BIOMETRICS 53, 1495–1502
December 1997

A Distribution of Tumor Size at Detection: An Application to Breast Cancer Data

Alexander D. Tsodikov,[1] Bernard Asselain,[2] and Andrej Y. Yakovlev[3]

[1]Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Liebigstrasse 27, 04103 Leipzig, Germany

[2]Biostatistiques, Institut Curie, 26 rue D'Ulm, 75231 Paris Cedex 05, France

[3]Huntsman Cancer Institute, Division of Public Health Science, University of Utah, 546 Chipeta Way, Suite 1100, Salt Lake City, Utah 84108, U.S.A.

SUMMARY

This paper discusses a method of estimating numerical characteristics of unobservable stages of carcinogenesis from data on tumor size at detection. To this end, a stochastic model of spontaneous carcinogenesis has been developed to allow for a simple pattern of tumor growth kinetics. It is assumed that a tumor becomes detectable when its size attains some threshold level, which is treated as a random variable. The model yields a parametric family of joint distributions for tumor size and age at detection. Some estimation problems associated with the proposed model appear to be tractable. This is illustrated with an application to the statistical analysis of data on primary breast cancer.

1. Introduction

The presently most popular two-stage model of carcinogenesis, also known as the Moolgavkar–Venzon–Knudson model (Moolgavkar and Venzon, 1979; Moolgavkar and Knudson, 1981), is focused on the events that precede the occurrence of the first malignant cell in a tissue. It is a common feature of many modern mechanistic models of carcinogenesis that they are devoid of the stage of tumor progression and mechanisms of tumor detection. The reason is that such models are basically intended for the analysis of time-to-tumor observations, which are likely to contain less information than is required for identification of a more complex model of tumor development. However, there is no relevant evidence that the time of tumor progression is negligibly small compared with the duration of earlier stages of carcinogenesis. This is fully appreciated by key contributors to this trend of research (Yang and Chen, 1991; Luebeck and Moolgavkar, 1994; Kopp-Schneider and Portier, 1995). A deterministic delay incorporated into some versions of the Moolgavkar–Venzon–Knudson model to describe the growth kinetics of malignant cells cannot be considered a cure for this difficulty (Kopp-Schneider and Portier, 1995). This pertains equally to the model of radiation carcinogenesis developed by Klebanov, Rachev, and Yakovlev (1993) and its generalization by Yakovlev and Polig (1996).

A mathematically appealing class of models allowing for a stochastic description of the progression stage is constituted by so-called threshold models. This name comes from the assurance that a tumor becomes detectable when its size attains some threshold level. As one example, tumor growth can be assumed to obey the postulates of a supercritical birth-and-death process with two absorbing states so that the first passage time with respect to the upper barrier will correspond to the time of tumor progression. The idea was explored with models for microbial infections with deterministic thresholds (Williams, 1965; Morgan and Watts, 1980). The results of Morgan and Watts (1980) suggest that incubation period data alone are insufficient for model identifiability. One is likely to face the same identifiability problem when attempting to develop a similar model for analysis of time-to-tumor observations.

Key words: Breast cancer; Carcinogenesis; Conditional distribution; Maximum likelihood estimation; Parametric analysis; Tumor size.

The problem calls for an additional source of information. As suggested by Yakovlev and Tsodikov (1996), this information might be provided by data on the primary tumor size at detection. The papers of Yakovlev et al. (1996) and Hanin et al. (1996) further explore this possibility. The authors proposed a threshold counterpart of the two-stage model of carcinogenesis, which makes it possible to estimate biologically meaningful parameters of initiation, promotion, and progression from data on tumor size at detection and age of the individuals (patients) diagnosed with a specific cancer. Their approach, which is briefly outlined in Sections 2–4, represents an alternative to some other endeavors to relate the chance of detecting a tumor to its size (Brown et al., 1984; Bartoszyński, 1987; Kimmel and Flehinger, 1991; Klein and Bartoszyński, 1991). The proofs of theoretical results are given in Hanin et al. (1996). The purpose of the present paper is to illustrate this approach with an application to the analysis of real data (Section 5).

2. The Model

The basic premises behind the proposed model can be briefly formulated as follows.

(i) The initiation event in the process of carcinogenesis is the formation of a primary intracellular lesion that, in the long run, is capable of producing an overt tumor. It makes sense to think of these precancerous lesions as initiated cells. Such primary events of lesion formation occur at random times, and their sequence in time is modeled as a homogeneous Poisson process with intensity θ.

(ii) A primary lesion remains dormant as long as it proceeds through the promotional stage of tumor development. Let $R(t)$ be the cumulative distribution function of this stage duration. All lesions are subject to promotion independently of each other.

The two assumptions given above are common to most of the modern two-stage stochastic models of carcinogenesis. To accommodate the data on tumor size at detection, Yakovlev et al. (1996) invoke the following additional assumptions.

(iii) Once the first malignant cell arises as a result of tumor promotion, its subsequent growth is irreversible and the progression stage begins. It is this clonogenic cell that gives rise to a detectable tumor after a lapse of time, which is thought of as a random variable with cumulative distribution function $F(t)$.

(iv) A tumor becomes detectable when its size attains some threshold value N, which is treated as a random variable. A linear pure birth process with the absorbing upper barrier N is used to model the dynamics of tumor growth. The critical number of tumor cells is represented as $N = cV$, where V is the volume of a tumor and c is the concentration of tumor cells per unit volume. The constant c is nonrandom and its values are typically large. The conditional progression time distribution function, given the threshold volume $V = v$, is

$$F(t \mid v) = (1 - e^{-\lambda t})^{cv}, \tag{1}$$

where λ is the birth rate. Equation (1) is derived under the assumption that tumor growth starts from a single malignant cell at time $t = 0$.

(v) The lengths of the promotion and progression stages are mutually independent.

Let $L(t)$ be the cumulative distribution function of the time it takes for the initiation and promotion processes to result in the first malignant cell. Derived from the above assumptions is the following expression for the corresponding survivor function $\bar{L}(t) = 1 - L(t)$:

$$\bar{L}(t) = \exp\left\{-\theta \int_0^t R(x)dx\right\} \tag{2}$$

(see Klebanov et al., 1993; Yakovlev and Polig, 1996). In Section 5, the two-parameter gamma distribution will be used to represent the function $R(x)$ in equation (2). Many analyses of experimental data lend support to this choice (Klebanov et al., 1993; Yakovlev, Tsodikov, and Bass, 1993; Yakovlev et al., 1995; Yakovlev and Tsodikov, 1996).

Let $g(t \mid v)$ stand for the conditional probability density function for the time of tumor latency measured from the date of birth of an individual. Then it follows from Assumption v and equations (1) and (2) that

$$g(t \mid v) = \lambda\theta cv \int_0^t e^{-\lambda(t-s)}(1 - e^{-\lambda(t-s)})^{cv-1} R(s) e^{-\theta \int_0^s R(x)dx} ds. \tag{3}$$

Introducing a marginal distribution $P(v)$ of the tumor volume V, we represent the probability density function of the detection time (age of the patient) as

$$g(t) = \int_0^\infty g(t \mid v)p(v)dv,$$

where $p(v)$ is the density of $P(v)$. Now the conditional probability density function of tumor volume at detection (given a tumor is detected at time t), hereafter denoted by $w(v \mid t)$, follows immediately from Bayes' formula as

$$w(v \mid t) = \frac{g(t \mid v)p(v)}{\int_0^\infty g(t \mid u)p(u)du} = \frac{g(t \mid v)p(v)}{g(t)}.$$

With t and cv tending to infinity, the conditional density $w(v \mid t)$ assumes a much simpler form, which is free from the promotion-time distribution $R(t)$. The following formula (see Hanin et al., 1996, for proof) holds for the limiting behavior of the conditional probability density function $w(v \mid t)$:

$$w(v \mid \infty) := \lim_{t\to\infty} w(v \mid t) = \frac{v^\mu p(v)}{\int_0^\infty u^\mu p(u)du}, \qquad cv \to \infty, \tag{4}$$

where $\mu = \quad \{1, \theta/\lambda\}$. A special case ($\mu = 1$) of this distribution is associated with what is known as a length-biased sampling from stationary point processes (Cox and Lewis, 1966). A sampling bias inherent in screening procedures under a stable disease model (Zelen and Feinleib, 1969) provides yet another example. Along similar lines, the case of induced carcinogenesis can be considered (Yakovlev et al., 1996; Hanin et al., 1996).

3. Estimation Problems with Bivariate Data

Suppose that clinical data are available as to the tumor size V and the age A at detection for patients diagnosed with a specific cancer. First, we assume that such data arise from the joint probability density function $g(v,t) = g(t \mid v)p(v)$. This assumption is warranted if the effect of data censoring due to competing risks is negligible (see Section 5). In such an event, the log-likelihood is represented as

$$\ell = \sum_i \log g(t_i \mid v_i) + \sum_i \log p(v_i) = \ell_1 + \ell_2. \tag{5}$$

It is clear that ℓ_1 and ℓ_2 can be maximized independently of each other, resulting in the empirical distribution function $\hat{P}(v)$ for estimation of the cumulative distribution function $P(v)$.

More generally, account must be taken of a competing risk that precludes tumor detection from occurring. To accommodate this censoring effect, we assume that the competing risk of death from all other causes is independent of the one of interest. The competing risk is characterized by its latent time Y. Let $S(y)$ be the survivor function for Y. It follows that

$$p_c(v) = p(v \mid A < Y) = \frac{p(v)\int_0^\infty g(u \mid v)S(u)du}{\int_0^\infty g(u)S(u)du} \tag{6}$$

and

$$g_c(v,t) = g(v,t \mid A < Y) = \frac{p_c(v)g(t \mid v)S(t)}{\int_0^\infty g(u \mid v)S(u)du}.$$

Since $S(t)$ is free from unknown parameters, the log-likelihood assumes the form

$$\ell_c = \sum_i \log g(t_i \mid v_i) - \sum_i \log \int_0^\infty g(u \mid v_i)S(u)du + \sum_i \log p_c(v_i), \tag{7}$$

which is similar to (5) in that the distribution $P(v)$ is irrelevant to the estimation of the parameters of $g(t \mid v)$. The log-likelihood (7) reduces to (5) if $S(t) = 1$ almost everywhere.

4. Inference Based on the Limiting Distribution $w(v \mid \infty)$

The adequacy of the limiting probability density function $w(v \mid \infty)$ can be explored indirectly through testing the hypothesis of conditional independence of V and A given $A > t^*$, where the value of t^* is to be estimated from a given sample. This will be put into effect in Section 5 by the application of Spearman's test. It is noteworthy that the presence of an independent competing risk leaves the form of $w(v \mid t)$ unaltered so that the parameter $\mu = \{1, \theta/\lambda\}$ can be estimated from a subsample of patients whose ages exceed t^*. Let n be this subsample size. Based on (4), the log-likelihood function is

$$\ell(\mu) = \mu \sum_{i=1}^{n} \log v_i + \sum_{i=1}^{n} \log p(v_i) - n \log \mathrm{E}(V^\mu).$$

Therefore, the maximum likelihood estimator $\hat{\mu}^*$ of the parameter μ can be obtained as a solution of the following equation:

$$\frac{\mathrm{E}(V^\mu \log V)}{\mathrm{E}(V^\mu)} = \frac{1}{n} \sum_{i=1}^{n} \log v_i. \tag{8}$$

The estimation equation (8) is unbiased since the left-hand side of (8) represents the expected value of the random variable $\log V$ conditional on $A > t^*$, the right-hand side being its empirical estimate. Using the Cauchy–Schwarz inequality, it can be shown that, except for the trivial case of a degenerate random variable V, the left-hand side of (8) is strictly monotone in μ, and thus the equation has a unique solution (Hanin et al., 1996).

The asymptotic variance of $\hat{\mu}^*$ is estimated by

$$D(\hat{\mu}^*) \approx \left[-\frac{\partial^2 \ell(\mu)}{\partial \mu^2}\bigg|_{\hat{\mu}^*} \right]^{-1} = \left\{ \frac{1}{n} \left[\sum_i \log v_i \right]^2 - n \frac{\mathrm{E}(V^\mu [\log V]^2)}{\mathrm{E}(V^\mu)} \right\}^{-1}.$$

We assume that the whole sample size is very large compared to the value of n, so the empirical distribution function $\hat{P}$ can be used as a substitute for the true distribution P when making inferences from the subsample of old individuals. Then the expected values in the left-hand side of (8) can be replaced with their empirical counterparts related to the whole sample. Denote by $\hat{p}_k$, $k = 1, \ldots, m$, the frequencies assigned by the empirical distribution function $\hat{P}$ to the m distinct observed values of tumor volume. For discrete or grouped data, the limiting distribution $W(v \mid \infty)$ is estimated by

$$\hat{w}_k^* = \frac{v_k^{\hat{\mu}^*} \hat{p}_k}{\sum_j v_j^{\hat{\mu}^*} \hat{p}_j}. \tag{9}$$

It should be noted that estimate (9) provides $\max_\mu \ell(\mu)$ conditionally on $\hat{P}$. Suppose that $\mu < 1$. Then the estimate $\hat{\mu}^*$ can be used to reduce the dimensionality of the likelihood ℓ_1 in (5) by imposing the constraint $\theta/\lambda = \hat{\mu}^*$.

5. Data Analysis

We analyze data on the primary tumor size for 2129 premenopausal patients diagnosed with clinical stage I–III unilateral invasive breast carcinoma and treated at the Curie Institute between 1981 and 1987. A comprehensive description of these patients is given in the paper of Rochefordiere et al. (1993). The data include clinical tumor size referring to one-dimensional measurements (in mm). It is shown to be highly correlated with the size measured by mammography; their mean values coincide quite closely. Tumors were assumed to be spherical and this approximation was used to measure their volumes in mm^3. It should be noted that the clinically measured tumor size usually results in wide grouping intervals (Figure 1). Mean age of the patients was 44.5 years.

First, we should make sure that the limiting density $w(v \mid \infty)$ is applicable to our data. Using Spearman's test, we tested the conditional independence of V and A given $A > t^*$ for various subsamples created by sampling from pooled adjacent age strata. This procedure resulted in the value of $t^* = 50$ years with 536 patients older than 50 in the data set. When applied to the whole sample, the Spearman test rejects the independence hypothesis at a significance level much lower than 0.001. Based on equation (8), the estimated value of μ is $\hat{\mu}^* = 0.1054$ (with the asymptotic

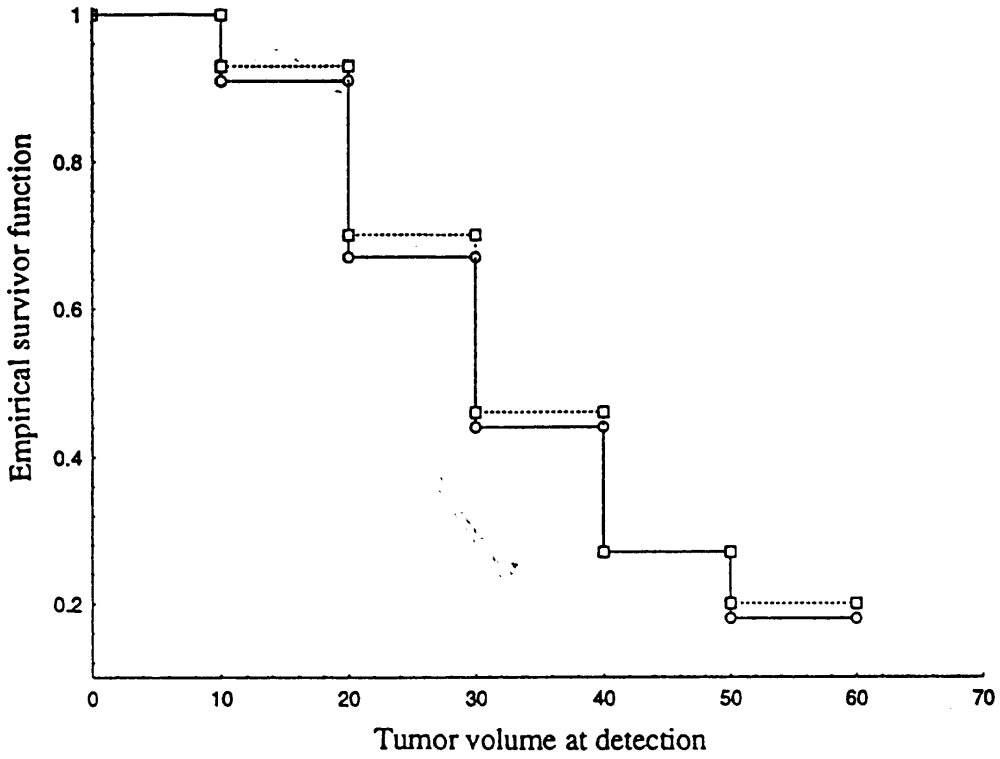

Figure 1. Empirical tail functions obtained from the whole sample (solid line) and from the subsample of patients older than 50 (dotted line).

95% confidence interval [0, 0.2290]), which is biologically plausible because it is natural to expect that the intensity of tumor cell proliferation is much higher than the rate of precancerous lesion formation. Recalling equation (4), we see that, with μ small, the unconditional distribution density $p(v)$ must be close to the function $w(v \mid \infty)$, which is insensitive to the presence of an independent competing risk. Since the value of $\hat{\mu}^*$ is small, the distribution $P(v)$ can be estimated from a subsample of patients whose age exceeds t^*. We denote the empirical counterpart of $P(v)$ thus constructed by $\hat{P}^*(v)$. The empirical tail function $1 - \hat{P}^*(v)$ is plotted in Figure 1 together with the corresponding estimate $1 - \hat{P}(v)$ obtained from the whole sample. It is seen in Figure 1 that the two estimates coincide very closely. This suggests that the discrepancy between the tumor size distribution $P(v)$ and the conditional distribution $P_c(v)$ is sufficiently small for the effect of data censoring to be ignored. In addition to the trivial case $S(t) = 1$, one can see from equations (3) and (6) that $p_c(v)$ closely approximates $p(v)$ if the parameter λ is sufficiently large. In our numerical experiments conducted with large values of λ, the ratio $p_c(v)/p(v)$ remained close to 1 when the mean value of a uniformly distributed censoring time was varied through a wide range. It may not be out of place to note the model stability under small perturbations in the probability density function $p(v)$ shown by Hanin et al. (1996).

Now the other biologically meaningful parameters can be estimated by maximizing the log-likelihood (5), which accounts for variations of the tumor volume at detection with the age of the patient. To maximize ℓ_1, use was made of a three-step nonlinear programming procedure based on random search (Zhigljavsky, 1992; Brooks and Morgan, 1995), the algorithm of Davidon, Fletcher, and Powell (Himmelblau, 1972), and the Zoutendijk algorithm (Himmelblau, 1972). The procedure is described at length in Yakovlev and Tsodikov (1996). The coefficient c was assigned a value of 10^6 cells/mm^3 (Klein and Bartoszyński, 1991). The promotion-time distribution was assumed to be gamma distributed with shape parameter α and scale parameter β. Incorporated in the model are two other parameters to be estimated from the data under study; these are the initiation rate θ and the rate of tumor cell proliferation λ. We first maximized the reduced likelihood ℓ_1 using 0.1054λ in place of θ. Having removed the constraint and using the solution of the constrained problem as the initial point in the search for the maximum of $\ell_1(\lambda, \theta, \alpha, \beta)$, we found the estimates changed, but very slightly. We obtained the following maximum likelihood estimates of the model parameters: $\hat{\theta} = 16.9$, $\hat{\lambda} = 159$, $\hat{\alpha} = 11$, $\hat{\beta} = 0.1027$. Given the large value of λ (short progression stage), the likelihood (7) yields almost the same estimates of the parameters θ, α, and β even in the presence of heavy censoring. Computed from these estimates, the expected value and the standard deviation of the promotion time are approximately equal to 107 and 32 years, respectively. The

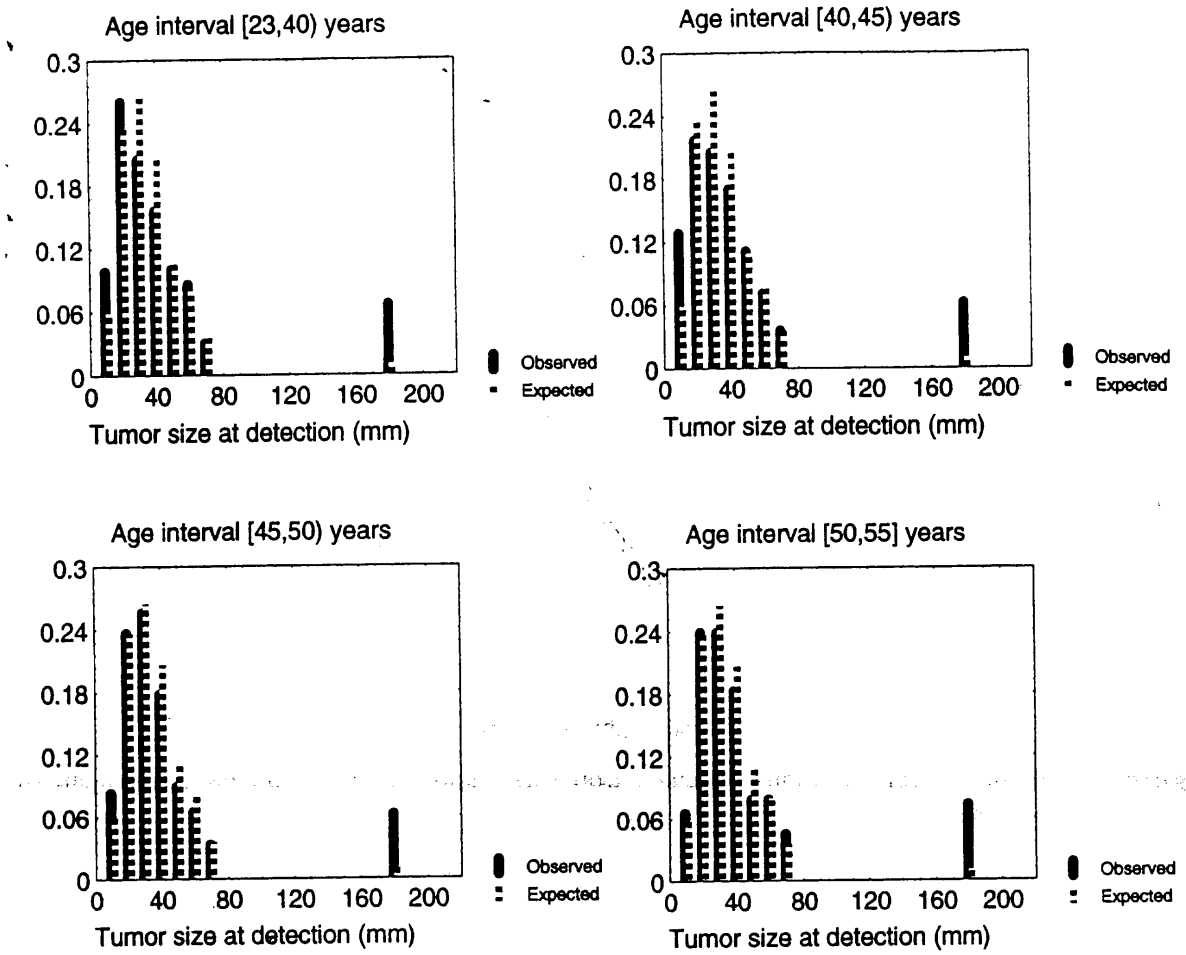

Figure 2. Observed versus expected frequencies of breast cancer size at detection for different age strata. The value of t is taken equal to the midpoint of each age interval.

long promotion stage indicates that only a proportion of tumors, characterized by a high promotion rate, manifest themselves during the lifetime of an individual. The value of $\hat{\mu} = \hat{\theta}/\hat{\lambda} = 0.1063$ is in good agreement with the estimate $\hat{\mu}^* = 0.1054$ based on equation (8).

Using the estimated parameter values, the corresponding parametric estimates of the densities $g(t \mid v)$ and $w(v \mid t)$ can be readily obtained. In doing so, we represent the conditional distribution density of tumor size at detection as

$$w_k(t) = \frac{g(t \mid v_k)p_k}{\sum_j g(t \mid v_j)p_j}$$

to contrast it with the corresponding observed frequencies. The resultant fit of the model to the data for different age strata is shown in Figure 2. There is a small proportion of large tumors that are not predicted by the model. This point is worth examining further.

In this example, the effect of data censoring appears to be small because of the short duration of the progression stage. For tumors with longer progression stages, additional information may be called on to estimate the function $S(t)$ involved in equations (6) and (7). A regression counterpart of the above presented model offers a natural classification of covariates in terms of their predominant effect on different stages of tumor development. This issue will be addressed in another paper.

Acknowledgements

We are very thankful to Drs R. Bartoszyński and D. Pearl for their valuable comments. The research of Dr Tsodikov was supported by grant LO 342/6-1 of the German Research Foundation. A large part of this research was carried out while Dr Yakovlev was visiting the German Cancer Research Center (Heidelberg, Germany) as recipient of an Alexander von Humboldt Award.

Résumé

Ce papier présente une méthode d'estimation des caractéristiques quantifiées des étapes non observables de la carcinogenèse à partir de données concernant la taille de la tumeur au moment

du diagnostic. Pour cela un modèle stochastique de carcinogenèse spontanée a été développé pour permettre une description simple de la cinétique de croissance cellulaire. Nous posons l'hypothèse qu'une tumeur devient détectable quand sa taille atteint un niveau seuil, qui est considéré comme une variable aléatoire. Le modèle repose sur une famille paramétrique de distributions jointes pour la taille tumorale et l'âge au diagnostic. Quelques problèmes d'estimation associés au modèle proposé ont pu être résolus. Une application du modèle à l'analyse de données dans le cancer primitif du sein est présentée.

References

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

Brooks, S. P. and Morgan, B. J. T. (1995). Optimisation using simulated annealing. *The Statistician* **44,** 241–257.

Brown, B. W., Atkinson, N. E., Bartozyński, R., and Montague, E. D. (1984). Estimation of human tumor growth rate from distribution of tumor size at detection. *Journal of the National Cancer Institute* **72,** 31–38.

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events.* London: Methuen, New York: John Wiley.

Hanin, L. G., Rachev, S. T., Tsodikov, A. D., and Yakovlev, A. Yu. (1996). A stochastic model of carcinogenesis and tumor size at detection. *Advances in Applied Probability,* in press.

Himmelblau, D. M. (1972). *Applied Nonlinear Programming.* New York: McGraw-Hill.

Kimmel, M. and Flehinger, B. J. (1991). Nonparametric estimation of the size–metastasis relationship in solid cancers. *Biometrics* **47,** 987–1004.

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

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

Kopp-Schneider, A. and Portier, C. J. (1995). Carcinoma formation in NMRI mouse skin painting studies is a process suggesting greater than two stages. *Carcinogenesis* **16,** 53–59.

Luebeck, E. G. and Moolgavkar, S. H. (1994). Simulating the process of malignant transformation. *Mathematical Biosciences* **123,** 127–146.

Moolgavkar, S. H. and Knudson, A. G. (1981). Mutation and cancer: A model for human carcinogenesis. *Journal of the National Cancer Institute* **66,** 1037–1052.

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

Morgan, B. J. T. and Watts, S. A. (1980). On modelling microbial infections. *Biometrics* **36,** 317–321.

Rochefordiere, A., Asselain, B., Campana, F., Scholl, S. M., Fenton, J., Vilcoq, J. R., Durand, J.-C., Pouillart, P., Magdelenat, H., and Fourquet, A. (1993). Age as prognostic factor in premenopausal breast carcinoma. *Lancet* **341,** 1039–1043.

Williams, T. (1965). The basic birth–death model for microbial infections. *Journal of the Royal Statistical Society, Series B* **27,** 338–360.

Yakovlev, A. and Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications.* Singapore: World Scientific Publishers.

Yakovlev, A. Y., Müller, W. A., Pavlova, L. V., and Polig, E. (1995). *A stochastic model of radiation carcinogenesis allowing for cell death and its biological implications.* Technical Report 567, Department of Statistics, Ohio State University, Columbus.

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

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

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. *Proceedings of the National Academy of Sciences, U.S.A* **93,** 6671–6675.

Yang, G. L. and Chen, C. W. (1991). A stochastic two-stage carcinogenesis model: A new approach to computing the probability of observing tumor in animal bioassays. *Mathematical Biosciences* **104,** 247–258.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic disease. *Biometrika* **56**, 601–614.

Zhigljavsky, A. (1992). *Theory of Global Random Search.* Dordrecht: Kluwer.

Received August 1995; *revised June* 1996 *and January and May* 1997; *accepted May* 1997.