaneous tumor detection serve as a valuable source of information for making inference about important biological parameters involved in the distribution of unobservable duration T of tumor latency discussed in "Non-identifiability of the Moolgavkar–Venzon–Knudson two-stage model of carcinogenesis" and "Identifiability properties of the Yakovlev–Polig model of carcinogenesis" sections. Spontaneous tumor detection occurs in the course of occasional medical checks or through onset of clinical symptoms of the disease and should be distinguished from screening based detection that comes as a result of disease specific medical exams scheduled at prescribed time moments. Let W be the time of spontaneous detection counted from the moment of tumor onset, then $U = T + W$. Let $f: [0,\infty) \to [1,\infty)$ be a deterministic function describing the law of tumor growth, then $S = f(W)$ is the corresponding tumor size at detection. The function f may depend on one or several parameters that reflects individual variability of tumor progression. The most important example is the exponential growth $f(w) = \mathrm{e}^{\lambda w}$, $\lambda > 0$, see Bartoszyński (1987) for substantiation. Another example is given by the two-parametric Gompertz family,

$$S(w) = \mathrm{e}^{A(1-\mathrm{e}^{-Bw})} \tag{35}$$

with constant parameters $A, B > 0$.

The process of spontaneous tumor detection will be characterized by the hazard function (detection rate) $r(t)$.

We proceed from the following assumptions.

(1) Function f is differentiable and $f' > 0$.
(2) RVs T and W are absolutely continuous and independent.
(3) The rate of spontaneous detection is proportional to tumor size: $r = \alpha S$, where $\alpha > 0$ is a constant.

Assumption (2) suggests that tumor progression is independent of the age of onset of the disease. Assumption (3) goes back to Brown *et al.* (1984), see also Klein and Bartoszyński (1991) and Bartoszyńsky *et al.* (2001). For a more detailed discussion of our hypotheses, the reader is referred to Bartoszyńsky *et al.* (2001) and Hanin *et al.* (2001). It will be assumed in what follows that $\lim_{t\to\infty} f(t) = \infty$; however, all results in this section are also true with minor notational changes for the case when the limit is finite. An example of saturated tumor growth is given by the Gompertz model (35).

According to assumption (2) and formula (3), the survival function for the RV W is given by

$$\bar{F}_W(w) = \exp(-\int_0^w r(u)\mathrm{d}u) = \exp(-\alpha\int_0^w f(u)\mathrm{d}u) = \mathrm{e}^{-\alpha\Phi(w)}, \quad w \geq 0,$$

where $\Phi(w) := \int_0^w f(u)\mathrm{d}u$. Hence,

$$f_W(w) = \alpha f(w)\mathrm{e}^{-\alpha\Phi(w)}, \quad w \geq 0.$$

Then for the tumor size S at detection we have

$$\bar{F}_S(s) = \bar{F}_W(g(s)) = \mathrm{e}^{-\alpha\Phi(g(s))}, \quad s \geq 1,$$

where we denote hereafter by g the inverse function to $f(g := f^{-1})$ which existence follows from assumption (1). Therefore,

$$f_S(s) = \alpha s g'(s)\mathrm{e}^{-\alpha\Phi(g(s))}, \quad s \geq 1. \tag{36}$$

In particular, for $f(t) = \mathrm{e}^{\lambda t}$,

$$\bar{F}_W(w) = \mathrm{e}^{-(\alpha/\lambda)(\mathrm{e}^{\lambda w}-1)}, \quad w \geq 0,$$

which justifies formula (34) for the PDF of RV W, and also

$$\bar{F}_S(s) = \mathrm{e}^{-(\alpha/\lambda)(s-1)}, \quad s \geq 1. \tag{37}$$

Equation (37) suggests that tumor size at detection follows a translated exponential distribution with parameter α/λ.

To compute the distribution of random vector $Y := (T + W, S)$, we look at Y as a transformation of the random vector $X := (T, W)$, $Y = \varphi(X)$, where $\varphi(t,w) = (t + w, f(w))$, $t, w \geq 0$. According to assumption (2), components of X are independent RVs. The inverse function $\psi = \varphi^{-1} : A \to [0,\infty) \times [0,\infty)$, where $A :=$

$\{(u,s): u \geq 0, 1 \leq s \leq f(u)\}$, is given by $\psi(u,s) = (u - g(s), g(s))$. Note that the Jacobian of ψ is g'. Then for the PDF of Y we have

$$\begin{aligned} f_Y(u,s) &= f_X(\psi(u,s))g'(s) \\ &= f_T(u - g(s))f_W(g(s))g'(s) \\ &= f_T(u - g(s))f_S(s), \quad u \geq 0, \quad 1 \leq s \leq f(u), \end{aligned} \tag{38}$$

and 0 otherwise. In the particular case of exponential tumor growth with rate $\lambda > 0$, we find using formula (37) that

$$f_Y(u,s) = \frac{\alpha}{\lambda} e^{-(\alpha/\lambda)(s-1)} f_T\left(u - \frac{\ln s}{\lambda}\right),$$

$$u \geq 0, \quad 1 \leq s \leq e^{\lambda u}.$$

Thus, the distribution of random vector Y is absolutely continuous, but the support of Y depends on unknown parameters involved in the law of tumor growth (λ in the case of exponential growth).

Let the distribution of tumor latency time T depend on a parameter set θ. Notationally, this will be reflected by setting $f_T(t) = h(t;\theta)$ with the understanding that $h(t;\theta) = 0$ for $t < 0$. Similarly, to show dependence of the law of tumor growth f and the distribution of tumor size at detection S on parameters, we will write $f(t) = f(t;\eta)$ and $f_S(s) = k(s;\xi)$. Observe that parameter sets η and ξ may have common elements. The following theorem shows that under natural conditions all parameters involved in the joint distribution of age and tumor size at detection are identifiable.

THEOREM 6 *Suppose that*

i) $f'(t;\eta) > 0$ *for all* $t \geq 0$ and η;
ii) $h(u,\theta) > 0$ *for all* $u > 0$ and θ;
iii) $k(s,\xi) > 0$ *for all* $s \geq 1$ and ξ;
iv) *parameters of functions* h, k, f *are identifiable.*

Then model (38) of the joint distribution of age and tumor size at detection is identifiable.

Proof Suppose that

$$h(u - g(s;\eta);\theta)k(s;\xi) = h(u - g(s;\tilde{\eta});\tilde{\theta})k(s;\tilde{\xi}) \tag{39}$$

for all $u \geq 0, s \geq 1$. Comparing the supports of both sides of this equality with conditions (i)–(iii) and formula (38) taken into account, we conclude that $f(t,\eta) = f(t;\tilde{\eta})$ for all $t \geq 0$. Therefore, in view of condition (iv), $\eta = \tilde{\eta}$. Next, integrating both sides of Eq. (39) for fixed s from $g(s;\eta)$ to infinity and observing that

$$\int_{g(s;\eta)}^{\infty} h(u - g(s;\eta);\theta)\mathrm{d}u = \int_0^{\infty} h(x;\theta)\mathrm{d}x = 1$$

we conclude that $k(s;\xi) = k(s;\tilde{\xi})$ for all $s \geq 1$. Hence, owing to condition (iv), $\xi = \tilde{\xi}$. Also, due to condition (iii), we derive from Eq. (39) that $h(u - g(s;\eta);\theta) = h(u - g(s;\eta);\tilde{\theta})$ for all $s \geq 1$ and $u \geq g(s;\eta)$ so that $h(x;\theta) = h(x;\tilde{\theta})$ for $x \geq 0$. Finally, applying condition (iv) we find that $\theta = \tilde{\theta}$. Theorem 6 is proved.

Remark Condition (i) is satisfied for the exponential and Gompertz laws of tumor growth. Condition (ii) is met for any tumor latency time model in which the hazard function is positive on $(0,\infty)$. In particular, this is the case for the gamma distribution, MVK model and also Y–P model provided that CDF F of the promotion time is positive on $(0,\infty)$ and the dose rate function h is positive almost everywhere on $(0,a)$ for some $a > 0$. Finally, it follows from Eq. (36) that condition (iii) of the theorem holds under assumptions (1) and (3).

IDENTIFICATION OF RANDOMIZED MULTIHIT MODELS OF IRRADIATED CELL SURVIVAL

It was discovered in the 1920s that biological effects of ionizing radiation are significantly different from those caused by other physical and chemical agents. In the spirit of the ideas of quantum mechanics, this was attributed to the discrete nature of ionizing radiation and stochastic character of radiation energy scattering on cell structures. The following two postulates that form the core of the "hit and target" theory were suggested to obtain a quantitative description of the biological effects of ionizing radiation.

(1) *Target principle.* Every cell contains a small sensitive region (called the *target*) that has to be effected for damage to result.
(2) *Hit principle.* There is a critical number m of hits in the cell target such that the cell is killed if its target is hit at least $m + 1$ times.

It is now commonly believed that cell target can be identified with cellular DNA, which single and double strand breaks constitute primary radiation-induced lesions. These lesions are subject to repair. Therefore, it is more biologically relevant to interpret m as the number of unrepaired (or irreparable) lesions (assumed to be identical) that a cell can bear without being killed. Alternatively, one may assume that a cell dies when it has at least one unrepaired lesion, but its repair capacity is limited to m lesions. For an at depth discussion of the foundations of the hit and target theory, the reader is referred to Dantzer (1934), Clifford (1972), Turner (1975), Hanin *et al.* (1993), Hanin *et al.* (1994) and Hanin *et al.* (1996).

It follows from the above two principles that survival probability of a cell exposed to an instantaneously delivered dose D of ionizing radiation is given by

$$S(D;m,x) = e^{-xD}\sum_{k=0}^{m}\frac{(xD)^k}{k!}, \tag{40}$$

where $x > 0$ is referred to as cell *radiosensitivity*. Since the expected number of hits in the cell target is equal to xD, parameter x can be viewed as a basic characteristic of the physical side of the damaging process. Formula (40) represents the classical *m hit model*.

Damage repair capacity of cells displays a high degree of variability. Therefore, it is appropriate to view the critical number m of hits in the target as a RV M taking non-negative integer values. By compounding Eq. (40) with the distribution of RV M we obtain the following randomized version of the multihit model (40):

$$S(D;Q,x) = \mathrm{e}^{-xD}\sum_{m=0}^{\infty}\Pr(M=m)\sum_{k=0}^{m}\frac{(xD)^k}{k!}$$

$$= \mathrm{e}^{-xD}\sum_{k=0}^{\infty}\frac{(xD)^k}{k!}Q(k), \tag{41}$$

where $Q(k) := \Pr(M \geq k)$, $k \geq 0$, is the survival sequence for the RV M. Denote by $\mathcal{Q}$ the class of all such sequences. It is easy to see that a sequence $\{Q(k)\}_{k=0}^{\infty}$ belongs to $\mathcal{Q}$ if and only if it is non-increasing, $Q(0) = 1$ and $\lim_{k\to\infty} Q(k) = 0$. The model (41) was extensively studied by Clifford (1972). It is used for describing dose-effect relationships for a single cell or a cell population that is homogeneous with respect to radiosensitivity x. A few biologically motivated examples of model (41) are given below.

(1) *The classical m-hit model (40).* It corresponds to the sequence $Q(k) = 1$ for $k = 0, 1, \ldots, m$ and $Q(k) = 0$ for $k > m$.

(2) *One-hit-to-kill model with misrepair.* This is the classical m-hit model with $m = 0$ with the additional feature that repair of a lesion may not be successful (which is called *misrepair*). Suppose that misrepair of a primary lesion occurs with probability q. Then the expected number of unrepaired lesions is $(1-q)xD$, and therefore

$$S(D;q,x) = \mathrm{e}^{-(1-q)xD} = \mathrm{e}^{-xD}\sum_{k=0}^{\infty}\frac{(xD)^k}{k!}q^k,$$

which is a randomized multihit model (41) with $Q(k) = q^k$, $k \geq 0$.

(3) *m-hit model with radiation-induced damage repair.* There is experimental evidence suggesting that exposure to radiation enhances lesion repair processes, so that every lesion that is not repaired by the background damage repair system can be repaired by its additional radiation-induced component. Let p be the probability of this event. Then the survival probability is given by

$$S(D;m,p,x) = \mathrm{e}^{-xD}\left[\sum_{k=0}^{m}\frac{(xD)^k}{k!} + \sum_{k=m+1}^{\infty}\frac{(xD)^k}{k!}p^{k-m}\right].$$

This function belongs to the class (41) with $Q(k) = 1$ for $k = 0, 1, \ldots, m$ and $Q(k) = p^{k-m}$ for $k > m$.

(4) *Multitarget model.* Suppose that a cell having $n \geq 2$ identical targets (for example, chromosomes) is killed when all its targets are destroyed, i.e. hit at least once. If x is radiosensitivity of the cell, then for its survival probability we have

$$S(D;n,x) = 1 - (1 - \mathrm{e}^{-xD/n})^n$$

$$= \mathrm{e}^{-xD}\sum_{k=0}^{\infty}\frac{(xD)^k}{k!}Q_n(k)$$

with

$$Q_n(k) = \sum_{i=0}^{n-1}(-1)^{n-i-1}\binom{n}{i}\left(\frac{i}{n}\right)^k.$$

It was shown by Clifford (1972) that this sequence belongs to the class $\mathcal{Q}$.

The use of the model (41) is hampered by the fact that parameter x is not identifiable from the dose-effect function $S(D)$. This remarkable fact, that was discovered by Clifford (1972), is a matter of the following statement.

PROPOSITION 3 *If a function $S(D)$ has representation (41) for some $x > 0$ with a sequence $Q = Q_x \in \mathcal{Q}$ then, for every $y > x$, there exists a sequence $Q_y \in \mathcal{Q}$ such that*

$$S(D) = \mathrm{e}^{-xD}\sum_{k=0}^{\infty}\frac{(xD)^k}{k!}Q_x(k)$$

$$= \mathrm{e}^{-yD}\sum_{k=0}^{\infty}\frac{(yD)^k}{k!}Q_y(k). \tag{42}$$

Specifically,

$$Q_y(k) = \sum_{i=0}^{k}\binom{k}{i}\left(1-\frac{x}{y}\right)^{k-i}\left(\frac{x}{y}\right)^i, \quad k \geq 0.$$

Estimates of the smallest value of x for which representation (41) of a given function $S(D)$ holds with some sequence $Q = Q_x \in \mathcal{Q}$ were obtained by Hanin *et al.* (1996). It was also shown in this work that each of the two parameters Q and x of the model (41) is uniquely determined by the dose-effect function $S(D)$ if the other parameter is fixed. Therefore, parametric dimension of the randomized multihit model (41) is equal to one.

Radiosensitivity x of cells in a cell population varies in wide limits and thus could be thought of as a RV. Denoting by μ its distribution and integrating Eq. (41) with respect to μ, we obtain the following randomized multihit model for the population dose effect function:

$$S(D;Q,\mu) = \int_0^{\infty}\mathrm{e}^{-xD}\sum_{k=0}^{\infty}\frac{(xD)^k}{k!}Q(k)\mathrm{d}\mu(x). \tag{43}$$

By contrast to the model (41), the two distributional parameters of the model (43) turn out to be identifiable in

several important cases listed in the following theorem. For the proof of this theorem, see Hanin *et al.* (1996).

THEOREM 7

(1) *For the classical m-hit model with a fixed non-random value m and random x, the distribution μ is identifiable.*
(2) *If μ is any given non-degenerate distribution ($\mu \neq \delta_0$) that has finite moments of all orders then the distribution of the hit parameter m is identifiable.*
(3) *Model (43) is completely identifiable in every family $\{(Q, \mu)\}$, where Q is the survival sequence for a distribution on $\mathbb{Z}_+$ with finite moments of all orders and μ is a gamma distribution.*

Acknowledgements

The paper was written in Summer 2001 while the author was visiting Huntsman Cancer Institute of the University of Utah. The author is thankful to the Huntsman Cancer Foundation that made this visit possible. The work was partially supported by the grant 1U01 CA88177-01 from the National Cancer Institute. The author is grateful to Dr A. Yu. Yakovlev for many fruitful discussions of and insightful comments on identification of stochastic models.

References

Armitage, P. and Doll, R. (1957) "The two-stage theory of carcinogenesis in relation to the age distribution of human cancers", *Br. J. Cancer* **11**, 161–169.

Bartoszyński, R. (1987) "A modeling approach to metastatic progression of cancer", In: Thompson, J.R. and Brown, B.W., eds, *Cancer Modeling* (Marcel Dekker, New York and Basel), pp 237–267.

Bartoszyński, R., Edler, L., Hanin, L.G., Kopp-Schneider, A., Pavlova, L.V., Tsodikov, A.D., Zorin, A.V. and Yakovlev, A.Yu. (2001) "Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis", *Math. Biosci.* **171**, 113–142.

Boucher, K.M. and Yakovlev, A.Yu. (1997) "Estimating the probability of initiated cell death before tumor induction", *Proc. Natl Acad. Sci. USA* **94**, 12776–12779.

Boucher, K.M., Pavlova, L.V. and Yakovlev, A.Yu. (1998) "A model of multiple tumorigenesis allowing for cell death quantitative insight into biological effects of urethane", *Math. Biosci.* **150**, 63–82.

Brown, B.W., Atkinson, N.E., Bartoszyński, R. and Montague, E.D. (1984) "Estimation of human tumor growth rate from distribution of tumor size at detection", *J. Natl Cancer Inst.* **72**, 31–38.

Clifford, P. (1972) "Nonthreshold models of the survival of bacteria after irradiation", *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **Vol. 4**, *Biology and Health*, (Unversity of California Press, Berkeley) pp 265–286.

Dantzer, H. (1934) "Über einige Wirkungen von Strahlen VII", *Z. Phys.* **89**, 421–434.

Hanin, L.G. and Boucher, K.M. (1999) "Identifiability of parameters in the Yakovlev–Polig model of carcinogenesis", *Math. Biosci.* **160**, 1–24.

Hanin, L.G. and Yakovlev, A.Yu. (1996) "A nonidentifiability aspect of the two-stage model of carcinogenesis", *Risk Anal.* **16**(5), 711–715.

Hanin, L.G., Rachev, S.T. and Yakovlev, A.Yu. (1993) "On the optimal control of cancer radiotherapy for non-homogeneous cell populations", *Adv. Appl. Prob.* **25**, 1–23.

Hanin, L.G., Pavlova, L.V. and Yakovlev, A.Yu. (1994) *Biomathematical Problems in Optimization of Cancer Radiotherapy* (CRC Press, Boca Raton).

Hanin, L.G., Klebanov, L.B. and Yakovlev, A.Yu. (1996) "Randomized multihit models and their identification", *J. Appl. Prob.* **33**, 458–471.

Hanin, L.G., Rachev, S.T., Tsodikov, A.D. and Yakovlev, A.Yu. (1997) "A stochastic model of carcinogenesis and tumor size at detection", *Adv. Appl. Prob.* **29**, 607–628.

Hanin, L.G., Tsodikov, A.D. and Yakovlev, A.Yu. (2001) "Optimal schedules of cancer surveillance and tumor size at detection", *Math. Comp. Model.* **33**, 1419–1430.

Heidenreich, W.F. (1996) "On the parameters of the clonal expansion model", *Radiat. Environ. Biophys.* **35**, 127–129.

Heidenreich, W.F., Luebeck, E.G. and Moolgavkar, S.H. (1997) "Some properties of the hazard function of the two-mutation clonal model", *Risk Anal.* **17**(3), 391–399.

Karlin, S. (1966) *A First Course in Stochastic Processes* (Academic Press, New York).

Klebanov, L.B., Rachev, S.T. and Yakovlev, A.Yu. (1993) "A stochastic model of radiation carcinogenesis: latent time distributions and their properties", *Math. Biosci.* **113**, 51–75.

Klein, M. and Bartoszyński, R. (1991) "Estimation of growth and metastatic rates of primary breast cancer", In: Arino, O., Axelrod, D.E. and Kimmel, M., eds, *Mathematical Population Dynamics* (Marcel Dekker, New York), pp 397–412.

Kopp-Schneider, A., Portier, C.J. and Sherman, C.D. (1994) "The exact formula for tumor incidence in the two-stage model", *Risk Anal.* **14**, 1079–1080.

Laczkovich, M. (1986) "Non-negative measurable solutions of difference equations", *J. Lond. Math. Soc.* **34**(2), 139–147.

Moolgavkar, S.H. and Knudson, A.G. (1981) "Mutation and cancer: a model for human carcinogenesis", *J. Natl Cancer Inst.* **66**, 1037–1052.

Moolgavkar, S.H. and Luebeck, G. (1990) "Two-event model for carcinogenesis: biological, mathematical, and statistical considerations", *Risk Anal.* **10**, 323–341.

Moolgavkar, S.H. and Venzon, D.J. (1979) "Two-event models for carcinogenesis: incidence curves for childhood and adult tumors", *Math. Biosci.* **47**, 55–77.

Schwartz, L. (1947) "Théorie générale des fonctions moyenne-périodiques", *Ann. Math.* **48**, 857–929.

Tsodikov, A. and Müller, W. (1998) "Modeling carcinogenesis under a time-changing exposure", *Math. Biosci.* **152**, 179–191.

Turner, M.M. (1975) "Some classes of hit-target models", *Math. Biosci.* **23**, 219–235.

Yakovlev, A.Yu. and Polig, E. (1996) "A diversity of responses displayed by a stochastic model of radiation carcinogenesis allowing for cell death", *Math. Biosci.* **132**, 1–33.

Yakovlev, A.Yu., Müller, W., Pavlova, L.V. and Polig, E. (1977) "Do cells repair precancerous lesions induced by radiation?", *Math. Biosci.* **142**, 107–117.

Yakovlev, A.Yu., Tsodikov, A.D. and Bass, L. (1993) "A stochastic model of hormesis", *Math. Biosci.* **116**, 197–219.

Yakovlev, A.Yu., Tsodikov, A.D. and Anisimov, V.N. (1995) "A new model of aging: specific versions and their application", *Biometrical J.* **37**, 435–448.

Yakovlev, A.Yu., Hanin, L.G., Rachev, S.T. and Tsodikov, A.D. (1996) "A distribution of tumor size at detection and its limiting form", *Proc. Natl Acad. Sci. USA* **93**, 6671–6675.

Zheng, Q. (1994) "On the exact hazard and survival functions of the MVK stochastic carcinogenesis model", *Risk Anal.* **14**, 1081–1084.