

A New Approach to Testing for Sufficient Follow-up in Cure-Rate Analysis

By LEV B. KLEBANOV

Department of Probability and Statistics, Charles University, and Institute of Informatics and Control of the National Academy of Sciences, Sokolovska 83, Praha-8, CZ-18675, Czech Republic
levkleb@yahoo.com

AND ANDREI Y. YAKOVLEV

Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, NY 14642, U.S.A.
Andrei_Yakovlev@urmc.rochester.edu

Summary

The problem of sufficient follow-up arises naturally in the context of cure rate estimation. This problem was brought to the fore by Maller & Zhou (1992, 1994) in an effort to develop nonparametric statistical inference based on a binary mixture model. The authors proposed a statistical test to help practitioners decide whether or not the period of observation has been long enough to detect the presence of cured (immune) individuals in the study population. The test is inextricably entwined with estimation of the cure probability by the Kaplan-Meier estimator at the point of last observation. While intuitively appealing, the test by Maller & Zhou does not provide a satisfactory solution to the problem because of its unstable and non-monotonic behavior when the duration of follow-up increases. The present paper introduces an alternative concept of sufficient follow-up allowing derivation of a lower bound for the expected proportion of immune subjects in a wide class of cure models. By building on the proposed bound, a new statistical test is designed to address the issue of the presence of immunes in the study population. The usefulness of the proposed approach is illustrated with an application to survival data on breast cancer patients identified through the NCI Surveillance, Epidemiology and End Results Database.

Some key words: Cure models; Generalized hazard; Statistical test; Sufficient follow-up; Survival analysis.

1. INTRODUCTION

The probability of cure, variously referred to as the cure rate or the surviving fraction, is defined as an asymptotic value of the survival function as time tends to infinity. Let X denote observed survival time. Statistical inference on cure rates relies on the fact that any improper survival function $S(t) = \text{pr}\{X \geq t\}$ can be represented in the form:

$$S(t) = a + (1 - a)S_o(t), \tag{1}$$

where $a = \text{pr}\{X = \infty\}$ is the probability of cure, and $S_o(t)$ is defined as the survival function for the time to failure conditional upon ultimate failure, i.e. $S_o(t) = \text{pr}\{X \geq t | X < \infty\}$. This representation is known in the literature as the (binary) mixture model. An alternative, but equally general, representation of an improper survival time distribution is given by the bounded cumulative hazard (BCH) model. The starting point for the development of the BCH model is the assumption that the cumulative hazard $\Lambda(t) = \int_0^t \lambda(t)dt$ has a finite positive limit, say θ , as t tends to infinity. In this case, one can write

$$S(t) = e^{-\theta F(t)}, \quad \theta > 0, \quad t \geq 0, \tag{2}$$

where $F(t) = \Lambda(t)/\theta$ is the cumulative distribution function of some non-negative random variable such that $F(0) = 0$. The probability of cure is equal to $\exp(-\theta)$ under this model. Both models and associated statistical methods are discussed at length in a recent paper by Tsodikov et al (2003).

Proceeding from representation (1), Maller & Zhou (1996) developed a theory of nonparametric estimation of the probability a . Suppose that the survival time distribution $G(t) = 1 - S(t)$ is absolutely continuous and let $t_1 < t_2 < \dots < t_n$ be a sample (subject to right censoring) of the ordered observed failure times. Maller & Zhou suggested estimating a by

$$\hat{a}_n = \hat{S}_n(t_n), \tag{3}$$

where $\hat{S}_n(t)$ is the Kaplan-Meier estimator (KME) of the underlying survival function $S(t)$. This nonparametric estimation of a was also advocated by other authors (Laska & Meisner, 1992; Sposto et al., 1992). The conditional survival function $S_o(t)$ is then estimated by

$$\hat{S}_{0n}(t) = (\hat{S}_n(t) - \hat{a}_n)/(1 - \hat{a}_n), \quad t \geq 0. \tag{4}$$

The model given by (1) is unidentifiable within the nonparametric framework and additional restrictions on the class of survival functions S_o are needed to make inference on the parameter

a. Maller & Zhou (1992) introduced a restriction in terms of the survival and censoring time distributions to address both the consistency question for the estimator \hat{a}_n and the issue of sufficiency of follow-up. Assuming the independent random censorship model with $0 < a \leq 1$ and the censoring time cumulative distribution function $C(t)$, Maller & Zhou (1992) proved that, under mild conditions, the estimator \hat{a}_n is consistent if and only if

$$\tau_{G_o} \leq \tau_C, \tag{5}$$

where τ_{G_o} and τ_C are the right extremes of $G_o(t) = 1 - S_o(t)$ and $C(t)$, respectively. This condition restricts the class of S_o to provide identifiability of the model. Maller & Zhou (1992) also suggested that the inequality (5) may be thought of as a quantification of “sufficient follow-up”. Indeed, if this condition is not true, then failures may occur after the maximum follow-up period and it is not possible to determine which proportion of the late censored data has been generated by cured subjects. As far as biomedical applications are concerned, this condition does not have any clear biological meaning (see Section 6 for further discussion).

Maller & Zhou (1994) proposed a statistical test of (5) based on the length $t_n - t_n^*$ of the interval between the largest uncensored failure time t_n^* and the largest overall failure time t_n . Intuitively, if this interval is large then the last failure has occurred well before the last censoring event so there has been sufficient follow-up. Let N_n be the number of uncensored failure times in the interval $(2t_n^* - t_n, t_n^*]$. Maller & Zhou showed that, under appropriate regularity conditions, the estimator $\alpha_n = (1 - N_n/n)^n$ is an approximate p -value for a test of $\tau_{G_o} \leq \tau_C$, and α_n converges to zero in probability if and only if $\tau_{G_o} \leq \tau_C$.

The original version of the test (α_n -test) by Maller & Zhou suggests the hypothesis of sufficient follow-up to be rejected whenever the α_n -value exceeds a prespecified critical value, say 0.05. This test does not control the significance level but rather estimates it.

Later, after conducting computer simulations, Maller & Zhou (1996) came to the conclusion that the α_n -test was far too conservative. They proposed a modification based on percentiles of the sampling distribution of the closely related statistic $q_n = N_n/n$. Unfortunately, this distribution is not known even in large samples. To surmount this obstacle, Maller & Zhou (1996) resorted to simulations based on certain parametric assumptions about survival and censoring time distributions. Specifically, they took G_o as the exponential distribution with mean 1 and C as a uniform distribution over the interval $[0, B]$ with several possible values of B . They generated percentiles of the sampling distribution of q_n covering a certain range of unknown parameters incorporated into the assumed parametric model. This approach can hardly

be a solution to the problem because it requires restrictive parametric assumptions, including the unknown values of a and B , thereby depriving the proposed test of all nonparametric advantages.

Any test designed to address the issue of sufficiency of follow-up is expected to display a monotonic behavior: the longer the follow-up, the more likely it is sufficient for making inferences about cure rates. We evaluated the performance of the asymptotic α_n -test in an application to survival data on female patients with breast cancer identified through the NCI Surveillance, Epidemiology and End Results Database. The women were divided into groups based on summary stage at diagnosis (localized, regional, or distant) and age at diagnosis (below 46, 46-55, 56-65, and 66 or older). The total follow-up period was equal to 30 years. We introduced an artificial censoring bound, t_c , so that all women who survived by the time t_c were considered as censored observations. Moving t_c to the right, we modeled periods of observation of various lengths and then computed the α_n -values for each of them. Shown in the first two columns of Table 1 are the results of this study for one category of patients (46-55 years old with localized breast cancer) characterized by a relatively high survival rate. This group includes 29,592 women with breast cancer. There are 8,947 uncensored observations in the data set. From Table 1, it is clear that the test falls far short of the desirable property of monotonicity even in this large sample study. The same behavior of the test was observed in other groups of patients.

Maller & Zhou (1996) conclude the exposition of their test as follows: “We do not expect q_n to be the last word, however. It is only a rough test which could be improved on. As a simple count of numbers of uncensored observations, it probably shares with similar nonparametric tests a lack of power in detecting alternatives... Test statistics similar in nature to q_n which use information in the KME relevant to its levelling at its right extreme more effectively are probably preferable”. In contrast to this opinion, we think that such an approach can hardly be improved on because it deals with observations close to the end of an observation period where the KME and related statistics are extremely unstable in the presence of censoring; see Pepe & Fleming (1989) Cantor & Shuster (1992) and Tsodikov (2001). Another snag is that the KME, as an estimated upper bound for a , is not directly relevant to the main problem in question. Indeed, what we would like to know is whether the follow-up has been long enough to provide strong evidence for the presence of immunes in the study population. Given the duration of follow-up, the corresponding statistical hypothesis is formulated as $H_o: a > 0$ (Maller & Zhou, 1995). Testing this hypothesis does not necessarily call for estimating a by means of formula

(3). In other words, condition (5) is relevant to testing the hypothesis H_o only insofar as the associated statistical test bears on the estimator given by formula (3).

In this paper, we suggest that the whole concept of sufficient follow-up be changed in order to overcome the above-discussed difficulties. Our new concept disregards condition (5) altogether while focusing solely on the hypothesis H_o . In other words, we no longer expect the proposed test to tell us whether a given follow-up is sufficient to provide a point estimate of the probability a . However, another minimal restriction on the class of proper distributions G_o is needed to make the model identifiable. We formulate such a restriction in terms of a new notion of φ -hazard rate that facilitates the construction of the proposed test and provides a more general mathematical framework for the problem under discussion.

2. A GENERALIZED HAZARD FUNCTION

Let $\varphi(u)$ be a non-negative strictly monotonically decreasing function defined for all $u \geq 0$. Suppose in addition that its first derivative φ' is continuous and $\varphi(0) = -\varphi'(0) = 1$. We define the φ -hazard rate $r(t)$ for the survival function $S(t)$ by the following relation

$$r(t) = \frac{d}{dt}\varphi^{-1}(S(t)), \quad (6)$$

where φ^{-1} is the inverse function for φ . It is clear that the function $r(t)$ thus defined reduces to the traditional hazard rate $\lambda(t)$ if one chooses φ of the form $\varphi(u) = e^{-u}$. Now it is worth discussing some useful properties of the φ -hazard rate.

Proposition 1. Suppose that the φ -hazard rate for a survival function $S(t)$ is non-decreasing. Then for any real values $x \geq t > 0$, we have

$$S(x) \leq \varphi\left(x \frac{\varphi^{-1}(S(t))}{t}\right). \quad (7)$$

Proof. Since the function $r(t)$ is non-decreasing, the function $\varphi^{-1}(S(t))$ is convex. In view of this fact and the condition $\varphi^{-1}(S(0)) = \varphi^{-1}(1) = 0$, we see that

$$\frac{\varphi^{-1}(S(x))}{x} \geq \frac{\varphi^{-1}(S(t))}{t}, \quad 0 < t \leq x \quad (8)$$

so that (7) follows from (8) and monotonicity of the function φ .

A similar argument yields the following property of the φ -hazard rate.

Proposition 2. Suppose that the φ -hazard rate for a survival function $S(t)$ is non-increasing. Then for any real values $x \geq t > 0$, we have

$$S(x) \geq \varphi\left(x \frac{\varphi^{-1}(S(t))}{t}\right). \quad (9)$$

It is the property (7) that will be exploited in the construction of the proposed test. Assuming this property amounts to stating that the survival function $S(t)$ has increasing (non-decreasing) φ -hazard rate average. This condition allows us to consider a general class of distributions that includes all distributions with increasing ordinary hazard rate average, also known as the IFRA class of distributions (Barlow & Proschan, 1981), as a sub-class. Proposition 1 provides a sufficient condition for a distribution to belong to this class. In what follows, we do not use this result but rather assume that condition (7) is met, thereby specifying the required class of distributions.

3. TESTING FOR SUFFICIENT FOLLOW-UP: TYPE 1 CENSORING

To illustrate the basic idea, we begin by considering the simplest case of Type 1 censoring mechanism (Kalbfleisch & Prentice 2002). A more interesting and practical case of random censorship will be discussed in Section 4. Within the nonparametric framework, the models given by (1) and (2) are equivalent, and we will use (1) as the starting point for simplicity. All the results that follow can be reformulated in terms of model (2). Now we adopt the following

Basic assumption: *The proper survival function $S_o(t)$ in the mixture model (1) has non-decreasing φ -hazard rate average.*

In the case of $\varphi(u) = e^{-u}$, the basic assumption amounts to stating that the survival function $S_o(t)$ has non-decreasing hazard rate average, which is quite plausible in many applications. This assumption is discussed further in the final section.

Let $t_1^* < t_2^* < \dots < t_k^* \leq T$ be uncensored failure times, where T stands for the end of follow-up. In this case, the value $S(T) = a + (1-a)S_o(T)$ is consistently estimated by $\hat{S}(T) = (n-k)/n$, where n is the total number of observations. We wish to test the null hypothesis

$$\mathbf{H}_o : a = 0. \tag{10}$$

Proceeding from the Basic Assumption we first derive a lower bound for the probability a . Let us choose a time point $t_o \in (0, T)$, which is an inner point of the support of $S_o(t)$, and consider

$$R(t_o) = \frac{1}{t_o} \varphi^{-1}(S(t_o)). \tag{11}$$

In like manner, we introduce

$$R_o(t_o) = \frac{1}{t_o} \varphi^{-1}(S_o(t_o)). \tag{12}$$

From formula (1) we have

$$S_o(T) \leq \varphi(TR_o(t_o)) \leq \varphi(TR(t_o)) \quad (13)$$

and

$$S(T) \leq \varphi(TR(t_o)) + a[1 - \varphi(TR(t_o))].$$

Therefore, we obtain

$$a \geq \frac{S(T) - \varphi(TR(t_o))}{1 - \varphi(TR(t_o))}. \quad (14)$$

Denote by $S_n(t)$ the empirical counterpart of the survival function $S(t)$. The function $\varphi(TR(t_o))$ is consistently estimated by $\varphi(TR_n(t_o))$, with

$$R_n(t_o) = \frac{1}{t_o} \varphi^{-1}(S_n(t_o)). \quad (15)$$

Then the statistic

$$a_n^* = \max \left\{ 1 - \frac{k}{n[1 - \varphi(TR_n(t_o))]}, \quad 0 \right\} \quad (16)$$

is a consistent estimator for the right-hand side of inequality (14) and we can claim that

$$\text{pr}\{a_n^* \leq a\} \rightarrow 1, \quad \text{as } n \rightarrow \infty. \quad (17)$$

The estimator a_n^* is of interest in its own right even if it is not directly involved in the testing procedure proposed below.

The choice of t_o will be addressed in Section 6. Leaving aside this question, we are in a position to design a statistical test for the hypothesis \mathbf{H}_o presented in formula (10). The following hypothesis

$$\mathbf{H}'_o : S(T) = S_o(T) \quad (18)$$

is obviously equivalent to \mathbf{H}_o . An appropriate test can be designed by constructing a conservative lower confidence limit for the difference: $S(T) - S_o(T)$. By virtue of the Basic Assumption, we can write

$$\varphi\left(\frac{T}{t_o} \varphi^{-1}(S_o(t_o))\right) \geq S_o(T). \quad (19)$$

We make inequality (19) stronger by writing

$$\varphi\left(\frac{T}{t_o} \varphi^{-1}(S(t_o))\right) \geq S_o(T) \quad (20)$$

in order to construct a lower confidence limit for

$$\Delta(T) = S(T) - \varphi\left(\frac{T}{t_o} \varphi^{-1}(S(t_o))\right). \quad (21)$$

Now we need one more mild assumption. Specifically, we assume that the function

$$\psi(u) = \varphi \left(\frac{T}{t_o} \varphi^{-1}(u) \right), \quad 0 \leq t_o \leq T \quad (22)$$

is convex in u . This assumption is met in the special case of the traditional hazard function, i.e. $\varphi(u) = e^{-u}$. Then we have

$$\begin{aligned} & \left| \varphi \left(\frac{T}{t_o} \varphi^{-1}(S_n(t_o)) \right) - \varphi \left(\frac{T}{t_o} \varphi^{-1}(S(t_o)) \right) \right| = |\psi(S_n(t_o)) - \psi(S(t_o))| \\ & \leq \max_{0 \leq u \leq 1} |\psi'(u)| |S_n(t_o) - S(t_o)| = \psi'(1) |S_n(t_o) - S(t_o)| = \frac{T}{t_o} |S_n(t_o) - S(t_o)| \end{aligned} \quad (23)$$

by monotonicity of ψ' and the conditions: $\varphi(0) = -\varphi'(0) = 1$. Since $S(t)$ is continuous, recourse can be made to the Kolmogorov goodness-of-fit statistic to contend that

$$\text{pr} \left\{ n^{\frac{1}{2}} \sup_{t \geq 0} |S_n(t) - S(t)| \leq D_\alpha(n) \right\} = 1 - \alpha, \quad 0 < \alpha < 1$$

where $D_\alpha(n)$ is the $(1 - \alpha)$ th percentile of the Kolmogorov distribution for a sample of size n . Therefore, we have

$$\text{pr} \left\{ n^{\frac{1}{2}} \left| \frac{n-k}{n} - S(T) \right| \leq D_\alpha(n) \right\} \geq 1 - \alpha. \quad (24)$$

Recalling (21), (23) and (24), we finally obtain

$$\text{pr} \left\{ \Delta(T) \geq \frac{n-k}{n} - \varphi \left(\frac{T}{t_o} \varphi^{-1}(S_n(t_o)) \right) - \left[1 + \frac{T}{t_o} \right] \frac{D_\alpha(n)}{n^{\frac{1}{2}}} \right\} \geq 1 - \alpha, \quad (25)$$

where $\Delta(T)$ is given by formula (21). If the lower confidence limit in (25) is greater than 0, the hypothesis \mathbf{H}'_0 is rejected at a significance level of less than α .

4. RANDOMLY CENSORED SAMPLES

Only an asymptotic inference is tractable in the case of randomly censored data. We will proceed from the right random censorship model as described in Kalbfleisch & Prentice (2002) and general results by Hall & Wellner (1980). Let X_1, \dots, X_n be independent positive random variables with continuous distribution function G and survival function $S = 1 - G$. Let Y_1, \dots, Y_n be independent positive random variables having left-continuous cumulative distribution function C , and suppose that the X 's and Y 's are stochastically independent. We observe the n pairs (\tilde{X}_i, δ_i) , with $\tilde{X}_i = \min(X_i, Y_i)$, and δ_i the indicator function of $X_i \leq Y_i$, $i = 1, \dots, n$. In the usual fashion, we use these data to construct the Kaplan-Meier product-limit estimate \hat{S}_n of the underlying survival function S .

Let T be the end of follow-up. Suppose $T_H = \inf \{t \geq 0 : H(t) = 1\} \leq \infty$, where $H = 1 - S(1 - C)$. Let us now define

$$\hat{R}_n(t_o) = \frac{1}{t_o} \varphi^{-1}(\hat{S}_n(t_o))$$

and

$$\hat{a}_n^* = \max \left\{ 1 - \frac{1 - \hat{S}_n(T)}{1 - \varphi(T \hat{R}_n(t_o))}, \quad 0 \right\}, \quad (26)$$

where $T < T_H$. If G and C are continuous and $T < T_H$ with $H(T) < 1$, then $\hat{S}_n(t) \rightarrow S(t)$ as $n \rightarrow \infty$ in probability for all $0 \leq t \leq T$ so that the result (17) remains valid for $T < T_H$.

Following Hall & Wellner (1980), introduce the notation

$$A_n(t) = n \sum_{\{i: \tilde{X}_i < t\}} \frac{\delta_i}{(n-i)(n-i+1)},$$

which is equivalent to their $C_N(t)$. From the basic result (Theorem 2, page 137) by Hall & Wellner (1980) it follows that

$$\lim_{n \rightarrow \infty} \text{pr} \left\{ \sup_{0 \leq t \leq T < T_H} \left(n^{\frac{1}{2}} |S(t) - \hat{S}_n(t)| \leq D_\nu \hat{S}_n(t) (1 + A_n(t)) \right) \right\} > 1 - \nu,$$

where $D_\nu = \lim_{n \rightarrow \infty} D_\nu(n)$ is the $(1-\nu)$ th percentile of the asymptotic Kolmogorov distribution.

By the argument of Section 3, we arrive at the following asymptotic lower confidence limit for $\Delta(T)$:

$$\text{pr} \left\{ \Delta(T) \geq \hat{S}_n(T) - \varphi \left(\frac{T}{t_o} \varphi^{-1} \left(\hat{S}_n(t_o) \right) \right) - \left[1 + \frac{T}{t_o} \right] \frac{D_\alpha}{n^{\frac{1}{2}}} \hat{S}_n(t_o) (1 + A_n(t_o)) \right\} > 1 - \alpha, \quad (27)$$

as $n \rightarrow \infty$, given $T < T_H$ and the assumptions adopted in the previous section are all met. Further results by Hall & Wellner (1980) provide a sharper bound (by reducing D_α) but its construction is computationally more expensive.

5. SHARPENING THE LOWER BOUND FOR THE PROBABILITY OF CURE

Proceeding from formula (4), let us define an infinite sequence of parameters a_k as follows

$$a_1 = \frac{S(T) - \psi(S(t_o))}{1 - \psi(S(t_o))},$$

$$a_k = \frac{S(T) - \psi(S_{k-1}(t_o))}{1 - \psi(S_{k-1}(t_o))},$$

where

$$S_k(t) = \frac{S(t) - a_k}{1 - a_k}, \quad k = 1, 2, \dots,$$

and

$$\psi(z) = \varphi\left(\frac{T}{t_o}\varphi^{-1}(z)\right).$$

It is clear that $a_1 \leq a_2 \leq \dots \leq a_k \leq \dots \leq a$, where $a \geq 0$ is the probability of cure. It is also obvious that $S_k(t) \geq S_o(t)$ for all $t \geq 0$ and any fixed k .

Then we have the following recurrence relation

$$a_k = \frac{S(T) - \psi\left(\frac{S(t_o) - a_{k-1}}{1 - a_{k-1}}\right)}{1 - \psi\left(\frac{S(t_o) - a_{k-1}}{1 - a_{k-1}}\right)}. \quad (28)$$

The sequence a_k , $k = 2, \dots$, is bounded and therefore it has a limit b^* satisfying the equation

$$b^* = \frac{S(T) - \psi\left(\frac{S(t_o) - b^*}{1 - b^*}\right)}{1 - \psi\left(\frac{S(t_o) - b^*}{1 - b^*}\right)}.$$

A consistent statistical estimator for b^* can be found as a solution \hat{b}_n^* of the following equation

$$\frac{\hat{S}_n(T) - \hat{b}_n^*}{1 - \hat{b}_n^*} = \psi\left(\frac{\hat{S}_n(t_o) - \hat{b}_n^*}{1 - \hat{b}_n^*}\right). \quad (29)$$

This equation can be solved numerically by the iterative procedure suggested by (28). The resultant estimator generally provides a sharper bound for a than the estimator \hat{a}_n^* (see Sections 6 and 7).

6. DATA ANALYSIS

Using the proposed test we analyzed the same set of data on breast cancer patients as described in Section 1. The analysis was designed along similar lines. We used the traditional hazard rate $\lambda(t)$ for $r(t)$, i.e. the function $\varphi(u)$ was chosen in the form: $\varphi(u) = e^{-u}$. The parameter t_o was initially chosen to belong to the interval $(0, T/2]$. This parameter was then optimized by maximizing the estimated lower bound \hat{a}_n^* given by formula (26). Maximizing \hat{a}_n^* does not affect our probabilistic argument used to obtain the lower confidence limit (27) because the latter limit is based on the Kolmogorov distance between $S(t)$ and $S_n(t)$, being uniformly valid for all values of $t > 0$. This data-driven choice of t_o allows us to increase the power of the test while providing control of the significance level.

The results of testing for sufficient follow-up for different values of the observation time t_c are presented in the fourth column of Table 1 (see Section 1). It is seen from the table that

the results are quite stable and meet the monotonicity requirement discussed in Section 1. The behavior of the estimated lower bound \hat{a}_n^* for the probability of cure a and of the estimate \hat{a}_n as functions of t_c can be inferred from Table 1. As a better illustration, we provide the corresponding plots in Figure 1. It is clear that the longer the follow-up the higher the lower bound \hat{a}_n^* and the more confident one can be about its sufficiency. The behavior of the KME is opposite to that of \hat{a}_n^* , which is where the main difference between the two approaches lies. A sharper lower bound is provided by the estimator \hat{b}_n^* (Section 5) which can be obtained as a numerical solution of equation (29). In the example presented above, the value of \hat{b}_n^* for $T = 360$ months is equal to 0.447 (compare to $\hat{a}_n^* = 0.317$), which is pretty close to the value of 0.469 provided by the KME at this late point (see Table 1).

When considering cause-specific survival of cancer patients, the concept of cure is biologically natural: it refers to those patients that have no distant metastases at the time of diagnosis and all clonogenic cells in their primary tumor are killed by treatment (Yakovlev & Tsodikov, 1996). The eventual death of a susceptible (incurable) patient is caused either by the propagation of primary tumor cells giving rise to local recurrence or by metastatic spread of the tumor, both processes being irreversible and cumulative in nature. Thus, it is also natural to assume that the survival function S_o defined for susceptible individuals has non-decreasing hazard average. While defining a very rich class of distributions, this property, is generally not preserved under the operation of forming a mixture (Barlow & Proschan, 1981). However, there are examples where the increasing hazard rate property is preserved under mixtures (Shaked and Spizzichino, 2001). Lynch (1999) provides some general conditions under which this property holds. It is obvious that one cannot model unobservable heterogeneity within the nonparametric framework. Therefore, it is advisable to identify study subjects so that their population is as homogeneous as possible when using the traditional hazard rate to construct the lower confidence bound given by formula (27). This can be accomplished by stratification of the study population by major clinical covariates as we did in the above analysis of breast cancer data. Some advantages offered by the concept of φ -hazard rate in studies of non-homogeneous groups of subjects are discussed in Section 8.

7. POWER PROPERTIES OF THE TEST: A SIMULATION STUDY

As seen in Table 1, the proposed test appears to be less conservative than the test by Maller & Zhou as it rejects the null hypothesis more frequently in this large sample study. To explore the power of the test in more detail, we conducted a simulation study. The survival time

distribution was generated by the mixture given by formula (1), where the function $S_0(t)$ was chosen to be a Weibull survival function, $W(\beta, \gamma)$, with shape parameter β and scale parameter γ . The parameter γ was set equal to 1 throughout all simulations. The censoring time was uniformly distributed over $[0, B]$. The test was applied to the simulated data at the point of the largest (censored) observation. The relationship between the parameters β and B controls the expected proportion of censored observations. For example, this proportion is equal to 0.55 when using $W(1.5, 1.0)$, $U[0, 1.5]$, and it is equal to 0.79 when using $W(1.5, 1.0)$, $U[0, 0.75]$. The significance level $\alpha = 0.05$ was used in conjunction with the lower confidence limit given by formula (27).

In the first set of simulations, 1000 samples of size n were generated for different values of the parameters a , β , B , and n . The test was applied to each sample at a significance level of 0.05 and the proportion of rejections was recorded. Although the test is conservative by design, the results shown in Tables 2 and 3 indicate that its power is quite high under the conditions of our simulations study. It is worth noting that the rejection probabilities for $a = 0$ in Tables 2 and 3 cannot be interpreted as the usual nominal significance level because the test is based on a lower bound for a parameter and not on the sampling distribution of a test statistic. This is also the reason why Type 1 error rate decreases concurrently with an increase in power when the sample size and/or the total follow-up increase. As evident from Tables 2 and 3, the power exceeds 80% when $a \geq 0.2$ and the sample size is as small as 500. Even for such a small departure from the null hypothesis as in the case of $a = 0.1$, the power is still greater than 70% with $n = 500$. The sample size $n = 200$ is typically considered far too small for cure-rate analysis and yet it provides a reasonably high power ($> 60\%$) even for such a close alternative as $a = 0.1$. As one would expect, the power tends to increase with the value of a and the sample size, and it decreases with the depth of censoring.

The second set of simulation experiments was concerned with the accuracy of \hat{b}_n^* which estimator is used as a sharp empirical lower bound of the cure probability a . These experiments are much more time-consuming than the above-described power studies, and we limited the number of simulations to 500. Since the lower bound becomes equal to the probability a when the survival function $S_o(t)$ is exponential, the Weibull distribution of the survival time is an appropriate choice to study departures from the baseline case. For $n = 500$, we studied two shapes of the Weibull hazard and two values of the censoring parameter B ($B = 1.0$ and $B = 1.5$) for each of the following three values of the cure probability: $a = 0.4, 0.5$ and 0.6 . Table 4 reports the mean value, M_b , of the estimator \hat{b}_n^* and the corresponding standard

deviation, σ_b , estimated from 500 simulations for each combination of the parameters a , β , and B . Based on the standard deviation σ_b as a measure of accuracy, the estimator \hat{b}_n^* of the lower bound for the probability a appears to be quite accurate and stable to the proportion of censored data.

To assess the effect of the sample size, we conducted a separate simulation experiment with $a = 0.4$ and $B = 1$. The results presented in Table 5 show that the estimated lower bound does not change much when the sample size increases from $n = 500$ to $n = 1,000$. We would like to note that the lower bound for a , estimated by \hat{b}_n^* , does not need to be very sharp for the proposed statistical test to perform well as discussed further in the next section.

8. DISCUSSION

As our discussion in Section 1 shows, the test by Maller & Zhou (1994, 1996) does not provide a satisfactory solution to the problem of testing for sufficient follow-up because of its unstable and non-monotonic behavior when the duration of follow-up increases. This also diminishes the utility of the KME as a point estimator of the probability of cure because condition (5) is difficult to test. However, the KME can be used as a nonparametric upper bound for this probability while the estimator \hat{a}_n^* (or \hat{b}_n^*) provides an appropriate lower bound (see Figure 1). It is important to note that both the KME and \hat{a}_n^* are consistent estimators of the upper and the lower bounds for the cure probability, respectively, for any value of T . The example presented in Sections 1 and 5 clearly demonstrates that the newly designed statistical test overcomes the conceptual difficulties discussed in Section 1. One may envision possible ways to further increase its power: resorting to a more precise result by Hall & Wellner (1980) is just one of them.

We would like to emphasize that the interpretation of the proposed test is narrower than that of the test by Maller & Zhou (1994; 1996). Our test addresses only the question about the presence of immunes in the study population. The statistical test for the hypothesis (10) and the estimation of the lower bound for the probability of cure are separate issues. For a good performance of the proposed test, the lower bound for a does not need to be very close to the corresponding true value as long as the power of the test is sufficiently high. The comparative analysis presented in Table 1 and the power studies of Section 7 indicate that the proposed test is less conservative than the test by Maller & Zhou.

In our application, we relied on the biological rationale given in Section 6 when choosing $\varphi(u) = e^{-u}$. This choice worked well for our purposes as evidenced by the monotonic behavior

of the test and some other considerations below. Let us consider the utility of other possible forms of the function $\varphi(u)$. If the function $S_o(t)$ were known, it would be computationally straightforward to find a pertinent transformation. Let X_o be a random variable with cumulative distribution function $G_o = 1 - S_o$. All we need is to compute the inverse G_o^{-1} of the cumulative distribution function G_o . Then the random variable

$$U = G_o^{-1}(X_o) \tag{30}$$

is uniformly distributed on $[0, 1]$, and the random variable

$$V = -\log(1 - U) \tag{31}$$

has a standard exponential distribution. Within the nonparametric framework, one might think of (30) and (31) as data transformations if a sample from G_o were available. However, we never have a sample from G_o . This brings up the question: How can we ensure that the basic assumption of Section 3 is consistent with the data at hand and how can we use the φ -hazard to correct its violations if necessary? Described below is a simple exploratory method that addresses this issue.

The condition of monotonicity of the hazard rate for $S_o(t)$ is equivalent to the requirement of convexity of $\Gamma(t) = -\log S_o(t)$. If we knew the cure probability a , we could use the equality $S_o(t) = [S(t) - a]/[1 - a]$, thereby formulating the problem in terms of the observable function $S(t)$. We do not know the true a , but one can estimate its upper bound by \hat{a}_n . Shown in Figure 2 is the function $\Gamma(t)$ (solid line) estimated from the data of Section 6 for $a = \hat{a}_n = 0.469$. In this case, $\Gamma(t)$ appears to be linear which is in agreement with the basic assumption in terms of the regular hazard rate. This indicates that our choice of $\varphi(u) = e^{-u}$ in the application presented in Section 6 was quite reasonable. Using, for example, $\varphi(u) = e^{-u^\kappa}$ one can gain more convexity in the corresponding φ -counterpart of $\Gamma(t)$ denoted by $\Gamma_\varphi(t)$. This is shown in Figure 2 where we used $1/\kappa = 1.5$ to construct $\Gamma_\varphi(t)$. If $a = \hat{a}_n$ does not result in a convex curve, one can try some other $a \in (0, \hat{a}_n)$ in an effort to attain these ends. Figure 3 presents the results obtained with $a = 0.458$ and the same value of κ . In general, we recommend trying several values of a before rejecting the convexity property of $\Gamma_\varphi(t)$. With $\varphi(u) = e^{-u^\kappa}$, we guarantee that the class of all distributions with non-decreasing φ -hazard contains all distributions with nondecreasing regular hazard functions as its proper subclass, and therefore this larger class is nonparametric as well.

We would like to emphasize that resorting to a φ -transformation is warranted only if violations of monotonicity are not too strong. The reason for this recommendation is that we want

to minimize potential losses in power that may have been caused by such a transformation. In oncological applications, we recommend resorting to the generalized hazard only to adjust for a possible heterogeneity of susceptible individuals. It is always a good idea to keep the φ -hazard rate as close to a constant as possible.

To gain greater insight into the concept of φ -hazard functions, let us discuss how diverse shapes of the traditional hazard rate can be accommodated in the construction of the lower bound for the probability of cure. A simple expedient of embracing the desired parametric families and a relevant example are discussed below. Note that all the results presented in this paper hold true for all such S_o for which $\psi(S_o(x))/x$ is a non-decreasing function, where $\psi = \varphi^{-1}$. Introduce the notation: $\Psi(t) = S_o^{-1}(x)$, $x \in (0, 1)$. If the inverse function Ψ exists, the requirement that the function $\psi(S_o(x))/x$ be non-decreasing is equivalent to the requirement that the function $\psi(x)/\Psi(x)$, $x \in (0, 1)$, should be non-increasing. Suppose that ψ and Ψ are differentiable. Then we have

$$\frac{\psi'(x)}{\psi(x)} \leq \frac{\Psi'(x)}{\Psi(x)}. \quad (32)$$

Suppose that the function Ψ belongs to some class (family) of functions denoted by \mathcal{F} . If for any $x \in (0, 1)$,

$$\inf_{\Psi \in \mathcal{F}} \frac{\Psi'(x)}{\Psi(x)} > -\infty, \quad (33)$$

the function ψ can be chosen as a solution of the inequality

$$\frac{\psi'(x)}{\psi(x)} \leq \inf_{\Psi \in \mathcal{F}} \frac{\Psi'(x)}{\Psi(x)}, \quad (34)$$

which solution exists as long as the condition (33) is met.

For the function ψ thus chosen, the function $\psi(S_o(x))/x$ will be non-decreasing for any survival function S_o with its inverse belonging to the class \mathcal{F} . If \mathcal{F} is a compact subset of the metric space $C^1(0, 1)$, representing all functions on $(0, 1)$ with continuous first derivative, then the condition (33) is satisfied. In particular, suppose we would like to include all two-component Weibull mixtures of the form

$$S_o(t) = \gamma e^{-at^\alpha} + (1 - \gamma)e^{-bt^\beta}, \quad (35)$$

in the construction of an appropriate φ -hazard function. Suppose in addition that the numerical parameters incorporated into expression (35) satisfy the following conditions: $a \in [a_1, a_2]$, $a_1 > 0$, $a_2 < \infty$, $b \in [b_1, b_2]$, $b_1 > 0$, $b_2 < \infty$, $\alpha \in [\alpha_1, \alpha_2]$, $\alpha_1 > 0$, $\alpha_2 < \infty$, $\beta \in [\beta_1, \beta_2]$, $\beta_1 > 0$, $\beta_2 < \infty$, where $a_1, a_2, b_1, b_2, \alpha_1, \alpha_2, \beta_1, \beta_2$ are fixed numbers. Under such restrictions, the class \mathcal{F} defined by (35) generates a wide variety of shapes of the traditional hazard function. For this

class, the condition given in (33) is met so that one can proceed with the constructive way of choosing the function ψ , and consequently φ , as described above. This approach is instrumental in overcoming nonidentifiability problems that may arise when using nonidentifiable mixtures of parametric families to specify $S_0(t)$.

ACKNOWLEDGMENTS

We are grateful to W. J. Hall and D. Oakes (University of Rochester) for their comments and suggestions. This research was supported by Czech Ministry of Education Grant MSM 113200008 and NIH/NCI grant UO1 CA88177.

REFERENCES

- Barlow, R.E. & Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing*, 2nd ed. Silver Spring: TO BEGIN WITH.
- Cantor, A.B. & Shuster, J.J. (1992). Parametric versus nonparametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine* **11**, 931-37.
- Hall, W.J. & Wellner, J.A. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67**, 133-43.
- Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. New Jersey: Wiley.
- Laska, E.M. & Meisner, M.J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics* **48**, 1223-34.
- Lynch, J.D. (1999). On conditions for mixtures of increasing failure rate distributions to have an increasing failure rate. *Probab. Eng. Inf. Sci.* **13**, 33-36.
- Maller, R.A. & Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika* **79**, 731-39.
- Maller, R.A. & Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. *J. Am. Statist. Assoc.* **89**, 1499-1506.

- Maller, R.A. & Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data. *Biometrics* **51**, 1197-1205.
- Maller, R.A. & Zhou, S. (1996). *Survival Analysis with Long-Term Survivors*, Chichester: Wiley.
- Pepe, M.S. & Fleming, T.R. (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* **45**, 497-507.
- Shaked, M. & Spizzichino, F. (2001). Mixtures and monotonicity of failure rate functions. In *Handbook of Statistics*, Vol. 20, Ed. Balakrishnan, N. and Rao, C.R., 185-198.
- Spoto, R., Sather, H.N. & Baker, S.A. (1992). A comparison of tests of the difference in the proportion of patients who are cured. *Biometrics* **48**, 87-99.
- Tsodikov, A. (2001). Estimation of survival based on proportional hazards when cure is a possibility. In *Modeling and Data Analysis in Cancer Studies*, Ed. Yakovlev, A.Y. and Moolgavkar, S.H., *Math. Comp. Modelling* **33**, 1227-36.
- Tsodikov, A.D., Ibrahim, J.G. & Yakovlev, A.Y. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *J. Am. Statist. Assoc.* **98**, 1063-78.
- Yakovlev, A.Y. & Tsodikov, A.D. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications*, Singapore: World Scientific.

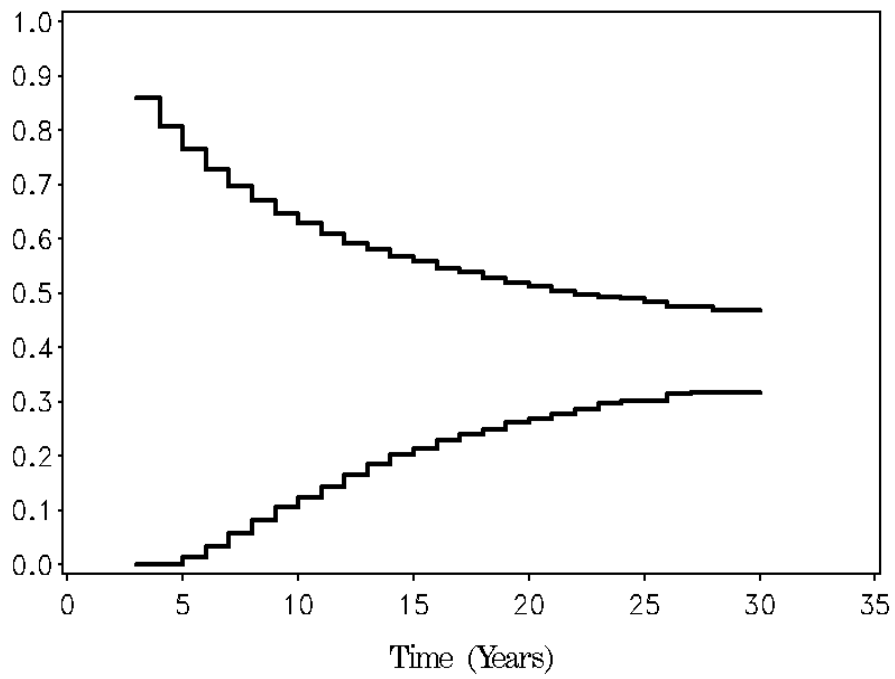


Figure 1: The Kaplan-Meier estimate (upper curve) and the estimated lower bound \hat{a}_n^* for cure probability (lower curve) as functions of the censoring bound t_c .

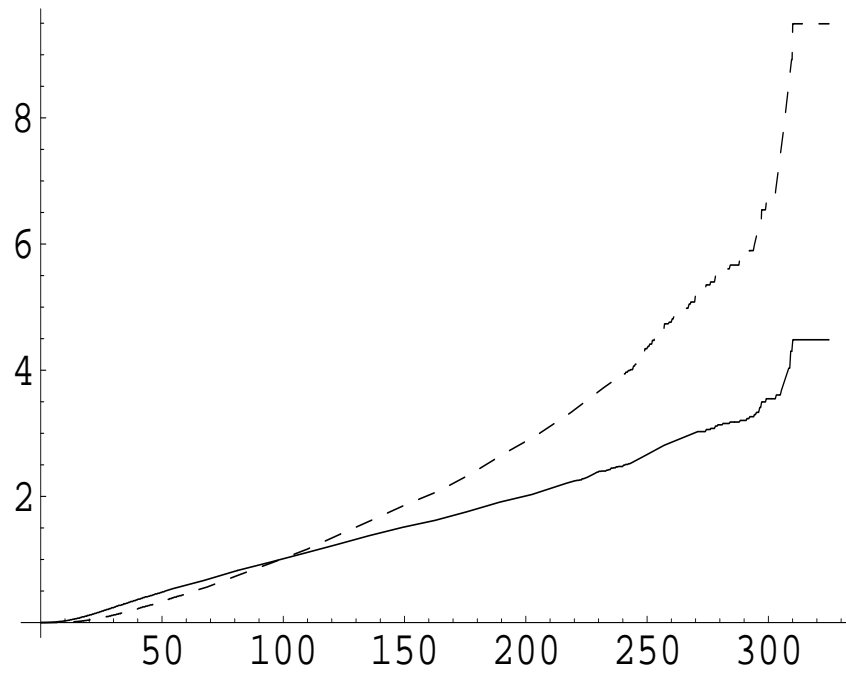


Figure 2: Plots of $\Gamma(t)$ and $\Gamma_f(t)$ constructed from breast cancer data for $a = 0.469$.

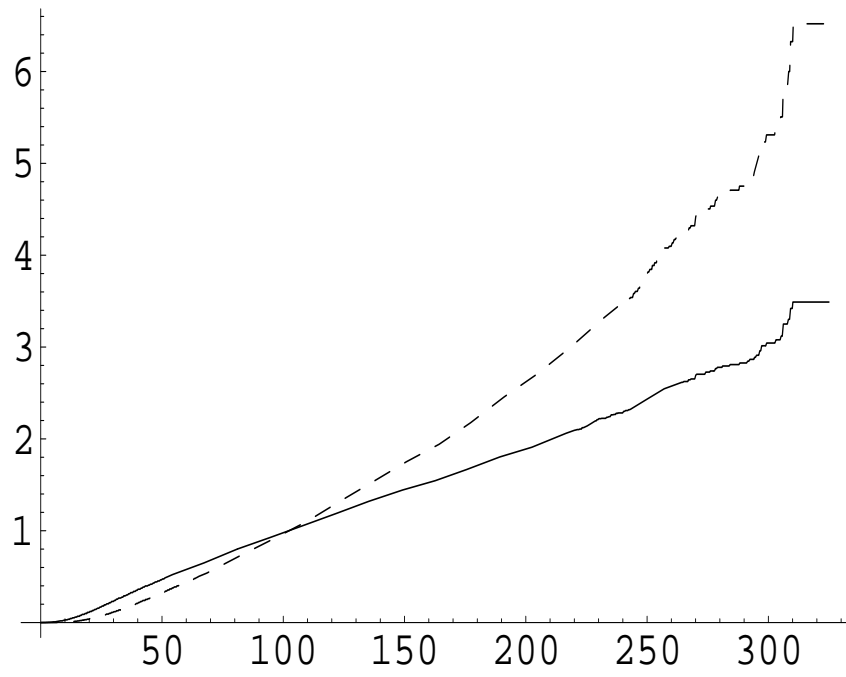


Figure 3: Plots of $\Gamma(t)$ and $\Gamma_f(t)$ obtained from breast cancer data for $a = 0.458$

Table 1: Testing the hypothesis of sufficient follow-up for different values of t_c in the application to breast cancer data. Rejection of the hypothesis is denoted by (-), its acceptance is denoted by (+). See text for further explanations.

Observation time t_c (months)	α_n -test	New test	\hat{a}_n (KME)	\hat{a}_n^*
36	+	-	0.859	0
48	-	-	0.806	0
60	-	-	0.764	0
72	-	+	0.728	0.014
84	-	+	0.696	0.034
96	+	+	0.670	0.058
108	-	+	0.647	0.080
120	-	+	0.628	0.106
132	+	+	0.608	0.123
144	-	+	0.592	0.144
156	+	+	0.580	0.166
168	+	+	0.568	0.185
180	-	+	0.558	0.201
192	-	+	0.546	0.213
204	-	+	0.538	0.229
216	+	+	0.528	0.239
228	+	+	0.519	0.250
240	-	+	0.513	0.262
252	-	+	0.505	0.269
264	-	+	0.498	0.276
276	-	+	0.493	0.287
288	-	+	0.491	0.298
300	-	+	0.484	0.302
312	+	+	0.475	0.303
324	+	+	0.475	0.314
336	-	+	0.469	0.317
348	-	+	0.469	0.317
360	-	+	0.469	0.317

Table 2: Power of the test for the survival time distribution $W(1.2, 1.0)$ and different parameters of the censoring time distribution $U[0, B]$.

a	a=0.0		a=0.1		a=0.2		a=0.3	
B	$n = 200$	$n = 500$	$n = 200$	$n = 500$	$n = 200$	$n = 500$	$n = 200$	$n = 500$
$B = 0.75$	3.7%	1.4%	57%	71%	81%	91%	85%	95%
$B = 1.0$	2.8%	0.4%	63%	71%	85%	88%	89%	91%
$B = 1.5$	0.5%	0.1%	75%	82%	80%	91%	94%	96%

Table 3: Power of the test for the survival time distribution $W(1.5, 1.0)$ and different parameters of the censoring time distribution $U[0, B]$.

a	a=0.0		a=0.1		a=0.2		a=0.3	
B	$n = 200$	$n = 500$	$n = 200$	$n = 500$	$n = 200$	$n = 500$	$n = 200$	$n = 500$
$B = 0.75$	3.9%	2.4%	61%	71%	74%	79%	75%	91%
$B = 1.0$	2.5%	0.3%	65%	76%	78%	82%	80%	89%
$B = 1.5$	0.8%	0.2 %	67%	80%	85%	90%	88%	95%

Table 4: The mean value M_b of the lower bound \hat{b}_n^* and the corresponding standard deviation σ_b for $n = 500$ and different parameters of the survival time and censoring time distributions.

Cure probability	a=0.4		a=0.5		a=0.6	
$U[0, B]$	$W(1.2, 1.0)$	$W(1.5, 1.0)$	$W(1.2, 1.0)$	$W(1.5, 1.0)$	$W(1.2, 1.0)$	$W(1.5, 1.0)$
$B = 1.0$	$M_b = 0.31$ $\sigma_b = 0.14$	$M_b = 0.25$ $\sigma_b = 0.13$	$M_b = 0.36$ $\sigma_b = 0.16$	$M_b = 0.29$ $\sigma_b = 0.14$	$M_b = 0.42$ $\sigma_b = 0.18$	$M_b = 0.34$ $\sigma_b = 0.16$
$B = 1.5$	$M_b = 0.31$ $\sigma_b = 0.13$	$M_b = 0.24$ $\sigma_b = 0.12$	$M_b = 0.38$ $\sigma_b = 0.14$	$M_b = 0.31$ $\sigma_b = 0.13$	$M_b = 0.45$ $\sigma_b = 0.16$	$M_b = 0.36$ $\sigma_b = 0.15$

Table 5: The mean value M_b of the lower bound \hat{b}_n^* and the corresponding standard deviation σ_b for $a = 0.4$, $B = 1.0$, and two different sample sizes.

Survival function	$n = 500$	$n = 1,000$
$W(1.2, 1.0)$	$M_b = 0.32$ $\sigma_b = 0.14$	$M_b = 0.34$ $\sigma_b = 0.15$
$W(1.5, 1.0)$	$M_b = 0.23$ $\sigma_b = 0.14$	$M_b = 0.30$ $\sigma_b = 0.15$