ELSEVIER

Statistics & Probability Letters 32 (1997) 223–230

STATISTICS & PROBABILITY LETTERS

Assessing risk with doubly censored data: an application to the analysis of radiation-induced thyropathy

Ilya L. Kruglikov[a], Nikolaj I. Pilipenko[b], Alexander D. Tsodikov[c], Andrej Yu. Yakovlev[d,*]

[a] *Forschungszentrum, Karlsruhe, Germany*
[b] *Institute of Medical Radiology, Kharkov, Ukraine*
[c] *IMISE, Universität Leipzig, Germany*
[d] *Huntsman Cancer Institute, University of Utah, 546 Chipeta Way, Suite 1100, Salt Lake City, UT 84108, USA*

Received March 1995; revised April 1996

Abstract

This paper deals with the statistical inference from doubly censored data on the incidence of thyropathy in a group of liquidators of the Chernobyl accident with special emphasis on the long-term risk assessment. In this study, all sample observations are either left or right censored. The prime objective is to estimate the disease onset distribution and the expected proportion of responders (long-term risk) from real data of this type. We give a solution to this problem using a parametric family of improper distributions derived from a recently proposed model of radiation carcinogenesis (Klebanov et al., 1993).

Keywords: Left and right censoring; Data grouping; Onset time distribution; Nonparametric estimation; Risk assessment; Parametric model; ML estimates; Radiation-induced thyropathy; Chernobyl accident

1. Introduction

Let X be the random time to an event of interest, $G(x)$ its cumulative distribution function, and $t_1, \dots, t_m$ a sequence of time points representing the time of examination which is thought of as a random variable. If every individual under study is examined only once and the diagnostic errors are negligible, the jth test, $j = 1, \dots, N$, at time t_i may have only two outcomes $\{X < t_i\}$ and $\{X \geqslant t_i\}$, $i = 1, 2, \dots, m$, and this is just the information to be employed for estimating $G(x)$ or the corresponding survivor function $S(x) = 1 - G(x)$. Our interest here is with this special case of doubly censored data. The situation where all observations are either left or right censored is not uncommon in practice as evidenced by a relevant example given in this paper.

[*] Corresponding author.

PII S0167-7152(96)00077-6

As early as in 1955, Ayer et al. constructed a nonparametric estimator to accomodate the above-described type of data. Over the years a great many articles on the analysis of doubly censored data have been published (Turnbull, 1974, 1976; Dinse and Lagakos, 1982; Kodell et al. 1982; Bergmann and Turnbull, 1983; Turnbull and Mitchell, 1984; Dewanji and Kalbfleisch, 1986; Portier, 1986; Dinse, 1988; to name a few), the vast majority of them dealing with more complicated observation designs and implying the presence of uncensored observations that supply estimation procedures with additional information. Turnbull (1976) extended the concept of self-consistency to develop an iterative algorithm converging to the maximum likelihood estimate of the distribution function with arbitrary grouped, censored and truncated data. Although derived from statistical considerations, the algorithms based on the self-consistency property have much in common with those that employ nonlinear programming methods. An example described in Section 3 of this paper motivated us to recall the problem in its original formulation with special focus on the long-term risk assessment.

Suppose a population of individuals is susceptible to a certain disease. The challenge is to estimate the distribution of this disease onset time and the expected proportion of affected individuals (long-term risk) from doubly censored observations. It is clear that nonparametric estimators are incapable of providing a projection of risk forward in time beyond a surveillance period. Even if the observation process extends over a long time, the proportion of unaffected individuals, also known as the surviving fraction (Miller, 1981; Yamaguchi, 1992; Yakovlev, 1994), is difficult to estimate nonparametrically with reasonable accuracy. The estimator by Ayer et al., much like the Kaplan–Meier estimator in presence of heavy right censoring (Pepe and Fleming, 1989; Cantor and Shuster, 1992), tends to fail in this situation. With a specific application presented in Section 3, we give a parametric solution to the problem, proceeding from a stochastic model of radiation carcinogenesis recently proposed by Klebanov et al. (1993).

2. Estimation procedures

2.1. *Traditional nonparametric estimator*

For a given surveillance strategy

$$0 < t_1 < t_2 < \cdots < t_m < \infty,$$

let a_i be the number of right censored and b_i be the number of left censored observations at time t_i, $i = 1, \ldots, m$. If there are no uncensored observations we have the likelihood

$$L = \prod_{i=1}^{m} (S(t_i))^{a_i}(1 - S(t_i))^{b_i}. \tag{1}$$

We consider $S(t)$ as a step-function, $S_i = S(t_i)$ being its value on $[t_i, t_{i+1})$, $i = 0, \ldots, m - 1$, and $S_m = S(t_m)$ being defined solely at the point t_m. Denote by E_m, the set of all vectors $e_m = (S_0, S_1, \ldots, S_m)$ satisfying the inequality

$$1 = S_0 \geqslant S_1 \geqslant \cdots \geqslant S_{m-1} \geqslant S_m \geqslant 0. \tag{2}$$

Let $\mathfrak{I}_m$ be the corresponding class of admissible survivor functions. As shown by Ayer et al. (1955), a consistent estimator for the survivor function $S(t)$ can be obtained by maximizing likelihood (1) over the class $\mathfrak{I}_m$:

$$\max_{S(t) \in \mathfrak{I}_m} L(e_m), \quad e_m \in E_m. \tag{3}$$

The estimator is given by the following simple procedure:

If the sequence $\{a_i/(a_i + b_i)\}_{i=1}^m$ is nonincreasing, then S_i is estimated by $\hat{S}_i = a_i/(a_i + b_i)$, $i = 1, \dots, m$, and $\hat{S}_0 = 1$.

If $a_i/(a_i + b_i) < a_{i+1}/(a_{i+1} + b_{i+1})$ for some $i = 1, \dots, m$, then both S_i and S_{i+1} are estimated by

$$\hat{S}_i = \hat{S}_{i+1} = \frac{a_i + a_{i+1}}{a_i + a_{i+1} + b_i + b_{i+1}}.$$

The procedure is repeated in arbitrary order until a sequence $\{\hat{S}_i\}_{i=0}^m$ satisfies (2), the resultant $\hat{S}(t) \in \mathfrak{J}_m$ being unique.

2.2. *Data grouping*

Since the ratios $a_i/(a_i + b_i)$ provide the maximum likelihood estimators, $\tilde{S}_i$, disregarding constraint (2), the question arises of how to handle the data if there are no tied observations in a given sample. Actually, in this case $\tilde{S}_i$ may take either the value 1 or the value 0 at all points specified by a sample. Clearly, this and less extreme situations call for a data smoothing procedure. The most straightforward way to do this is through data grouping. It is worth noting that the multinomial likelihood (1) is of the same form as the one resulting from grouped observations. However, the question arises of how to group such data. It turns out that the traditional nonparametric algorithm by Ayer et al. can be interpreted in terms of *data* grouping while it originally pools the *estimates* to satisfy (2). The meaning of the following procedure is that the set of grouping points $\{\tau_i\}$ is considered as a part of the set of parameters to be estimated by the maximum likelihood method (see Tsodikov (1995) for a general discussion).

Let $T = \max_{1 \leqslant i \leqslant m} t_i$. Denote by τ_i, $i = 0, 1, \dots, n$, $n \leqslant m$, the time points that specify a partition of the interval $[0, T]$ into a set of grouping intervals $\{[\tau_{i-1}, \tau_i)\}_{i=1}^n$. It is natural to assume that τ_i take their values on the set $\{t_i\}_{i=1}^m$ and satisfy the inequalities

$$0 = \tau_0 < \tau_1 < \cdots < \tau_n = t_m.$$

Proceeding from the original sample $\{t_i, a_i, b_i\}_{i=1}^m$, one may consider the number of right censored observations, a_i', and the number of left censored observations, b_i', entering the interval $[\tau_{i-1}, \tau_i)$, so that

$$a_i' = \sum_{j:\tau_{i-1} \leqslant t_j < \tau_i} a_j, \qquad b_i' = \sum_{j:\tau_{i-1} \leqslant t_j < \tau_i} b_j. \tag{4}$$

Denote by D_n the set of all possible partitions $d_n = \{\tau_i\}_{i=1}^n$ of the interval $[0, T]$ into n subintervals. Let S_i', $i = 0, \dots, n-1$, stand for the value of the step-function $S(t)$ on $[\tau_i, \tau_{i+1})$. A natural class $\mathfrak{R}$ of admissible survivor functions is defined by the triple $(E_n', D_n, n \leqslant m)$, where E_n' is the set of all possible $e_n' = (S_0', S_1', \dots, S_n')$ subject to the constraint

$$1 = S_0' > S_1' > \cdots > S_n' \geqslant 0. \tag{5}$$

The sought-for estimator $\{\hat{S}_i', \hat{\tau}_i, \hat{n}\}$ can be obtained as a solution to the following problem:

$$\max_{S(t) \in \mathfrak{R}} L'(e_n', d_n, n), \quad e_n' \in E_n', \tag{6}$$

where

$$L' = \prod_{i=1}^{n} (S_i')^{a_i'}(1 - S_i')^{b_i'},$$

and a_i', b_i' are given by (4).

Owing to the property

$$\prod_{j:\tau_{i-1}\leqslant t_j<\tau_i}(S)^{a_j}(1-S)^{b_j}=(S)^{\sum_{j:\tau_{i-1}\leqslant t_j}a_j}(1-S)^{\sum_{j:\tau_{i-1}\leqslant t_j}b_j}=(S)^{a_i'}(1-S)^{b_i'}, \tag{7}$$

there exists a one-to-one correspondence between the class $\mathfrak{J}$ and $\mathfrak{R}$, the pairs of their elements being characterizaed by equal likelihoods L and L', respectively. This fact ensures a solution of problem (6) similar to that of problem (3):

If $a_i'/(a_i'+b_i')\leqslant a_{i+1}'/(a_{i+1}'+b_{i+1}')$ for some i, then the point τ_i is eliminated from the current partition, the other points being newly enumerated from i to $n_{\text{new}}=n-1$.

It should be noted that the stepwise curves $\hat{S}(t)$ and $\hat{S}'(t)$ do not necessarily coincide. The difference is attributable to the difference between the classes $\mathfrak{J}$ and $\mathfrak{R}$ and to the left continuity convention. If some interval $[\tau_i,\tau_{i+1})$ contains a number of pooled subintervals $[t_j,t_{j+1})$ then the estimate $\hat{S}$ by Ayer et al. will have a step at the end of the first subinterval, while the estimate $\hat{S}'$ will have a step at τ_{i+1}, filling the entire interval $[\tau_i,\tau_{i+1})$ with the value $\hat{S}_i'$. In fact, the estimate by Ayer et al. has the same value of the likelihood L' as the estimate $\hat{S}'$ and may as well represent a solution of problem (6). The reverse is not true. The special convenience of the estimator $\hat{S}'(t)$ is that is shows explicitly how to group data subjected to either left or right censoring.

2.3. *Parametric estimation*

Within the nonparametric framework, the values of $\hat{S}_i$ (or $\hat{S}_i'$) affect the estimates on adjacent intervals solely by virtue of imposing restriction (2) (or (5)) on the estimated survivor function. The greatest impact upon the surviving fraction estimation is expected from late observations, first and foremost from those related to t_m. Suppose that all the persons examined at this time were found to be ill (or even that the last observation is distinct and left censored) whereas the estimates $\hat{S}_0,\hat{S}_1,\ldots,\hat{S}_{m-1}$ appeared to satisfy the monotonicity condition. Does it mean that the surviving fraction estimate should be taken equal to zero? Clearly, the positive answer might be very wide of the truth. In less extreme cases one can meet with this difficulty as well. It is clear that a much better solution to the problem can be obtained, providing a pertinent parametric model is available.

It is obvious that a more reliable and substantive inference from real biomedical data is provided by constructing biologically-based models, rather than by selecting a suitable distribution among standard parametric families. In an effort to develop a method for the radiation-induced cancer risk assessment, Klebanov et al. (1993) proposed a simple stochastic model that yields an improper distribution for the time of tumor latency under different irradiation conditions. In the application that follows, use is made of the following version of the model:

$$S(t)=\exp\{-\theta F(t)\}, \tag{8}$$

where θ is the expected number of precancerous lesions induced by irradiation, and $F(t)$ is the cumulative distribution function of the progression time, i.e., the time it takes for a single lesion to produce a detectable tumor. The parameter θ may be taken to be proportional to the dose value D, i.e., $\theta=\theta_0 D$, where θ_0 is the expected number of lesions per unit dose. Further parametrizations with respect to D are also feasible.

To put formula (8) to practical use, it remains to specify the progression time distribution $F(t)$. In the analysis presented in Section 3, the progression time is assumed to be gamma distributed with shape parameter λ and scale parameter μ so that the mean, τ, and the standard deviation, σ, of the progression time are computed as follows: $\tau=\lambda/\mu$, $\sigma=\sqrt{\lambda}/\mu$. This flexible parametric family, very simple as it is, reflects a multi-stage structure of the process of tumor progression. Computer simulations conducted with a comprehensive model of tumor development are in favor of such a choice (Ivankov et al., 1992).

We proceed from the likelihood L' and the mode of data grouping furnished by the fore-going nonparametric procedure (Section 2.2) for the purpose of contrasting the parametric estimate with the nonparametric one. In order to maximize L' with S_i' represented in the parametric form (8), we use a 3-step nonlinear programming procedure based on random search, the algorithm of Davidon, Fletcher and Powell, and the Zoutendijk algorithm (see Hoang et al. (1995) for details and references).

3. Application to data on radiation-induced thyropathy

Preliminary screening studies of the children exposed to the radioactive iodine from the Chernobyl fallout have demonstrated an increase in the frequency of thyroid diseases, hyperplasia of the thyroid gland and benign tumors included (Bolshova et al., 1991). Another important risk category is represented by the persons (we call them liquidators), who helped to clean up the Chernobyl nuclear plant. This population as a whole is highly heterogeneous regarding dose and time of irradiation, spectrum of the exposed radionuclides, etc. A more homogeneous subpopulation of liquidators is comprised of those of them who might be irradiated only while working in the contaminated zone of the Chernobyl accident. In particular, data have been collected on liquidators whose domicile is Kharkov (Ukraine) or its adjacent territory, i.e., Kharkov district. From the radiation-hygienic viewpoint, Kharkov, situated 500 km east from Chernobyl, is recognized as safe after the accident since no significant radioactive fallouts have been registered in the area. All liquidators were male and there was no special selection procedure of their employment.

To study the incidence of thyropathy, a group of liquidators, residing in Kharkov or Kharkov district, was examined by clinical and ultrasound methods in the Kharkov Institute of Medical Radiology. The disease was assumed to be irreversible, and detectable only in the course of medical examinations. With rare exception, every liquidator was examined only once at some random time after his job had been completed. The Chernobyl accident happened on 26 April 1986.

Since no reliable information was available regarding individual absorbed doses of external and internal irradiation, the following two categories of liquidators were specified to allow for levels of radiation on a relative scale.

Group 1: those who had left the zone of the Chernobyl accident before 31 July 1986 ($N = 334$);

Group 2: those who started and finished their work in the zone in 1987 ($N = 282$).

The two groups of liquidators were similar with respect to the age distribution. Using the procedure given in Section 2.2 the data were grouped as shown in Table 1.

From the general viewpoint one could expect an increased frequency of thyropathies in Group 1, since at least for some time these people worked at the Chernobyl NPS during the "iodine" period of the accident. Liquidators of Group 2 were exposed to the doses of external irradiation whose cumulative values were at least an order of magnitude less than those for Group 1, and they did not inhale the radioactive iodine.

It seems natural to proceed from the assumption that the observed thyropathies were induced by irradiation of liquidators in the zone of accident. However, both the nonparametric and parametric estimates (Fig. 1A, B) are in conflict with the assumption, suggesting that the liquidators of Group 2 develop thyropathies much earlier than those of Group 1. In other words, the latent period appears to be longer for persons who worked in the zone almost immediately after the Chernobyl accident, which is unrealistic. The estimated parameters are $\hat{\theta}_1 = 0.71$, $\hat{\tau}_1 = 6.90$, $\hat{\sigma}_1 = 1.30$ in Group 1, and $\hat{\theta}_2 = 0.78$, $\hat{\tau}_2 = 1.42$, $\hat{\sigma}_2 = 0.73$ in Group 2. These estimates are in good agreement with those obtained from the ungrouped data: $\hat{\theta}_1 = 0.71$, $\hat{\tau}_1 = 6.60$, $\hat{\sigma}_1 = 1.28$ in Group 1, $\hat{\theta}_2 = 0.78$, $\hat{\tau}_2 = 1.28$, $\hat{\sigma}_2 = 0.52$ in Group 2. The likelihood ratio test rejects the hypothesis of homogeneity (equality of the model parameters) for these groups at a significance level of nearly 0.05, the disparity between the two survivor functions being attributable to the progression time parameters. In a recent paper, Andersen and Rønn (1995) proposed a nonparametric test for comparing two

Table 1
The data on thyropathy incidence after grouping

Group 1, time after work in the zone									
Grouping points	5	6	37	46	50	56	57	60	66
# Left censored cases	0	1	28	23	18	52	15	26	2
# Right censored cases	3	4	36	28	18	48	13	15	0
Group 1, time after the accident									
Grouping points	7	14	39	49	57	60	63		
# Left censored cases	0	5	21	27	68	53	1		
# Right censored cases	4	10	30	28	61	26	0		
Group 2, time after work in the zone									
Grouping points	1	12	15	36	39	60			
# Left censored cases	2	13	3	45	22	60			
# Right censored cases	7	19	3	42	17	40			
Group 2, time after the accident									
Grouping points	21	29	56	61					
# Left censored cases	8	7	66	66					
# Right censored caases	16	10	62	47					

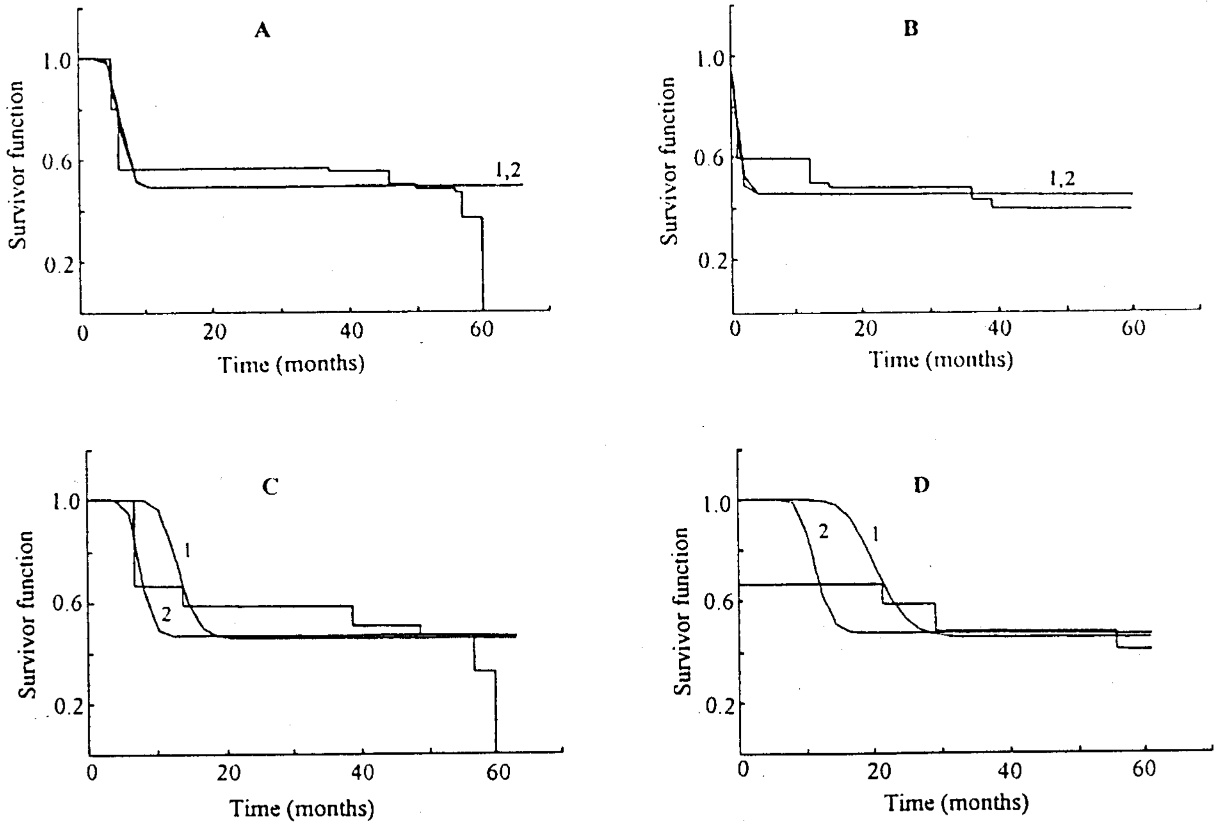

Fig. 1. The parametric and nonparametric estimates for the thyropathy onset distribution. (A and B) time is measured from the departure of workers from the zone of accident, (A) liquidators of Group 1, (B) liquidators of Group 2; (C and D) time is measured from the date of accident, (C) liquidators of Group 1, (D) liquidators of Group 2. Stepwise curve is the estimate by Ayer et al., solid lines 1 and 2 refer to grouped and ungrouped data, respectively.

Table 2
Maximum likelihood estimates for the model parameters. The onset time is measured from the date of the accident; l is the log-likelihood value

	Group 1				Group 2			
Data	$\hat{\theta}$	$\hat{\tau}$	$\hat{\sigma}$	l	$\hat{\theta}$	$\hat{\tau}$	$\hat{\sigma}$	l
Ungrouped	0.76	8.23	1.38	− 228.458	0.74	11.72	1.84	− 195.015
Grouped	0.78	13.99	2.24	− 226.896	0.78	21.17	3.81	− 193.254

samples where all observations are either left or right censored. When applied to Groups 1 and 2 of liquidators, the one-sided version of this asymptotic two-sample test rejects the null hypothesis at a significance level of slightly higher than 0.1.

Everything falls into place as soon as we assume that it is the time of the accident that should be taken as the starting point in our consideration. The model yields reasonable results and seems consistent with data when the time to tumor onset is measured from the date of the disaster. Notwithstanding some discrepancy between the mean progression time estimates obtained from grouped and ungrouped data (Table 2), both analyses reveal no significant difference between the two groups of liquidators in this case. This is corroborated by the likelihood ratio test. The two-sided test of Andersen and Rønn results in a significance level of higher than 0.9. The corresponding estimates of the survivor functions are depicted in Fig. 1(C) and (D). It is noteworthy that the estimated values of θ for Groups 1 and 2 coincide very closely, the same evidently being valid for the surviving fractions.

This epidemiological finding still awaits its biomedical interpretation. Since no other explicit causes of the thyroid diseases have been revealed in the Kharkov region so far, we may assume that all liquidators were affected by the Chernobyl accident, irrespective of whether or not they were in the zone of accident in April–July, 1986. An added irradiation of liquidators during their work in the zone had little if any effect on the dynamics of thyroid hyperplasias in this population. In other words, the Chernobyl accident shows itself as a general radioecological factor though its specific mechanisms have yet to be understood. The "intact" population of residents should be examined carefully to detect an increase of morbidity suggested by the above observation. In this connection it is interesting to note that a distinct increase in the incidence of thyroid cancer in the whole of the Ukraine has been reported (Likhtarev et al., 1995).

Acknowledgements

We are very grateful to Drs. N. Afanas'eva and S. Rozdil'skiy for their help in managing the data and to the reviewer for most valuable comments which led to a substantial improvement in the manuscript. A 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. The research of Dr. Tsodikov is supported by Grant LO 342/6-1 of the German Research Foundation.

References

Andersen P.K. and B.B. Rønn (1995), A nonparametric test for comparing two samples where all observations are either left- or right-censored, *Biometrics* **51**, 323–329.

Ayer, M., H.D. Brunk, G.M. Ewing, W.T. Reid and E. Silverman (1955), An empirical distribution function for sampling with incomplete information, *Ann. Math. Statist.* **26**, 641–647.

Bergman, S.W. and B.W. Turnbull (1983), Efficient sequential design for destructive life testing with application to animal serial sacrifice experiments, *Biometrika* **70**, 305–314.

Bolshova, O., D. Derevianko and O. Boiarska (1991), Selection of thyroid risk categories among children who were exposed to radioiodines following the Chernobyl NPS disaster, *Health Phys.* **61**, 153.

Cantor, A.B. and J.J. Shuster (1992), Parametric versus nonparametric methods for estimating cure rates based on censored survival data, *Statist. Med.* **11**, 931–937.

Dewanji, A. and J.D. Kalbfleisch (1986), Nonparametric methods for survival/sacrifice experiments, *Biometrics* **42**, 325–341.

Dinse, G.E. (1988), Estimating tumor incidence rates in animal carcinogenicity experiments, *Biometrics* **44**, 405–441.

Dinse, G.E. and S.W. Lagakos (1982), Nonparametric estimation of lifetime and disease onset distributions from incomplete observations, *Biometrics* **38**, 921–932.

Hoang, T., A. Tsodikov, A. Yakovlev and B. Asselain (1995), Modeling breast cancer recurrence, in: O. Arino, D. Axelrod, M. Kimmel, eds., *Mathematical Population Dynamics: Analysis of Heterogeneity, Vol.* 2, *Carcinogenesis and Cell and Tumor Growth* (Wuerz Publications, Winnipeg, Manitoba, Canada) pp. 283–296.

Ivankov, A., T. Hoang, M. Loeffler, A. Tsodikov and A. Yakovlev (1992), A distribution of clonogens progression time – A computer simulation study, in: B. Bru, C. Huber, B. Prum, eds., *Statistique des Processus en Milieu Médical* (Paris, Université René Descartes) pp. 287–294.

Klebanov, L.B., S.T. Rachev and A.Yu. Yakovlev (1983), A stochastic model of radiation carcinogenesis: latent time distributions and their properties, *Math. Biosci.* **113**, 51–75.

Kodell, R.L., G.W. Shaw and A.M. Johnson (1982), Nonparametric joint estimators for disease resistance and survival functions in survival/sacrifice experiments, *Biometrics* **38**, 43–58.

Likhtarev, I.A., B.G. Sobolev, I.A. Kairo, N.D. Tronko, T.I. Bogdanova, V.A. Oleinic, E.V. Epshtein and V. Beral (1995), Thyroid cancer in the Ukraine, *Nature* **375**, 365.

Miller, R.G. (1981), *Survival Analysis* (Wiley, New York).

Pepe, M.S. and T.R. Fleming (1989), Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data, *Biometrics* **45**, 497–507.

Portier, C. (1986), Estimating the tumor onset distribution in animal carcinogenesis experiments, *Biometrika* **73**, 371–378.

Tsodikov, A.D. (1995), The empirical distribution function with incomplete data, to appear in: *Comput. Statist. Data Anal.*

Turnbull, B.W. (1974), Nonparametric estimation of a survivorship function with doubly censored data, *J. Amer. Statist. Assoc.* **69**, 169–173.

Turnbull, B.W. (1976), The empirical distribution function with arbitrarily grouped, censored and truncated data, *J. Roy. Statist. Soc., Sci., B* **38**, 290–295.

Turnbull, B.W. and T.J. Mitchell (1984), Nonparametric estimation of the distribution of time to onset for specific diseases in survival/sacrifice experiments, *Biometrics* **40**, 41–50.

Yakovlev, A.Yu. (1994), Letter to the Editor, *Statist. Med.* **13**, 983–986.

Yamaguchi, K. (1992), Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of permanent employment in Japan, *J. Amer. Statist. Assoc.* **87**, 284–292.