Statistics Surveys
Vol. 18 (2024) 82–138
ISSN: 1935-7516
https://doi.org/10.1214/24-SS147

Mixture cure model methodology in survival analysis: Some recent results for the one-sample case

Ross Maller[1], Sidney Resnick[2], Soudabeh Shemehsavar[*3,4], and Muzhi Zhao[5]

[1]*Research School of Finance, Actuarial Studies & Statistics, The Australian National University, Canberra, Australia*
e-mail: Ross.Maller@anu.edu.au

[2]*School of Operations Research & Information Engineering, Cornell University, Ithaca, N.Y., USA*
e-mail: sir1@cornell.edu

[3]*School of Mathematics, Murdoch University, Perth, Western Australia*

[4]*Department of Mathematics & Statistics, University of Tehran, Tehran, Iran*
e-mail: Soudabeh.Shemehsavar@murdoch.edu.au

[5]*Research School of Finance, Actuarial Studies & Statistics, The Australian National University, Canberra, Australia*
e-mail: Muzhi.Zhao@anu.edu.au

Abstract: The mixture cure model in survival analysis has received large and growing attention in the last few decades. Restricting ourselves mainly to the one-sample case, we present here an overview drawing together some recent significant advances and earlier results, and pointing out areas where further work is needed. New results presented include a discussion of testing for the presence of long term survivors in the null case (when there are no cures present), the probability that an individual is cured (when cures are present), and further analysis of the idea of sufficient followup. Extreme value methods play a key role. We draw attention to some challenging open problems.

MSC2020 subject classifications: Primary 62N01, 62N02, 62N03, 62E10, 62E15, 62E20, G2G05; secondary 62F03, 62F05, 62F12, 62G32.
Keywords and phrases: Censored survival data, cure model, immune or cured individuals, long term survivors, Kaplan-Meier estimator, sufficient followup, probability of being cured, extreme value methods.

Received April 2023.

Contents

*Corresponding author.

1. The mixture cure model

In the analysis of time-to-event data, we often encounter survival curves which plateau (level off) at the right hand end. This may indicate the presence of a proportion of individuals in the population who will not suffer the event, no matter how long they are followed up. We refer to them as "cured of" or "immune to" the cause of the event, and methods are now well developed to deal with this kind of data, generally known as *cure model analysis*. As well as providing significant extra information beyond that of a standard survival analysis, ignoring the presence of cures in an analysis can result in biased and misleading conclusions, sometimes with profound consequences for diagnostic prognostications and evaluations. Cure models have received large and growing attention in the last few decades and it seems timely now to provide a review of their development leading up to present day applications. Various types of cure models have been formulated but here we concentrate on *mixture cure models*.

The first recognition of the need for and implementation of a cure model seems to have been by Boag (1949). He collected data from a number of centres in England, for various sites of the disease and treatment methods, and observed that the distributions of life-lengths (measured from the beginning of treatment) of those dying appeared to follow quite well a lognormal distribution. But, he wrote *"if (a) sample consisted mainly of patients treated while the disease was still in an early and localized form, an analysis made ten years later would yield a distribution of survival times... similar in form to that of the foregoing sample [i.e., to those dying of the disease]... together with a large group of patients who were still alive and symptom-free... In this instance we should conclude that a proportion of the patients was permanently cured by the treatment."*

Accordingly, he proposed a model in which *"A proportion ... of all patients treated is permanently cured. Patients in the remaining fraction ... are liable to die of cancer if they do not previously die from other causes and the survival times of patients in this group follow a lognormal distribution."* He went on to fit by maximum likelihood a lognormal distribution with mass at infinity – a mixture cure model – to followup data on 121 women with breast cancer, finding a significant "cured" proportion in the data.[1] We revisit Boag's data and analysis in later sections.

Since the prospect of a cure is surely the hope of many or most medical procedures, the importance of Boag's insight can hardly be overstated. Following his groundbreaking paper a number of researchers, including Berkson and Gage (1952), Haybittle (1965), Farewell (1977a,b, 1982, 1986), Pocock et al. (1982), Goldman (1984, 1991), Rutquist and Wallgren (1984, 1985), Larson and Dinse (1985), Halpern and Brown (1987), Struthers and Farewell (1989), Sposto et al. (1992), followed up with various aspects and analyses of the model, but the first systematic treatment of what is now called the long term survivor or cure mix-

[1] The lognormal was not in fact the best fitting model for this data – Boag found that an exponential mixture distribution was slightly better – but he favoured the lognormal because it also described well the suite of other cancer data types he considered. We fitted the Weibull as being more general than the exponential; see Subsection 7.7.

ture model seems to have been in Maller and Zhou (1996). That book combines nonparametric and parametric theoretical formulations and proofs with many practical applications and examples of the model.

There has been an upsurge in interest in the model since the 1990s, with many applications areas explored, especially in medical statistics, and some substantial theoretical advances made. Correspondingly, computational facilities have improved tremendously, and with modern capabilities a wide variety of parametric or semi-parametric models of censored data with long term survivors can now be fitted routinely with the statistical package R; see for example Cai et al. (2012), Jackson (2016), Nui et al. (2018), Amdahl (2020) and López-Cheda et al. (2021). There have also been a number of review/overview and methodological articles, for example, Morbiduccia et al. (2003), Amica and Van Keilegom (2018), Patilea and Van Keilegom (2020), Musta et al. (2021), and the book by Peng and Yu (2021), summarising some of these aspects. A special issue of *Statistical Methods in Medical Research* is devoted to cure rate modelling, with an introduction by Balakrishnan (2017).

It seems appropriate now to present an overview drawing together earlier and some more recent developments as well as pointing out areas where further work is needed. The literature has grown too large and diverse to summarise completely here, and we confine ourselves to a selection reflecting our own main interests. In particular, we restrict our discussion to the *mixture* cure model in this survey. A number of papers deal with non-mixture cure models, for example Yin and Ibrahim (2005), Koutras and Milienos (2017), Leão et al. (2020), Milienos (2022) and Wang and Pal (2022). But the mixture model is easy to formulate and easy for practitioners to interpret, and it generalises naturally to competing risks setups; see Maller and Zhou (2002) and Subsection 4.5 herein.

Before leaving this introduction we mention variants of the mixture model for cure discussed in McLachlan and Peel (2000), Lee et al. (2017), McLachlan et al. (2019), Tawiah et al. (2020a) and Lee et al. (2021). The use of the EM algorithm in the mixture cure model is discussed in Sy and Taylor (2000), Yu et al. (2004b), McLachlan and Krishnan (2008) and Lee et al. (2021). Bayesian methods are considered in Morbiduccia et al. (2003) (classification of individuals into diagnostic classes), and in Gupta et al. (2016). Non-parametric cure models are considered in Peng and Dear (2000), Peng and Carriere (2002) and Peng (2003). Balakrishnan and his coworkers have analysed a large class of generalisations based on the mixture cure model; see Balakrishnan and Barui (2023) and their references.

1.1. Notation, assumptions and distributions

This subsection introduces the notation to be used throughout. We adopt the notation in Maller et al. (2022, 2023) and postulate an independent and identically distributed (iid) censoring model with right censoring. Thus a sample of size n consists of observations on the sequence of iid 2-vectors $\big(T_i = T_i^* \wedge U_i, C_i = \mathbf{1}(T_i^* \le U_i);\ 1 \le i \le n\big)$. The T_i^* with continuous cumulative distribution func-

tion (cdf) F^* on $[0,\infty)$ represent the times of occurrence of an event under study, such as the death of a person, or the onset of a disease, etc. The U_i, iid with continuous[2] cdf G on $[0,\infty)$, are censoring random variables, independent[3] of the T_i^*. In a sample from a population containing long-term survivors we observe the censored random variables $T_i = T_i^* \wedge U_i$ with censor indicators $C_i = \mathbf{1}(T_i^* \le U_i)$.

In the general mixture cure model, the censoring distribution G of the U_i is always assumed proper (total mass 1), but the distribution F^* of the T_i^* may be improper, of the form

$$F^*(t) = pF(t),\ t \ge 0, \tag{1.1}$$

where $0 < p \le 1$ and F is a proper distribution (has total mass 1). F is the distribution of the lifetimes of susceptible individuals in the population; only these can experience the event of interest and have a potentially uncensored failure time. The remainder are immune to the event of interest or cured of it. The presence of cured subjects is signalled by a value of $1-p > 0$, where $1-p$ is the proportion of the population that is cured. In this case the distribution F^* is improper, with total mass p. Observations on cured or immune individuals are always censored; those on susceptibles may or may not be according as the corresponding $T_i^* > U_i$ or not. We deal with the usual situation when we do not know which individuals in the sample may be cured or immune; we only have the survival times and censor indicators to work with.[4] The notation $\overline{F}^*(t) = 1 - F^*(t)$, $t \ge 0$, is used for the survival function (tail function) of F^*, and similarly $\overline{F}(t) = 1 - F(t)$ and $\overline{G}(t) = 1 - G(t)$. Let $H(t) := P(T_1 \le t)$ be the distribution of the observed survival times $T_i = T_i^* \wedge U_i$, with tail $\overline{H}(t) = 1 - H(t) = P(T_i^* \wedge U_i > t) = \overline{F}^*(t)\overline{G}(t)$, $t \ge 0$.

1.2. Data display: the Kaplan-Meier Estimator

We take a practical point of view whereby the data has prominence and the methodological developments flow from the inferences to be drawn from it. So suppose we have at hand a single sample of survival data which is to be analysed statistically. For visual display of the data the Kaplan and Meier (1958) empirical distribution function estimator (KME) of the lifetimes is commonly used, and we briefly review its properties here.[5]

[2]The assumption of continuity of F^* and G can be dropped; see Subsection 9.2.

[3]We discuss "informative censoring" (when the U_i and T_i^* are dependent) in Subsections 3.3 and 4.3.

[4]In some applications, some subjects surviving after a specified finite time threshold are considered cured. Safari et al. (2022, 2023) consider models in which cure status is only partially observed. We deal only with cases where survival times are either censored, or not.

[5]Also in use is the hazard function estimator of Nelson (1972), but the KME is most evocative for our purposes. A modified version of the KME was suggested by Beran (1981), but we use the original form. The importance of making preliminary visual assessments of the data is stressed in Yu et al. (2013): "We recommend that, regardless of the model used, the underlying assumptions for cure and model fit should always be graphically assessed".

The KME is a highly informative data display which shows clearly in visual form the features we want to investigate. To define it, order the lifetimes $(T_i)_{1\le i\le n}$ as $T_n^{(1)} < T_n^{(2)} < \cdots < T_n^{(n)}$, with associated censor indicators $C_n^{(1)}, C_n^{(2)}, \ldots, C_n^{(n)}$. Let $M(n) = T_n^{(n)} = \max_{1\le i\le n} T_i$ be the largest survival time and let $M_u(n) = \max_{1\le i\le n} C_i T_i$ be the largest observed *uncensored* survival time. An explicit definition of the KME is

$$\widehat{F}_n(t) := 1 - \prod_{1\le i\le n:\, T_n^{(i)}\le t} \Big(1 - \frac{C_n^{(i)}}{n-i+1}\Big), \text{ for } 0 < t \le M(n), \tag{1.2}$$

with $\widehat{F}_n(0) := 0$ and $\widehat{F}_n(t) := \widehat{F}_n(M(n))$ for $t > M(n)$. In (1.2), $n-i+1$ is the number of subjects "at risk" at a time just prior to $T_n^{(i)}$. Recall we assume F^* and G are continuous so with probability 1 there are no tied survival times in the data. Let

$$\widehat{p}_n := \widehat{F}_n(M(n)) \tag{1.3}$$

be the value of the KME at its right extreme. Equivalently, we can take $\widehat{p}_n := \widehat{F}_n(M_u(n))$, since the KME stays constant in $(M_u(n), M(n))$.

As an example, Fig. 1 shows the KME of the survival distribution (lifetime measured from the time of first diagnosis of the disease), with 95% confidence intervals, for Boag's 121 breast cancer patients.[6] The KME jumps only at the (uncensored) death times in the data, remaining constant at censored times, as indicated on the figure. In Fig. 1 it appears to level off at a value less than 1, consistent with Boag's observation of a possible cured component. This is very typical of the kind of KME plot that can be seen in much of the medical literature.

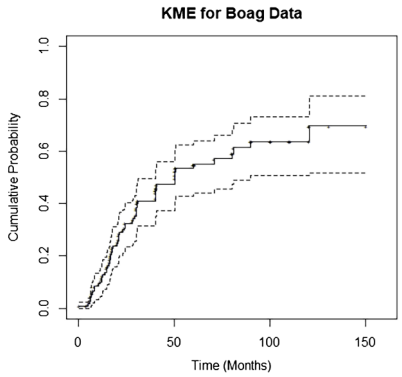


FIG 1. *KME for Boag data with 95% confidence intervals.*

The first step in a statistical analysis of survival data with possible cures is to assess and test for their existence in the population. A nonparametric estimate of

[6]The KME is usually displayed as the tail of the survival distribution, but we will use the term KME to describe the cumulative distribution function (cdf) of the lifetimes.

the population proportion p susceptible to dying from the disease is given by the maximum value of the KME, that is, $\widehat{p}_n$ as defined by (1.3), and its complement is the estimated cure proportion. As can be seen in Fig. 1, an estimate of the cure proportion for Boag's data is 0.30, with a 95% confidence interval (CI) of $[0.19, 0.48]$ calculated using the Greenwood and Irwin (1939) estimate. This interval excludes 0, in general agreement with Boag's observation of a possible cured component. This confidence interval assessment though indicative is only approximate, however, as the fact that $\widehat{p}_n$ is calculated from the KME at a random (not deterministic) time should be taken into account. Rigorous tests using $\widehat{p}_n$ are considered in Section 3.

The KME contains further evidence about the existence of a cured component. The length of the level stretch at the righthand end of the KME is indicative of the amount of followup in the data. We see in Fig. 1 a tendency for the KME to remain constant at lifetimes greater than 90 months, except for one late death at 120.6 months. So we might be inclined to pronounce a patient as "cured" of the disease if she survives more than about 90 months from first being diagnosed with it. But again an estimate like this comes with associated variability which should be included in any recommendation. One way to approach this is to estimate the probability a patient is cured having been followed up for a nominated event-free amount of time. We show how to estimate such an individual probability and assess its reliability in Section 8, and discuss a parametric approach to it using Boag's data in Subsection 8.3.

We emphasise that an "improper" sample KME, that is, one having right extreme less than 1, is suggestive but not definitive evidence of the presence of cured individuals in the population. Even in the absence of cures, it's possible for the right extreme of the KME to be less than 1 just by chance; Maller and Zhou (1993) calculate the probability of this event under the assumption of iid censoring. So we need rigorous tests for whether the right extreme of the KME is significantly less than 1, and for how this is related to the length of the level stretch at the righthand end of the KME.

Important information is also contained in the magnitudes of the largest survival time, $M(n)$, and the largest *uncensored* survival time, $M_u(n)$, and the numbers of observations (censored and uncensored) in the two time intervals $[0, M_u(n)]$ and $(M_u(n), M(n)]$. Much of the methodology is set out in Maller and Zhou (1996), which can be read as background to the present paper. Some important issues are left unresolved in that book, which we address here.

1.3. The role of the right extremes

The right extremes of the survival and censoring distributions play a special role in our analysis. Let $\tau_{F^*} = \inf\{t > 0 : F^*(t) = 1\}$ (with the inf of the empty set equal to ∞) be the right extreme of the survival distribution F^*, and similarly τ_F, τ_G and τ_H are the right extremes of F, G and H. The quantity τ_{F^*} represents the largest possible survival time of an individual in the population, but in a sample we can only observe times up to a maximum of $\tau_H := \min(\tau_{F^*}, \tau_G)$, due

to the censoring. We always have $H(\tau_H) = 1$, $G(\tau_G) = 1$ and $F(\tau_F) = 1$. When $p = 1$, so that $F^* \equiv F$, then F^* has total mass 1 and $\tau_{F^*} = \tau_F$; when $p < 1$ we have $\tau_{F^*} = \infty$, and $\tau_F \leq \tau_{F^*}$, with the possibility that $\tau_F < \tau_{F^*}$.

The inequality $\tau_F \leq \tau_G$ quantifies "sufficient followup" in the sense that it allows the largest possible susceptible survival times to be observed; in the contrary situation, $\tau_G < \tau_F$, censoring is so heavy that the data is truncated at a level below the maximum possible survival time. Besides expressing that followup is "sufficient" in this sense, the condition $\tau_F \leq \tau_G$ arises in a number of theoretical results. For example, the KME is biased downwards, but the bias tends to 0 in large samples, if and only if $\tau_F \leq \tau_G$; and this is also necessary and sufficient for the KME $\widehat{F}_n$ to be consistent for F^* on the whole line, that is, for $\sup_{t\geq 0} |\widehat{F}_n - F^*(t)| \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$ (Maller and Zhou, 1996, Thms. 3.8, 3.13 and 4.2). These are true for all $0 < p \leq 1$.

The convergence of the integral

$$\int_{\{0<t<\tau_H\}} \frac{dF(t)}{1 - G(t)} \tag{1.4}$$

is a required assumption in Thm. 4.2.3, p.82, of Gill (1980), which gives a functional limit theorem for the KME, from which its asymptotic normality at each $t > 0$ can be deduced. The "sufficient followup" condition $\tau_F \leq \tau_G$ is necessary for the integral in (1.4) to be finite, and subsequently plays an important role in many of the large-sample results in Gill (1980) and in the literature. For example, (1.4) is also assumed in Thm. 4.3 of Maller and Zhou (1996) which gives the asymptotic normality of $\widehat{p}_n$ in the case $p < 1$ (see Subsection 3.1).

These considerations highlight the need for information on, or assumptions about, the right hand endpoints of F^*, F and G, and, especially, whether they are finite or not. Most realistic is to assume $\tau_G < \infty$ since observation must always cease at some finite point. In many cases the assumption $\tau_F < \infty$ may also be natural. Certainly in real survival data no individual lives forever, but we would set $\tau_F = \infty$ for example when studying the occurrence of an infectious disease where an immune individual would never contract the disease no matter how long the follow-up. This can certainly be the case in epidemics such as the COVID virus pandemic, for example; and in Cairns et al. (2013) an analysis with children immune to malaria is given.

In practice, it is not uncommon to use a distribution with infinite right endpoint as the lifetime distribution; exponential, Weibull, lognormal, or Gumbel distributions are often used, for example in engineering reliability studies. In doing this we accept that the probability of seeing an extremely long lifetime under the assumed model is negligible, so the theoretical approximation is good enough for practical purposes. Alternatively, we could truncate the survival distribution at a (large) finite value, thus creating a distribution with $\tau_F < \infty$, as is often done in simulations. The truncation value is usually chosen so that the probability in the tail of the original distribution is negligible, say, less than 0.01.

2. Basic methodology

In this section we review some of the methodological advances of recent years. Subsection 2.1 gives a "splitting" result for the iid censoring model which is an intuitively attractive way of looking at the sample and furthermore facilitates the calculation of exact distributions of statistics such as $M(n)$ and $M_u(n)$ in Subsection 2.1 and Q_n, a statistic for assessing sufficiency of followup, in Subsection 4.2. Calculation of exact distributions allows for a rigorous investigation of their properties and makes unnecessary the need for simulations of percentage points. Having formulae for exact distributions also provides the basis for the asymptotic analyses we summarise in Sections 4, 5 and 8. Again the right extremes of the relevant distributions play an important role.

2.1. Splitting the sample at the largest uncensored observation

A key structural result obtained in Maller et al. (2022), Thm. 2.1, is that, conditional on the value of the largest uncensored survival time, and knowing the number of censored observations that exceed the largest uncensored lifetime, the sample partitions into two independent subsamples, each having the distribution of an iid sample of censored survival times, of reduced size, from the distribution of truncated random variables.

Recall the notation in Section 1.1. We keep $n \geq 3$. The splitting theorem tells us that the sample $S_n := \{T_i, 1 \leq i \leq n\}$ partitions into disjoint random sets:

$$S_n = S_n^{<} \cup \{M_u(n)\} \cup S_n^{>}, \tag{2.1}$$

where the component sets are

$$S_n^{<} = \{T_i : i \leq n \text{ and } T_i < M_u(n)\} \text{ and } S_n^{>} = \{T_i : i \leq n \text{ and } T_i > M_u(n)\}.$$

On $\{M_u(n) > 0\}$, let

$$N_c^{>}(M_u(n)) = |S_n^{>}| = \text{number of censored observations exceeding } M_u(n), \tag{2.2}$$

and

$$\{N_c^{>}(M_u(n)) = 0\} = \{M_u(n) = M(n)\} = \{\text{largest observation uncensored}\}$$

and $\{N_c^{>}(M_u(n)) = n\} = \{\text{all } n \text{ observations censored}\}$. On $\{N_c^{>}(M_u(n)) = n\}$, set $M_u(n) = 0$. Then, conditional on $\{M_u(n) = t > 0\}$ and $\{N_c^{>}(M_u(n)) = r\}$, $S_n^{<}$ consists of $n - r - 1$ iid variables with distribution that of T_1, conditional on $T_1 < t$; and $S_n^{>}$ consists of r iid variables with tail function

$$P(T_1^{>,c}(t) > x) := \frac{\int_x^\infty \overline{F}^*(s)G(\mathrm{d}s)}{\int_t^\infty \overline{F}^*(s)G(\mathrm{d}s)}, \quad x \geq t,$$

which is the distribution tail of a censored observation conditional on being bigger than t. Furthermore, $S_n^{<}$ and $S_n^{>}$ are conditionally independent given

$M_u(n) = t$ and $N_c^>(M_u(n)) = r$. Note that observed lifetimes less than $M_u(n)$ may be either censored or uncensored but observed lifetimes greater than $M_u(n)$, i.e., those in $S_n^>$, are necessarily censored.

The splitting result remains true if conditioning is done on the 3-vector $(M_u(n), M(n), N_c^>(n))$ rather than just on $(M_u(n), N_c^>(n))$. That result is used in Maller et al. (2022) to derive the finite sample joint distribution of $M(n)$ and $M_u(n)$, as in the next theorem.

2.2. Distributions of the largest censored and uncensored lifetimes

Theorem 2.1. *(i) The joint distribution of $M_u(n)$ and $M(n)$ is given by*

$$P\big(0 \le M_u(n) \le t,\, 0 \le M(n) \le x\big) = \begin{cases} \big(\int_{z=0}^{x} \overline{F}^*(z)\mathrm{d}G(z)\big)^n, & \text{if } t = 0,\ 0 \le x \le \tau_H; \\ H^n(x), & \text{if } 0 \le t \le \tau_H;\ 0 \le x \le t; \\ \big(\int_{z=t}^{x} \overline{F}^*(z)\mathrm{d}G(z) + H(t)\big)^n, & \text{if } 0 \le t < x \le \tau_H. \end{cases} \tag{2.3}$$

(ii) The distribution of $M_u(n)$ is given by

$$P\big(M_u(n) \le t\big) = J^n(t),\ t \ge 0, \tag{2.4}$$

where $J(t)$ is the distribution of an uncensored lifetime:

$$J(t) = \begin{cases} 1 - \int_{z=0}^{\tau_H} \overline{G}(z)\mathrm{d}F^*(z) = \int_{z=0}^{\tau_H} \overline{F}^*(z)\mathrm{d}G(z),\ t = 0; \\ 1 - \int_{z=t}^{\tau_H} \overline{G}(z)\mathrm{d}F^*(z) = \int_{z=t}^{\tau_H} \overline{F}^*(z)\mathrm{d}G(z) + H(t),\ 0 \le t \le \tau_H. \end{cases} \tag{2.5}$$

(iii) The distribution of $M(n)$ is given by

$$P(M(n) \le x) = H^n(x),\ x \ge 0. \tag{2.6}$$

Remarks. Note that (2.3) is consistent with the fact that $0 \le M_u(n) \le M(n) \le \tau_H$. There is no probability mass outside the region $[0, \tau_H] \times [0, \tau_H]$ so the distribution in (2.3) equals 1 for $t > \tau_H$, $x > \tau_H$. Likewise the distribution in (2.5) equals 1 for $t > \tau_H$.

Note also that Lines 2 and 3 on the RHS of (2.3) include the value for $t = 0$; there is mass on the interval $\{t = 0\} \times [0 \le x \le \tau_H]$, as given by the first line on the RHS of (2.3). $M_u(n)$ has the distribution of the maximum of n iid copies of a rv with distribution J on $[0, \infty)$. This distribution has mass $\big(\int_{z=0}^{\tau_H} \overline{F}^*(z)\mathrm{d}G(z)\big)^n$ at 0 corresponding to all observations being censored. (It may seem pedantic to include these degenerate cases but they are important for checking that distributions are proper.) Recall that τ_H is the right endpoint of the support of the distribution H. The right extreme τ_J of the distribution J may be strictly less than τ_G; in fact, we have $\tau_J = \tau_F \wedge \tau_G$. In general, $\tau_J \ne \tau_H = \tau_{F^*} \wedge \tau_G$.

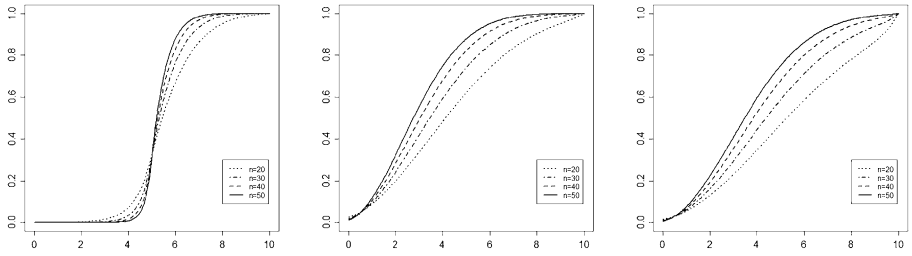


FIG 2. *The cdf of* $M(n) - M_u(n)$ *from* (2.7) *with* $F = U[0, a]$, $G = U[0, 10]$, $p = 0.3$. *Left to right:* $a = 5, 10, 15$.

Illustrative plots of the distributions of $M(n)$ and $M_u(n)$ are in the Supplementary Material to the paper, and asymptotic distributions of $M(n)$ and $M_u(n)$ are in Section 5.

Also important are the length of the time interval between the largest uncensored survival time and the largest survival time, and the ratio of those times. For them we have the following distributions.

Theorem 2.2. *We have for* $0 \le u \le \tau_H \le \infty$

$$P\big(M(n) - M_u(n) \le u\big) = n\int_{t=0}^{\tau_H} \Big(\int_{z=t}^{(t+u)\wedge\tau_H} \overline{F}^*(z)\mathrm{d}G(z) + H(t)\Big)^{n-1} \overline{G}(t)\mathrm{d}F^*(t) + \Big(\int_{z=0}^{u} \overline{F}^*(z)\mathrm{d}G(z)\Big)^n, \tag{2.7}$$

with $P\big(M(n) - M_u(n) \le u\big) = 1$ *for* $u > \tau_H$. *We have for* $v \ge 1$

$$P\big(M(n) \le vM_u(n) | M_u(n) > 0\big) = \frac{\int_{t=0}^{\tau_H} \Big(\int_{z=t}^{(tv)\wedge\tau_H} \overline{F}^*(z)\mathrm{d}G(z) + H(t)\Big)^{n-1} \overline{G}(t)\mathrm{d}F^*(t)}{\int_{t=0}^{\tau_H} \Big(\int_{z=t}^{\tau_H} \overline{F}^*(z)\mathrm{d}G(z) + H(t)\Big)^{n-1} \overline{G}(t)\mathrm{d}F^*(t)},$$

with $P\big(M(n) \le vM_u(n) | M_u(n) > 0\big) = 0$ *for* $0 \le v < 1$.

Remarks. Setting $u = 0$ in (2.7) we see that the distribution of the difference $M(n) - M_u(n)$ has mass at 0 of

$$P\big(M(n) - M_u(n) = 0\big) = P\big(M(n) = M_u(n)\big) = n\int_{t=0}^{\tau_H} H^{n-1}(t)\overline{G}(t)\mathrm{d}F^*(t). \tag{2.8}$$

This is the probability that the largest observation is uncensored (Maller and Zhou (1993)).

Figures 2 and 3 illustrate the distribution of the difference $M(n) - M_u(n)$ for $F = U[0, a]$ and $G = U[0, 10]$, with $p = 0.3, 0.7$ and $a = 5, 10, 15$. In this

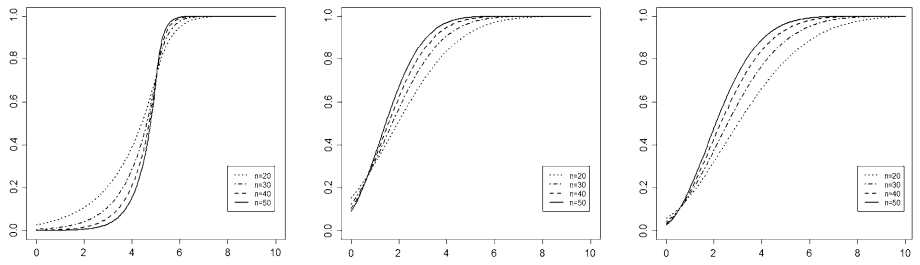


FIG 3. *The cdf of* $M(n) - M_u(n)$ *from* (2.7) *with* $F = U[0, a]$, $G = U[0, 10]$, $p = 0.7$. *Left to right:* $a = 5, 10, 15$.

scenario, $a = 5$ represents sufficient follow-up, in which case the distribution concentrates closely around $10 - 5 = 5$, whereas for $a = 10$ (a marginal case), and $a = 15$ (insufficient follow-up) the distributions are much more spread out, representing more uncertainty in the length of the plateau. We note that not all of the plots show an atom at 0. Plots were done using the statistical package R.

2.3. Conditional multinomial distribution of numbers

Besides the definition of $N_c^>(M_u(n))$ in (2.2), we need notation for the numbers of censored observations smaller or greater than $M_u(n)$. Let $N_u(n)$ be the total number of uncensored observations in the sample, and, when $N_u(n) > 1$, define

$$N_u^<(M_u(n)) = N_u(n) - 1$$
$$= \text{number of uncensored observations strictly less than } M_u(n)$$

and

$$N_c^<(M_u(n)) = \text{number of censored observations less than } M_u(n).$$

On $\{N_u(n) = 1\}$, set $N_u^<(M_u(n)) = 0$. When $N_u(n) = 0$, we do not define $N_u^<(M_u(n))$ or $N_c^<(M_u(n))$. Let

$$N_c(n) = \text{total number of censored observations in the sample.}$$

We also use the notation $\mathbb{N}_n := \{1, 2, \ldots, n\}$, $n = 1, 2, \ldots,$.

With these definitions and conventions, on $\{N_u(n) \geq 1\}$ the $N_u^<(M_u(n))$, $N_c^<(M_u(n))$ and $N_c^>(M_u(n))$ take values in $\mathbb{N}_{n-1} \cup \{0\}$, satisfying $N_u^<(M_u(n)) + N_c^<(M_u(n)) + N_c^>(M_u(n)) = n - 1$ and $N_c(n) = N_c^>(M_u(n)) + N_c^<(M_u(n))$. We also have

$$\{N_u(n) = 0\} = \{\text{all } n \text{ observations censored}\} = \{M_u(n) = 0\}.$$

The next result concerns the vector $(N_c^>(M_u(n)), N_c^<(M_u(n)), N_u^<(M_u(n)))$. This vector is not as might be thought at first multinomially distributed, but

it is, conditional on the value of $M_u(n)$. We proved it as another application of the splitting property. We need more notation. Define the functions

$$\begin{aligned} p_c^{>}(t) &= \frac{\int_{y=t}^{\tau_H} \overline{F}^*(y)\mathrm{d}G(y)}{\int_{y=t}^{\tau_H} \overline{F}^*(y)\mathrm{d}G(y) + H(t)}, \\ p_c^{<}(t) &= \frac{\int_{y=0}^{t} \overline{F}^*(y)\mathrm{d}G(y)}{\int_{y=t}^{\tau_H} \overline{F}^*(y)\mathrm{d}G(y) + H(t)}, \text{ and} \\ p_u^{<}(t) &= \frac{\int_{y=0}^{t} \overline{G}(y)\mathrm{d}F^*(y)}{\int_{y=t}^{\tau_H} \overline{F}^*(y)\mathrm{d}G(y) + H(t)}, \end{aligned}$$

which are non-negative and add to 1 for each $t \in (0, \tau_H)$.

Theorem 2.3. *(i) We have for $t > 0$, $0 \le r, s, k \le n-1$, $r+s+k = n-1$, the multinomial probability*

$$\begin{aligned} &P\big(N_c^{>}(M_u(n)) = r,\, N_c^{<}(M_u(n)) = s,\, N_u^{<}(M_u(n)) = k \big| M_u(n) = t\big) \\ &= \frac{(n-1)!}{r!\,s!\,k!} \times (p_c^{>}(t))^r (p_c^{<}(t))^s (p_u^{<}(t))^k. \end{aligned}$$

(ii) Consequently, conditional on $M_u(n) = t$, the marginal rvs $N_c^{>}(M_u(n))$, $N_c^{<}(M_u(n))$ and $N_u^{<}(M_u(n))$ are trinomial with $n-1$ as the number of trials and success probabilities $p_c^{>}(t)$, $p_c^{<}(t)$ and $p_u^{<}(t)$ respectively.

(iii) Conditional on $M_u(n) = t > 0$, the number of censored observations $N_c(n) = N_c^{>}(M_u(n)) + N_c^{<}(M_u(n))$ is Binomial $(n-1, p_c(t))$, where $p_c(t) = p_c^{<}(t) + p_c^{>}(t)$.

(iv) Conditional on $N_c(n) = \ell$ and $M_u(n) = t$, the number $N_c^{>}(M_u(n))$ is Binomial $(\ell, p_c^{+}(t))$, where

$$p_c^{+}(t) := \frac{\int_{y=t}^{\tau_H} \overline{F}^*(y)\mathrm{d}G(y)}{\int_{y=0}^{\tau_H} \overline{F}^*(y)\mathrm{d}G(y)}.$$

Remarks. Since $t > 0$ in Theorem 2.3, the conditioning on $M_u(n) = t$ implies $M_u(n) > 0$, thus $N_u(n) \ge 1$, and there is at least one uncensored observation. Thus $N_u^{<}(M_u(n)) + N_c^{<}(M_u(n)) + N_c^{>}(M_u(n)) = n-1$.

An application of Theorem 2.3 is to derive in Section 4 the finite sample distribution of the statistic Q_n, used as a test for sufficient followup. But first we want to test for the presence of cures in the population.

3. Testing for the presence of cures using $\widehat{p}_n$

The KME was of course not available to Boag in 1949 and he used a parametric approach, inferring the existence of cures in his population from his sample estimate of the proportion cured and its standard error, obtained from a lognormal

mixture fit to the data. This confidence interval assessment and the implied one-sided hypothesis test of $H_0 : p = 1$ applied to a parameter estimate overlooks the restriction of p to $[0, 1]$. Section 5.3 of Maller and Zhou (1996) contains a discussion of this issue as it relates to a parametric analysis of the cure model.

The advent of the KME in 1958 was a major advance in the visualisation and analysis of survival data, and especially with reference to assessing the presence or otherwise of cures. But how do we formalise conclusions drawn from the visual information displayed in the KME graph?

That the medical literature did and still does wrestle with this problem is illustrated for example by the "mini-review" paper of Damuzzo et al. (2019), from which we quote: *Some anti-cancer treatments (e.g., immunotherapies) determine, on the long term, a durable survival in a small percentage of treated patients; in graphical terms, long-term survivors typically give rise to a plateau in the right tail of the survival curve. ... To capture the presence of a survival plateau by quantitative methods, two approaches have thus far been proposed: the milestone method and the area-under-the-curve (AUC) method. ...*

The problem is that with cures (possibly) present, the KME is improper and theoretical quantities such as the mean survival time or expected "area-under-the-curve" are, formally, infinite. A remedy is to restrict the calculation of such properties to the (proper) survival distribution of the susceptibles, but then we must estimate this distribution.

We can start by estimating the proportion of cured subjects in the population. This proportion is the complement of the susceptible proportion, of which perhaps the simplest and most intuitive estimate is $\widehat{p}_n$, the maximum value of the KME. The properties of $\widehat{p}_n$ are explored in Ghitany et al. (1995) and more extensively in Maller and Zhou (1996), though there are still unknown features; we discuss one such in Subsection 3.2. With an estimate of the proportion of cured subjects, we can rescale the KME or a fitted parametric distribution to estimate the survival distribution of the susceptibles.

As foreshadowed in Section 1.3, the sufficient followup condition $\tau_F \leq \tau_G$ plays an important role. Under our present assumptions (continuity of F and G), $\widehat{p}_n$ is consistent for p if and only if $\tau_F \leq \tau_G$, see Theorem 4.1 of Maller and Zhou (1996), and when $0 < p < 1$ and the integral in (1.4) is finite: $\widehat{p}_n$ is asymptotically normally distributed, as stated in the next subsection.

3.1. Asymptotic distribution of $\widehat{p}_n$, cured present

We keep $p < 1$ throughout this subsection. This means that $\tau_{F^*} = \infty$ and consequently $\tau_H = \tau_{F^*} \wedge \tau_G = \tau_G$. When the integral in (1.4) is finite, the sufficient followup condition $\tau_F \leq \tau_G$ holds, as well as the finiteness of the function

$$v(t) := \int_{[0,t]} \frac{dF^*(s)}{(1 - F^*(s))^2(1 - G(s))} \tag{3.1}$$

for all $t < \tau_H = \tau_G$. The largest uncensored survival time $M_u(n) < \tau_F$ with probability 1, so $v(t)$ in (3.1) can be evaluated at time $M_u(n)$.

The next theorem, Theorem 4.3 of Maller and Zhou (1996), is based on Theorem 4.2.3 of Gill (1980), which is stated for convenience in the Supplementary Material to the paper.

Theorem 3.1. *Assume $p < 1$ and the integral in (1.4) is finite. Then*

$$\sqrt{n}\big(\widehat{p}_n - p\big) \xrightarrow{\mathrm{D}} N\big(0, (1-p)^2 v(\tau_G)\big), \text{ as } n \to \infty. \tag{3.2}$$

To apply Theorem 3.1 in practice, we need a sample estimate for the population quantity $v(\tau_G)$ in (3.2). A consistent estimator of $v(t)$ is, under our assumptions,

$$v_n(t) := \sum_{i:T_i>t} \frac{nC_n^{(i)}}{(n-i+1)(n-i+1-C_n^{(i)})}, \ t \geq 0 \tag{3.3}$$

(see Theorem 4.4, p.73, and the discussion following it in Maller and Zhou (1996)), and correspondingly a consistent estimator of $v(\tau_G)$ is

$$v_n := \sum_{i=1}^{n-1} \frac{nC_n^{(i)}}{(n-i+1)(n-i+1-C_n^{(i)})}. \tag{3.4}$$

3.2. Asymptotic distribution of $\widehat{p}_n$, cured not present

In the case $p = 1$, where there are no immunes in the population, the asymptotic distribution of $\widehat{p}_n$ is currently unknown in complete generality, and finding it still remains a challenge. Here we give a partial but enlightening result.

From (1.2) we can write the complement of $\widehat{p}_n$ as

$$1 - \widehat{p}_n = 1 - \widehat{F}_n(T_n^{(n)}) = \prod_{i=1}^{n} \Big(1 - \frac{C_n^{(i)}}{n-i+1}\Big) = \prod_{i=1}^{n} \Big(1 - \frac{C_n^{(n-i+1)}}{i}\Big), \tag{3.5}$$

where, recall that $T_n^{(1)} < T_n^{(2)} < \cdots < T_n^{(n)}$ are the ordered lifetimes with associated censor indicators $C_n^{(1)}, C_n^{(2)}, \ldots, C_n^{(n)}$. Now, we are only interested in data for which the largest observation is censored, i.e., $C_n^{(n)} = 0$, and in this case we can take logs and turn the product into a sum. Thus from (3.5), on the event $\{C_n^{(n)} = 0\}$, we get

$$|\log(1 - \widehat{p}_n)| = \sum_{i=2}^{n} \Big|\log\Big(1 - \frac{C_n^{(n-i+1)}}{i}\Big)\Big| = \sum_{i=2}^{n} a_i C_n^{(n-i+1)}, \tag{3.6}$$

where $a_i := |\log(1 - 1/i)|$, $i \geq 2$. Since $p = 1$ we have $\widehat{p}_n \xrightarrow{\mathrm{P}} 1$, so $|\log(1 - \widehat{p}_n)| \xrightarrow{\mathrm{P}} \infty$ as $n \to \infty$, and we need to determine the rate of this divergence. Despite its simple representation, it seems hard to analyse (3.6) in full generality. So we turn to a special case to get some intuition.

3.3. The Koziol-Green model

The model, also known as the proportional hazards model (not to be confused with Cox's proportional hazards model), assumes the iid censoring model of Subsection 1.1 but with the distributions F^* and G functionally related by

$$\overline{G}(t) = (\overline{F}^*(t))^\beta \tag{3.7}$$

for some $\beta > 0$ (Koziol and Green (1976)).[7] Kirmani and Dauxois (2004) give real-data examples where the model appears to apply and some where it does not. There is also a substantial literature in which the model is used to gain theoretical insight into the behaviour of survival models (e.g., Cheng and Lin (1987), Chang (1996)). We use it here similarly to get some valuable insight. Applied in the next lemma, the Koziol-Green property greatly simplifies the analysis of $\widehat{p}_n$ when $p = 1$. Allen (1963) showed that T_i and C_i are independent for each i in this model (see also Chen et al. (1982)), and this transfers easily to the ordered values.

Lemma 3.1. *In the Koziol-Green model (3.7), with iid censoring, $(C_n^{(i)})_{1\le i\le n}$ are iid, each having the same distribution as C_1, namely, a Bernoulli $(1/(\beta+1))$ distribution.*

The proof of Lemma 3.1 is omitted. As a corollary we get from (3.6), on the event $\{C_n^{(n)} = 0\}$,

$$|\log(1-\widehat{p}_n)| \overset{\mathrm{D}}{=} \sum_{i=2}^{n} a_i C_n^{(n-i+1)}, \tag{3.8}$$

where $a_i = |\log(1-1/i)|$ and the $C_n^{(n-i+1)}$ are iid, $2 \le i \le n$, for each $n > 1$.

Using Lemma 3.1, we can give a quite explicit representation for the limiting distribution of $\widehat{p}_n$ in the Koziol-Green model, when $p = 1$. Note that this implies $F^* \equiv F$. Recall that, even in the absence of cures, it's possible for the right extreme of the KME to be less than 1 just by chance; using (2.8) and (3.7) we can calculate the complementary probability as

$$\mathrm{P}(\widehat{p}_n = 1) = \mathrm{P}(C_n^{(n)} = 1) = n\int_{t=0}^{\tau_H} H^{n-1}(t)\overline{G}(t)\mathrm{d}F^*(t) = \frac{\beta}{\beta+1}. \tag{3.9}$$

So in the Koziol-Green model the distribution of $\widehat{p}_n$ has mass of $\beta/(\beta+1)$ at 1 (for all $n \in \mathbb{N}$). Alternatively, conditional on the event $\{\widehat{p}_n < 1\} = \{C_n^{(n)} = 0\}$, we have the following limit distribution.

Theorem 3.2. *In the Koziol-Green model, with iid censoring and $p = 1$,*

$$\lim_{n\to\infty} P\big(n^{1/(\beta+1)}(1-\widehat{p}_n) \le x | C_n^{(n)} = 0\big) = P(e^{-Y} \le x), \text{ for } x > 0, \tag{3.10}$$

[7] Koziol and Green (1976) attribute the model formulation to Breslow and Crowley (1974), but it seems to have first appeared in Armitage (1959); see aso Cox (1959).

where Y is a random variable defined by

$$Y = \sum_{i\geq 2} a_i\big(Y_i - E(Y_i)\big). \tag{3.11}$$

Here the Y_i are iid as Bernoulli $(1/(\beta+1))$ and $a_i = |\log(1-1/i)|$, $i \geq 2$. The rv Y is infinitely divisible with $E(Y) = 0$ and $\mathrm{Var}(Y) = \sum_{i\geq 2} a_i^2/(\beta+1)$.

Proof of Theorem 3.2: From (3.8) we can calculate (with $(C_i)_{2\leq i\leq n}$ iid)

$$\begin{aligned} & E\big(|\log(1-\widehat{p}_n)|;\, C_n^{(n)} = 0\big) = \sum_{i=2}^{n} a_i E(C_i) = \frac{1}{\beta+1}\sum_{i=2}^{n} a_i \\ = \;& -\frac{1}{\beta+1}\sum_{i=2}^{n}\log(1-1/i) = \frac{1}{\beta+1}\sum_{i=2}^{n}\big(\log i - \log(i-1)\big) \\ = \;& \frac{1}{\beta+1}\log n = \log n^{1/(\beta+1)}. \end{aligned}$$

Then argue as follows: for $x > 0$,

$$\begin{aligned} & P\big(n^{1/(\beta+1)}(1-\widehat{p}_n) \leq x | C_n^{(n)} = 0\big) \\ & = P\big(n^{1/(\beta+1)}(1-\widehat{p}_n) \leq x,\, \widehat{p}_n < 1 | C_n^{(n)} = 0\big) \\ & = P\big(|\log(1-\widehat{p}_n)| - \frac{1}{\beta+1}\log n \geq -\log x | C_n^{(n)} = 0\big) \\ & = P\big(\sum_{i=2}^{n} a_i(C_i - E(C_i)) \geq -\log x | C_1 = 0\big) \\ & = P\big(\sum_{i=2}^{n} a_i(C_i - E(C_i)) \geq -\log x\big). \end{aligned} \tag{3.12}$$

Now use the Kolmogorov convergence criterion (e.g., Lemma 3.16, p.47, of Kallenberg (2021)), which gives that a sum $\sum_{i\geq 2}\xi_i$ of independent random variables $(\xi_i)_{i\geq 2}$ having mean 0 and satisfying $\sum_{i\geq 2} E(\xi_i^2) < \infty$, converges a.s. to a finite rv. Apply this to (3.12) with $Y_i = C_i$ and $\xi_i := a_i(Y_i - E(Y_i))$, $i \geq 2$. Then the Y_i are iid Bernoulli and since $a_i \leq 2/(i-1)$ for $i \geq 2$, the convergence of the series $\sum_{i\geq 2} E\big(a_i(Y_i - E(Y_i))\big)^2$ is clear and we deduce that the rv Y in (3.11) is finite a.s. and has the specified variance. Thus the RHS of (3.12) converges to $P\big(Y \geq -\log x\big) = P\big(e^{-Y} \leq x\big)$, as required in (3.10). The infinite divisibility of Y follows from Corollary 13.7, p.251, of Kallenberg (2021), since the sequence $(\xi_i)_{1\leq i\leq n}$ forms a null array (Kallenberg (2021), p.249). □

Remarks. (i) (3.9) and Theorem 3.2 show that the limiting distribution of $\widehat{p}_n$, centered at 1 and normalised by $n^{1/(\beta+1)}$, is a mixture of the distribution of e^{-Y} and a point mass at 0 in proportions $1/(\beta+1)$ and $\beta/(\beta+1)$. Consequently it depends strongly on the parameter β in (3.7), and so it seems we cannot expect any sort of universal limiting result for $\widehat{p}_n$, in general, when $p = 1$. We do get $\sqrt{n}$ convergence to e^{-Y} in (3.9) when $\beta = 1$ (the symmetric case).

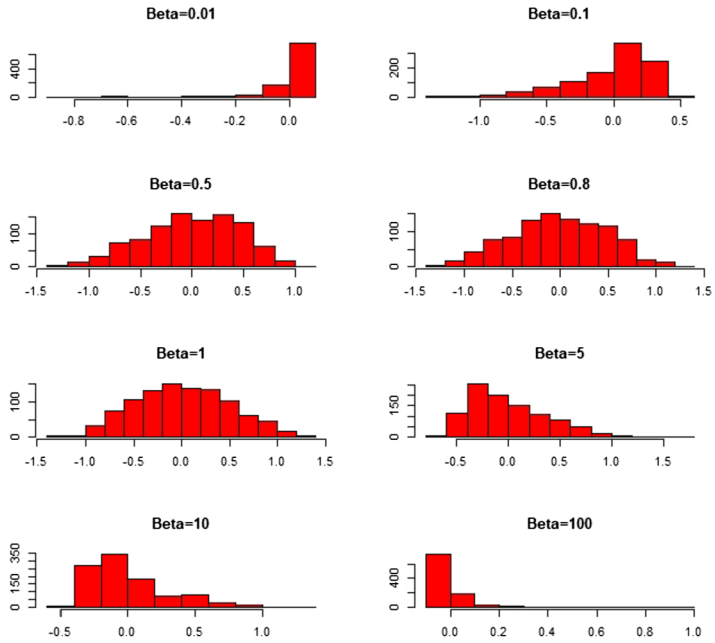


FIG 4. *Distribution of Y for various β.*

(ii) Varying β in the model covers a range of possible censoring scenarios. When $\beta = 0$, (3.10) gives $n(1-\widehat{p}_n) \xrightarrow{\mathrm{D}} e^{-Y}$, as $n \to \infty$, and since Var $(Y) = 0$ in this case, Y degenerates at 0, and we get $n(1-\widehat{p}_n) \xrightarrow{\mathrm{P}} 1$. When $\beta = 0$, (3.7) gives $\overline{G} \equiv 1$, thus G degenerates at ∞, indicating that there is no censoring at all. Since we conditioned on seeing $\widehat{p}_n < 1$, the result $n(1-\widehat{p}_n) \xrightarrow{\mathrm{P}} 1$ is consistent with the fact that in this case the KME is simply the empirical distribution function estimator of F, which jumps $1/n$ at each sample point, and at the right extreme jumps from $1-1/n$ to 1.

Formally setting $\beta = \infty$, we again get Var $(Y) = 0$, and again Y degenerates at 0. In this case, (3.7) gives $\overline{G} \equiv 0$, thus G degenerates at 0, indicating that all observations are censored. Since there are then no susceptibles present, (3.10) correctly signals $1-\widehat{p}_n \xrightarrow{\mathrm{P}} 1$, that is, $\widehat{p}_n \xrightarrow{\mathrm{P}} 0$.

(iii) Fig. 4 shows sample pdfs of the distribution of Y for various β. The distribution of Y is skewed to the left for $\beta < 1$, skewed to the right for $\beta > 1$, symmetrical for $\beta = 1$, and degenerates to 0 when $\beta \downarrow 0$ or $\beta \to \infty$.

(iv) Centering $|\log(1-\widehat{p}_n)|$ at its expectation as we do in Theorem 3.2 gives the right order of magnitude in some other situations, too, and we conjecture that this is so very generally. With uniform censoring and exponential survival (a uniform distribution for G and an exponential distribution for F), for example, we can show with some calculations (details omitted) that, conditional on $\{C_n^{(n)} = 0\}$, $|\log(1-\widehat{p}_n)| - E|\log(1-\widehat{p}_n)| = O_P(1)$ (is bounded in probability),

as $n \to \infty$. The same result holds if we assume the pairwise independence of the $C_n^{(i)}$, i.e., that $P(C_n^{(i)} = 1, C_n^{(j)} = 1) = P(C_n^{(i)} = 1)P(C_n^{(j)} = 1)$, $j > i$. But this pairwise independence is not true in general for the iid censoring model.

4. Testing for sufficient followup using Q_n

In Section 1.3 we quantified the idea of "sufficient followup" as the condition $\tau_F \leq \tau_G$. In this section we formalize a test for this condition based on the statistic Q_n proposed in Maller and Zhou (1994) (some alternatives are discussed in Section 10). Recall the notations in Section 2 of $M(n)$ for the largest and $M_u(n)$ for the largest uncensored survival time in a sample of size n. To calculate Q_n, take the length of the interval $[M_u(n), M(n)]$, measure back this distance from $M_u(n)$, and count the number of uncensored observations seen (omitting $M_u(n)$). A heuristic rationale for this procedure is in Maller and Zhou (1996), p.84. To formulate it, set $\Delta_n := 2M_u(n) - M(n)$ and define[8]

$$\begin{aligned} Q_n &= \frac{1}{n}\#\{\text{uncensored observations in } [\Delta_n, M_u(n))\} \\ &= \frac{1}{n}\#\{\text{uncensored observations exceeding } 2M_u(n) - M(n)\}. \quad (4.1) \end{aligned}$$

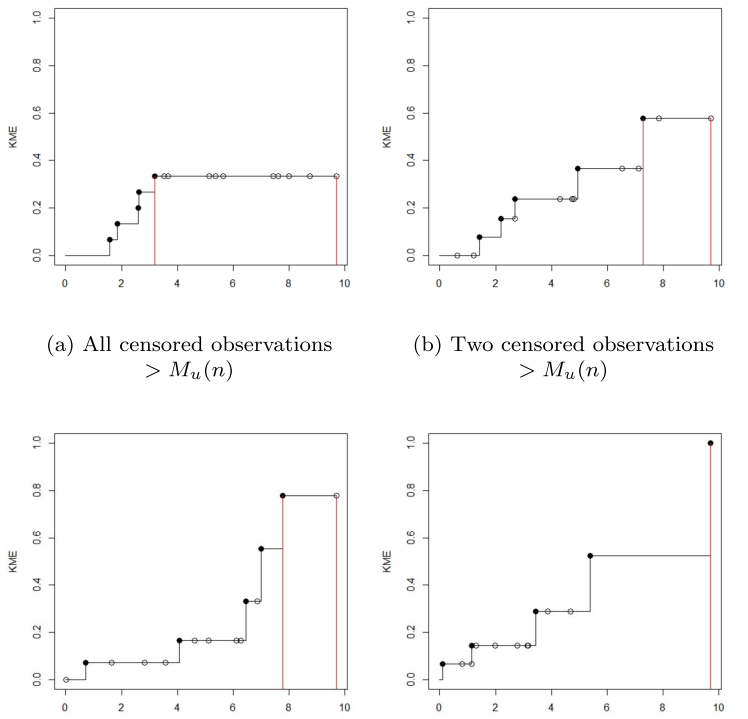
Assume the hypothesis $H_0 : \tau_G < \tau_F$, that followup is *insufficient*. Under this assumption it follows from the results in Subsection 5.4 that $Q_n \xrightarrow{\text{P}} 0$ as $n \to \infty$, so the probability of seeing a large value of Q_n is small. Consequently we reject H_0 and conclude that follow-up is sufficient if the observed value of the test statistic exceeds a nominated quantile of its distribution under H_0. When H_0 is not true a test based on large values of Q_n will reject the hypothesis of insufficient follow-up with probability approaching 1 as $n \to \infty$.

Using the splitting theorem we can give a formula for the exact distribution of Q_n under the iid censoring model from which the asymptotic distribution is derived in Section 5 under appropriate conditions.

4.1. Understanding the sample properties of Q_n

The value of Q_n depends in a complicated way on the numbers of censored and uncensored observations, the way they happen to occur below or above $M_u(n)$, and on the relative magnitudes of $M_u(n)$ and $M(n)$. In order to calculate its distribution under the iid censoring model we need to understand how it varies with these things. For this we consider hypothetical sample situations, vary the mentioned quantities and see how the value of Q_n changes. Recall that $N_u(n)$ is the number of uncensored survival times, necessarily in $[0, M_u(n)]$,

[8]Note that we exclude $M_u(n)$ when counting the number of uncensored observations greater than Δ_n; after all $M_u(n)$ is an uncensored observation (the largest one). Excluding it as we do simplifies formulae by allowing Q_n to take minimum value 0 rather than $1/n$ as would be the case if we counted $M_u(n)$ in Q_n. So we also have the case $k = 0$ in (4.2). The distinction is minor in practice and disappears as $n \to \infty$.



(a) All censored observations $> M_u(n)$

(b) Two censored observations $> M_u(n)$

(c) 1 censored observation $> M_u(n)$

(d) Largest observation uncensored

FIG 5. *Schematic KME diagrams.*

$N_c^<(n)$ is the number of censored survival times in $[0, M_u(n))$, and $N_c^>(n)$ is the number of censored survival times in $(M_u(n), M(n)]$. Thus there is a total of $N_c(n) = N_c^<(n) + N_c^>(n) = n - N_u(n)$ censored survival times in the sample.

We begin by considering possible values of Δ_n, noting that we always have $\Delta_n = 2M_u(n) - M(n) \leq 2M_u(n) - M_u(n) = M_u(n)$. Possible values of Δ_n range from $\Delta_n = -M(n)$ if $M_u(n) = 0$, that is, if all observations are censored, to $\Delta_n = M_u(n) = M(n)$ if $M_u(n) = M(n)$, that is, if the largest observation is uncensored. We have $\Delta_n = 0$ if it happens that $M_u(n) = M(n)/2$. Thus we may have $\Delta_n < 0$, $\Delta_n = 0$, or $\Delta_n > 0$. When $\Delta_n \leq 0$ then $[\Delta_n, M_u(n)) \supseteq [0, M_u(n))$ and (4.1) gives $nQ_n = N_u(n) - 1$, as 1 less than the number of uncensored observations in the sample. At the other extreme, the interval $[\Delta_n, M_u(n))$ may be empty, and this is certainly so when $\Delta_n = M_u(n)$. Whenever this occurs we set $Q_n = 0$.

Now think of the way Q_n changes if we rearrange the conformation of the censored observations less than or greater than $M_u(n)$, by keeping $M(n)$ and $N_u(n) > 0$ fixed and varying $N_c^<(n)$ and $N_c^>(n)$, thus moving $M_u(n)$ between 0 and $M(n)$. It helps to visualise the various situations with schematic KME diagrams in the different cases, as in Figure 5.

We start with an extreme case.
Case 1: $N_c^<(n) = 0$, $N_c^>(n) > 0$ (see Fig. 5(a)). In this conformation all the censored observations in the sample form a level stretch of the KME between $M_u(n)$ and $M(n)$. In this case $M_u(n)$ takes the minimum possible value for the sample under this kind of rearrangement, $M(n) - M_u(n)$ takes the maximum possible value, $\Delta_n = 2M_u(n) - M(n) = M_u(n) - (M(n) - M_u(n))$ takes the minimum possible value, and Q_n takes the maximum possible value under this kind of rearrangement for the sample. We reject $H_0 : \tau_G < \tau_F$ and conclude there is sufficient follow-up if we observe large values of Q_n, so this arrangement accords with our intuition that a (long) level stretch on the KME between $M_u(n)$ and $M(n)$ indicates there is sufficient follow-up.
Case 2: $N_c^<(n) > 0$, $N_c^>(n) > 0$ (see Fig. 5(b)). As censored observations are moved to the left of $M_u(n)$, $M_u(n)$ tends to increase and $M(n) - M_u(n)$ tends to decrease (it cannot increase). So Δ_n will tend to increase and consequently Q_n will tend to decrease. This accords with our intuition that a decrease in the number of censored observations above $M_u(n)$ and in the length of the level stretch of the KME between $M_u(n)$ and $M(n)$ makes it less likely to reject H_0, the hypothesis of insufficient follow-up.

Ultimately, continuing this process, we reach:
Case 3: $N_c^<(n) > 0$, $N_c^>(n) = 1$ (see Fig. 5(c)). The one censored observation above $M_u(n)$ is $M(n)$ itself and $\Delta_n = 2M_u(n) - M(n)$ will be close to or equal to $M_u(n)$. The interval $[\Delta_n, M_u(n))$ is small and Q_n is small, possibly equal to 0 (this certainly occurs when $\Delta_n = M_u(n)$). This accords with our intuition that a short level stretch of the KME between $M_u(n)$ and $M(n)$ indicates via a small value of Q_n that there is insufficient follow-up.

In these scenarios, Q_n decreases monotonically from a sufficient follow-up situation to one with insufficient follow-up.

The actual values taken on by Q_n in these scenarios depend on the relative magnitudes of $M_u(n)$ and $M(n)$. The possibilities are as follows. Note that since $N_u(n) > 0$, we have $M_u(n) > 0$.

(a) When $0 < M_u(n) \le \frac{1}{2}M(n)$, then $\Delta_n \le 0$ and $[\Delta_n, M_u(n)) \supseteq [0, M_u(n))$. In this case

$$Q_n = \frac{1}{n}\#\{\text{uncensored observations other than } M_u(n)\} = \frac{N_u(n) - 1}{n}.$$

This is the largest value Q_n can take for a given sample.

(b) When $\frac{1}{2}M(n) < M_u(n) < M(n)$, then $\Delta_n > 0$ and the interval $[\Delta_n, M_u(n))$ contains, say, k observations. We have $k \ge 0$ and $k \le n - 1$ since there is at least one censored observation greater than $M_u(n)$, namely, $M(n)$. So we can write

$$Q_n = \frac{k}{n} = \frac{1}{n}\#\{\text{uncensored observations in } [\Delta_n, M_u(n))\}, \tag{4.2}$$

where k decreases from its maximum value when $M_u(n)$ is near $\frac{1}{2}M(n)$, reaching 0 when $M_u(n)$ is near $M(n)$.

There are also two other extreme cases to consider.

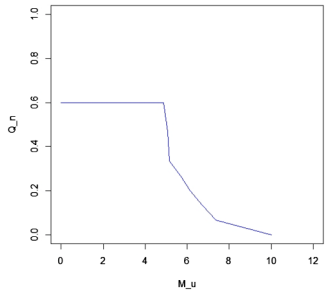


FIG 6. *Possible Values for* Q_n.

(c) When $N_c^{>}(n) = 0$, then $M_u(n) = M(n)$, and the largest observation is uncensored (see Fig. 5(d)). Then $\Delta_n = M_u(n)$, the interval $[\Delta_n, M_u(n))$ is empty, and $Q_n = 0$. Here the level stretch has length 0 and the low value of Q_n correctly reflects sufficient follow-up. (This case includes also the possibility that all observations are uncensored, corresponding to $N_u(n) = n$, and $k = n$.) But this Case (c) means there is no evidence of immunes and hence no issue of sufficient or insufficient follow-up. We condition on the non-occurrence of this event when calculating the distribution of Q_n.

(d) When $N_u(n) = 0$, all observations are censored, and, formally, $Q_n = 0$. This anomalous or ambiguous case is of no interest and we condition on its non-occurrence also, when calculating the distribution of Q_n.

In this thought experiment let

$$W_n := \max\{M_u(n) : Q_n > 0\}.$$

Then $\frac{1}{2}M(n) < W_n < M(n)$, and we may have $W_n = M(n)$ if the largest observation is uncensored. Possible scenarios for values of Q_n are illustrated schematically in Fig. 6. We see that Q_n decreases monotonically as $M_u(n)$ moves between $\frac{1}{2}M(n)$ and $M(n)$ but remains constant on $[0, \frac{1}{2}M(n)]$.

4.2. Finite sample distribution of Q_n

Theorem 4.1 gives a formula for the finite sample distribution of Q_n assuming the iid censoring model. The formula and the associated Lemma 4.1 are derived in Maller et al. (2023) as an application of the splitting result outlined in Section 2.1. We keep $n > 2$, $0 < t < x \le \tau_H$ and $1 \le r \le n-1$, condition on the event $\{M_u(n) = t, M(n) = x, N_c^{>}(M_u(n)) = r\}$, and consider separately the cases $2t - x \le 0$ (Case A) and $0 < 2t - x \le \tau_H$ (Case B). For Case A define

$$\pi^A(t) := \frac{P(0 < T_1^* \le t, T_1^* \le U_1)}{P(T_1^* \wedge U_1 \le t)} = \frac{\int_0^t \overline{G}(y)\mathrm{d}F^*(y)}{H(t)}$$

and for Case B define

$$\pi^B(t,x) := \frac{P(2t-x < T_1^* \le t, T_1^* \le U_1)}{P(T_1^* \wedge U_1 \le t)} = \frac{\int_{2t-x}^{t} \overline{G}(y)\mathrm{d}F^*(y)}{H(t)}.$$

Define also the probability

$$p_c^{>}(t,x) = \frac{\int_{y=t}^{x} \overline{F}^*(y)\mathrm{d}G(y)}{\int_{y=t}^{x} \overline{F}^*(y)\mathrm{d}G(y) + H(t)},$$

and let

$$\rho^A(t,x) := (1-p_c^{>}(t,x))\pi^A(t) \text{ and } \rho^B(t,x) := (1-p_c^{>}(t,x))\pi^B(t,x). \quad (4.3)$$

Let $Bin(n,\rho)$ denote a binomial random variable with parameters $n \in \mathbb{N}$ and $\rho \in (0,1)$.

Lemma 4.1. ***Part (i):*** *We have for* $1 \le r \le n-1$, $0 < t < x \le \tau_H$,

$$P\big(N_c^{>}(M_u(n)) = r \big| M_u(n) = t, M(n) = x\big) = P\big(Bin(n-2, p_c^{>}(t,x)) = r-1\big)$$

(with the lefthand side taken as 0 when $r = 0$*).*
Part (ii): *For* $0 \le k \le n-2$,

$$P\big(nQ_n = k \big| M_u(n) = t, M(n) = x\big) = P\big(Bin(n-2,\, \rho(t,x)) = k\big), \quad (4.4)$$

where $\rho(t,x) = \rho^A(t,x)$ *in Case A and* $\rho(t,x) = \rho^B(t,x)$ *in Case B (see (4.3)).*

We need one more formula: from (2.3) we have

$$\begin{aligned} &P_n(\mathrm{d}t, \mathrm{d}x) := P\big(M_u(n) \in \mathrm{d}t, M(n) \in \mathrm{d}x\big) \\ &= n(n-1)\Big(\int_{y=t}^{x} \overline{F}^*(y)\mathrm{d}G(y) + H(t)\Big)^{n-2} \overline{G}(t)\mathrm{d}F^*(t)\,\overline{F}^*(x)\mathrm{d}G(x). \quad (4.5) \end{aligned}$$

Theorem 4.1. *Assume the iid censoring model in Subsection 1.1. Then for* $n > 2$, $k = 0, 1, 2, \ldots, n-2$,

$$P\big(nQ_n = k \big| 0 < M_u(n) < M(n)\big) = \frac{A_n(k) + B_n(k)}{D_n}, \quad (4.6)$$

where

$$A_n(k) = \int_{t=0}^{\tau_H/2} \int_{x=2t}^{\tau_H} P\big(Bin(n-2,\, \rho^A(t,x)) = k\big) P_n(\mathrm{d}t, \mathrm{d}x)$$

and

$$B_n(k) = \Big[\int_{t=0}^{\tau_H/2} \int_{x=t}^{2t} + \int_{t=\tau_H/2}^{\tau_H} \int_{x=t}^{\tau_H}\Big] P\big(Bin(n-2,\, \rho^B(t,x)) = k\big) P_n(\mathrm{d}t, \mathrm{d}x).$$

The denominator in (4.6) is

$$\begin{aligned} D_n &= P\big(0 < M_u(n) < M(n)\big) \\ &= 1 - \Big(\int_{t=0}^{\tau_H} \overline{F}^*(t)\mathrm{d}G(t)\Big)^n - n\int_{t=0}^{\tau_H} H^{n-1}(t)\overline{G}(t)\mathrm{d}F^*(t). \end{aligned}$$

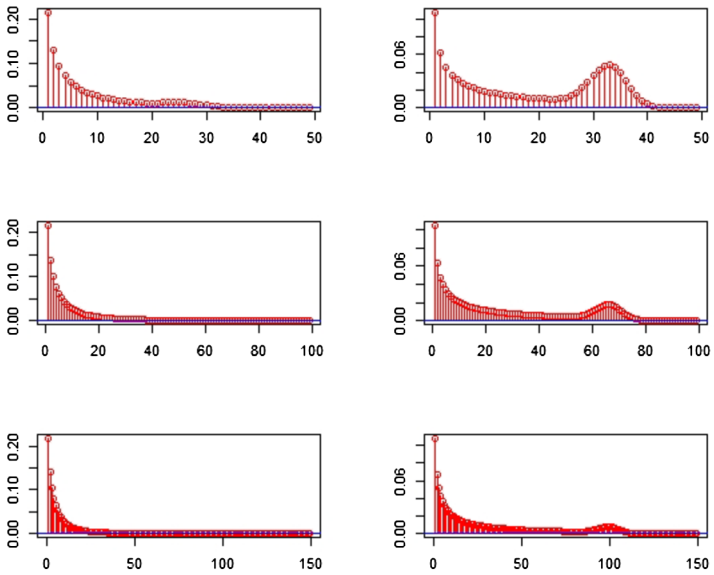


FIG 7. *Probability mass functions for* nQ_n. $F = exp(1)$, $p = 0.8$. *Left column:* $G = U[0,3]$; *Right column:* $G = U[0,6]$. *Top, middle, bottom row:* $n = 50, 100, 150$.

Figure 7 shows graphs of the probability mass functions (pmfs) of nQ_n calculated from (4.6) for various scenarios with F exponential, G uniform, and $n = 50, 100, 150$. For small n the pmfs are bimodal, reflecting the two components on the RHS of (4.6). The bimodality is least prominent when censoring is heavy and disappears altogether as $n \to \infty$.

4.3. Dependent censoring

The independence assumptions inherent in the iid censoring model may not be tenable in some situations and there have been a number of studies where it has been relaxed. See for example the competing dependent risks of leukaemia relapse and graft versus host disease analysed in Kalbfleisch and Prentice (2003) and Kovar et al. (2018) and the approaches in Tawiah et al. (2020a,b). To assess the effect of departures from independence we consider the distribution of Q_n in a model where there is dependence between survival and censoring.

As in Subsection 1.1 we assume a sample consists of observations on the 2-vectors $\big(T_i = T_i^* \wedge U_i,\ C_i = \mathbf{1}(T_i^* \le U_i);\ 1 \le i \le n\big)$, where now the T_i^* and U_i are dependent with a joint continuous distribution having marginal distributions F^* and G on $[0, \infty)$. We considered a kind of functional dependence between F^* and G in the Koziol-Green model of Subsection 3.3. In the present subsection we model actual dependence between T_i^* and U_i using a copula to connect the marginal distributions with the joint distribution.[9]

[9]For other discussions on the independence of survival and censoring, see Peterson (1976), Lagakos and Williams (1978), Lagakos (1979), Leung et al. (1997), Rufibach et al. (2023).

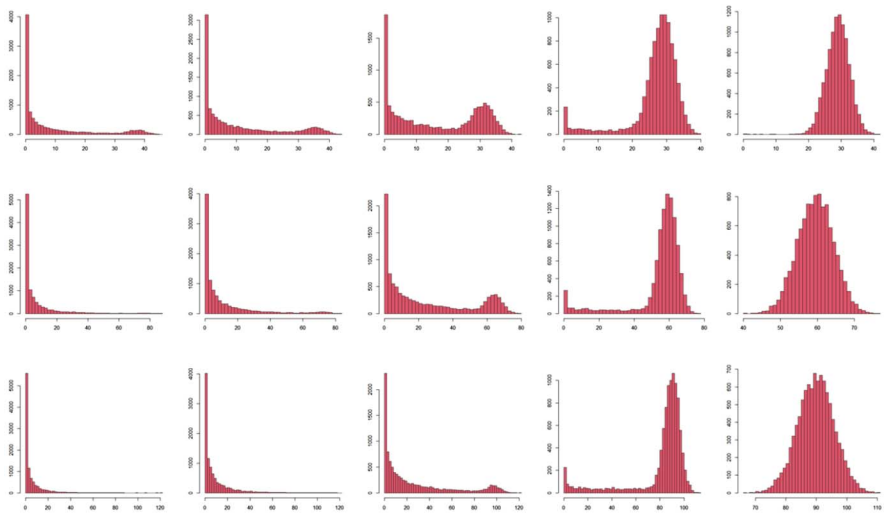


FIG 8. *Probability mass functions for nQ_n with Frank copula for dependence. $F = exp(1)$, $p = 0.8$, $G = U[0,6]$. Top, middle, bottom panel: $n = 50, 100, 150$. Left to right: $\theta = 300, 6, 0, -6, -300$.*

Numerous copulas are defined and described in Nelsen (2006), to which we refer for background. By virtue of Sklar's theorem (Sklar (1959)), the bivariate distribution of (T_i^*, U_i) with specified continuous marginals F^* and G can be expressed in a unique way via a 2-copula J

$$J(w_1, w_2, \theta) := P(W_1 \le w_1, W_2 \le w_2),$$

for two uniform random variables W_1, W_2 and a copula parameter θ which quantifies the dependence between them. We restrict our discussion to the class of *Archimedean copulas*, considering one such, the *Frank* (Frank (1979)) copula. The corresponding copula function is

$$J_{\text{Frank}}(w_1, w_2) = -\frac{1}{\theta} \log\left(1 + \frac{(e^{-\theta w_1} - 1)(e^{-\theta w_2} - 1)}{e^{-\theta} - 1}\right).$$

In order to simulate an observation on (T_i^*, U_i), it is sufficient to simulate a vector $(W_1, W_2) \sim J$ with values w_1 and w_2, where the rvs W_1 and W_2 are independent Uniform $[0, 1]$. Then

$$t^* = F^{*,\leftarrow}(w_1), \quad u = G^{\leftarrow}(w_2),$$

is an observation on (T^*, U) having the required joint distribution (see Salvadori et al. (2007), Appendix A).

We simulated samples of size $n = 50, 100, 150$, from the function J for the Frank copula for various values of θ, taking $F^* = pF$, where F is exponential with parameter 1, $G = U[0, 6]$ and $p = 0.8$. In each sample we calculated the value of Q_n and repeated this 10000 times to draw up the pmfs of Q_n in Fig. 8. In the figure the pmf for $\theta = 0$ corresponds to independence. Viewing

from the centre panel left we see that introducing high positive dependence tends to concentrate the mass near small values of Q_n and the bimodality of the distribution becomes less; introducing negative dependence also tends to decrease the bimodality but shifts the pmfs closer to normal as sample size increases. This intuition can be useful in practice to assess how dependence between T^* and U affects percentage points of the distribution of Q_n.

4.4. Sufficient followup and identifiability

Cure models suffer from identifiability issues when F^* is estimated nonparametrically. Laska and Meisner (1992) note that the KME is the generalized maximum likelihood estimator (GMLE) of F^*, and its maximum value, $\widehat{p}_n$ (in our notation), which is the GMLE of p in (1.1), is *uniquely defined* when the largest survival time is uncensored – but not when it is censored. Of course this is just the situation we are in when investigating the possibility of cures. The convention according to Laska and Meisner (1992) is then to take the GMLE of p to be the maximum value of the KME, which is precisely our $\widehat{p}_n$.

Survival data in practice usually have a vector of covariates $\mathbf{x}$ associated with each failure or censored time. In the mixture cure model we can allow the susceptible proportion $p = p(\mathbf{x})$ (the "incident" model) to depend on the covariate information. A natural and common choice is to use the well understood and easily interpretable logistic regression formulation for this. Regarding the cdf $F(t; \mathbf{x})$ of the susceptible survival times (the "latency" model), covariates can be introduced into most of the usual parametric models (see Section 7) in natural ways. Vu et al. (1998) discuss some of the issues, and Zhao (2023) has example analyses.

Li et al. (2001) discuss the identifiability of cure models when F is parametric, possibly with covariates $\mathbf{x}$. Under regularity conditions, the model is identifiable regardless of whether a corresponding model for $p = p(\mathbf{x})$ is parametric or not. With a logistic regression for the incident model, and some structure (e.g. proportional hazards or a parametric distribution) assumed for the latency model, identifiability is not a problem in practice *provided follow-up is sufficient.*

When follow-up is insufficient, however, cure models do suffer from identifiability issues. If the KME is improper but still tending to increase near its right hand endpoint, it's understandable that the model has difficulty distinguishing between a possible presence of cures and a situation in which failures will continue beyond the range of the data to the extent that all subjects eventually fail. In other words, difficulty in distinguishing between the cases $p = 1$ and $p < 1$. Yu et al. (2004a) and Peng and Taylor (2014) refer to this as a "near non-identifiability problem", suggesting that it manifests as a relatively flat likelihood surface. A consequence is that parameters will tend to be imprecisely estimated. Examples illustrating this are in Zhao (2023). See also Parsa and Van Keilegom (2023). In Section 6 we discuss a way of adjusting estimates when follow-up is insufficient.

4.5. Sufficient followup with competing risks

With competing risks data, failure is classified into more than one cause, e.g., death from cancer, or heart attack, etc., and there may be subjects with long followup not having died, thus, suggesting the possibility of immune or cured individuals in the population besides those susceptible to the risks of dying. The cumulative incidence function (CIF) plays the role of the Kaplan Meier estimator in this arena: Choi and Zhou (2002), Kalbfleisch and Prentice (2003), and a natural generalisation of the mixture model can be used: Larson and Dinse (1985), Maller and Zhou (2002), Jeong and Fine (2006). It's clear that a consideration of sufficient followup is important here, but to our knowledge no rigorous study has been done.

Identifiability issues are also important and have been addressed by Tsiatis (1975) and Lemdani and Pons (1997, 2003). The latter obtain consistency and asymptotic normality of the estimators in a parametric setup under some stringent assumptions, including identifiability assumptions, treating the boundary value case when individuals are "totally susceptible" to death, among other results. They also address the problem of the elimination of a cause, providing an ingenious approach to it in their context. This question goes back a long way. D'Alembert in 1761 and Daniel Bernoulli in 1766 gave formulae for the gain in lifetime to be expected after elimination of smallpox as a cause of death – hypothetical at that time, but since become a reality: Fenner et al. (1988).

5. Asymptotics of largest censored/uncensored lifetimes, and Q_n

In practice, samples of survival data are often large enough that asymptotic methods are appropriate. In this section we list some recently obtained results for $M(n)$, $M_u(n)$ and Q_n. Equations (2.4) and (2.6) suggest the use of extreme value methods to find limiting distributions of $M(n)$ and $M_u(n)$ after rescaling by nonstochastic sequences, and this turns out to be appropriate for Q_n too.

The case when G has a finite right endpoint is particularly important, because, as explained in Section 4, when testing for sufficient followup we proceed by assuming $H_0 : \tau_G < \tau_F$, that followup is *insufficient*, and this implies a finite τ_G. But for theoretical as well as modelling purposes we want to allow infinite endpoints for F and G, so we consider all cases. We refer to Embrechts et al. (1997), de Haan and Ferreira (2006) and Resnick (2008) for general background, and for the domain of attraction results we use.

5.1. Maximum domains of attraction

A distribution F belongs to the maximum domain of attraction of an extreme value distribution if the maximum of a sample of n from F converges in distribution as $n \to \infty$ to a finite nondegenerate random variable, after appropriate centering and norming. There are three possible limiting types: Fréchet, Gumbel and reverse Weibull. A condition for being in one of these domains specifies in

some way the rate at which the tail $\overline{F}(t)$ of F approaches 0 as $t \uparrow \tau_F$. Each of the three domains is relevant to some aspect of the large sample behaviour of $M(n)$, $M_u(n)$ or Q_n, as we outline in what follows.

The standard Fréchet cdf is $\Phi_\gamma(x) = \exp(-x^{-\gamma})$, where $\gamma, x > 0$. A distribution F belongs to the maximum domain of attraction of Φ_γ if and only if $\overline{F}(t) > 0$ for all large $t > 0$ and $\overline{F}(t)$ is regularly varying with index $-\gamma$ as $t \to \infty$; thus,

$$\lim_{t\to\infty} \frac{\overline{F}(\lambda t)}{\overline{F}(t)} = \lambda^{-\gamma} \text{ for each } \lambda > 0 \tag{5.1}$$

(Resnick (2008), p.54, de Haan and Ferreira (2006), p.10). For the Fréchet domain, necessarily $\tau_F = \infty$. Subsection 6.1 employs this type of convergence.

The standard Gumbel cdf is $\Lambda(x) = \exp(-e^{-x})$, $x \in \mathbb{R}$. The necessary and sufficient condition for F to belong to the Gumbel maximum domain of attraction becomes

$$\lim_{t\uparrow\tau_F} \frac{\overline{F}(t + yf(t))}{\overline{F}(t)} = e^{-y}, \tag{5.2}$$

for $y \in \mathbb{R}$ and a positive *auxiliary function* $f(t)$. In the Gumbel case, τ_F can be finite or infinite.

A slightly more stringent condition than (5.2) is to assume F is a *von Mises distribution*; i.e., F is absolutely continuous with tail function satisfying

$$\overline{F}(x) = 1 - F(x) = k_1 \exp\Big\{-\int_{x_0}^{x} \frac{1}{f(u)} du\Big\}, \quad x_0 < x < \tau_F, \tag{5.3}$$

for some $k_1 > 0$ and $x_0 \in (0, \tau_F)$, where f is a positive differentiable function on $[x_0, \tau_F)$ with derivative f' satisfying $\lim_{x\uparrow\tau_F} f'(x) = 0$. The difference between (5.3) and (5.2) is that in (5.2) we replace the k_1 of (5.3) with a function $c(x) \to c_0 \in (0, \infty)$. Differentiation of (5.3) shows that, in this formulation, f is the *reciprocal hazard function* of F on $[x_0, \tau_F)$. A possible choice of f is

$$f(t) = \frac{1}{\overline{F}(t)} \int_t^{\tau_F} \overline{F}(x) dx, \ 0 < t < \tau_F. \tag{5.4}$$

Equivalent to (5.3) is that F has a finite negative second derivative F'' for all t in some left neighbourhood of τ_F satisfying

$$\lim_{t\uparrow\tau_F} \frac{F''(t)\overline{F}(t)}{(F'(t))^2} = -1. \tag{5.5}$$

See Resnick (2008), p.46, or de Haan and Ferreira (2006), Thm. 1.2.1, p.19. Subsections 5.3 and 6.2 employ this type of convergence.

The standard reverse Weibull cdf is $\Psi_\gamma(x) = \exp(-|x|^\gamma)$, $x < 0$, $\gamma > 0$, with $\Psi_\gamma(x) = 1$ when $x \geq 0$. Rewriting Prop. 1.1.3 of Resnick (2008), p.59, we have that F belongs to the maximum domain of attraction of Ψ_γ if and only $\tau_F < \infty$ and $\overline{F}(\tau_F - t)$ is regularly varying with index γ as $t \downarrow 0$, $t < \tau_F$. We then have the convergence

$$\lim_{n\to\infty} P\big(a_n(\tau_F - \max_{1\leq i\leq n} X_i) \leq x\big) = \big(1 - \exp(-x^\gamma)\big) 1_{\{x\geq 0\}} \tag{5.6}$$

for a sample $(X_i)_{1\le i\le n}$ from F, with a norming sequence $a_n \to \infty$. Subsections 5.2 and 5.4 employ this type of convergence.

5.2. Asymptotics of extremes, right extreme of G finite

In this subsection we present the joint asymptotic distributions of $M(n)$ and $M_u(n)$ when $\tau_G < \infty$, for various values of p and τ_F, derived using the formulae in Theorem 2.1 and the ideas in Subsection 5.1. We consider scenarios of sufficient or insufficient followup, and assume appropriate extreme value domain of attraction conditions on F and G. The rescaled $M(n)$ and $M_u(n)$ are then asymptotically independent with asymptotic Weibull distributions. Recall that τ_J is the right extreme of the distribution J in (2.4).

Theorem 5.1. Case 1: *Assume $\tau_F < \tau_G < \infty$ and $0 < p < 1$, and in addition, as $z \downarrow 0$, $0 < z < \tau_G$,*

$$\overline{G}(\tau_G - z) = a_G(1+o(1))z^\gamma L_G(z) \text{ and } \overline{F}(\tau_F - z) = a_F(1+o(1))z^\beta L_F(z), \quad (5.7)$$

where a_G, a_F, γ, β are positive constants and $L_G(z)$ and $L_F(z)$ are slowly varying as $z \downarrow 0$. Then for some deterministic norming sequences $a_n, b_n \to \infty$,

$$\begin{aligned}&\lim_{n\to\infty} P\big(a_n(\tau_G - M(n)) \le u,\, b_n(\tau_F - M_u(n)) \le v\big)\\ &= \big(1 - e^{-(1-p)u^\gamma}\big)\big(1 - e^{-p\overline{G}(\tau_F)v^\beta}\big), \text{ for } u, v \ge 0. \end{aligned} \quad (5.8)$$

Case 2: *Assume $\tau_F < \tau_G < \infty$ and $p = 1$, and in addition, as $z \downarrow 0$, $0 < z < \tau_G$,*

$$\overline{G}(\tau_G - z) = a(1+o(1))z^\beta L(z) \text{ and } \overline{F}(\tau_F - z) = a(1+o(1))z^\beta L(z), \quad (5.9)$$

where a and β are positive constants and $L(z)$ is slowly varying as $z \downarrow 0$. Then there exists a deterministic sequence $a_n \to \infty$ such that

$$\lim_{n\to\infty} P\big(a_n(\tau_F - M(n)) \le u,\, a_n(\tau_F - M_u(n)) \le v\big) = 1 - e^{-\overline{G}(\tau_F)u^\beta}, \text{ for } u, v \ge 0. \quad (5.10)$$

Case 3: *Assume $\tau_G < \tau_F < \infty$ and $0 < p < 1$, and assume the first relation in (5.7) holds. Suppose in addition that, in a neighbourhood of τ_G, F has a density f which is positive and continuous at τ_G. Then there exist deterministic sequences $a_n \to \infty$ and $b_n \to \infty$ such that*

$$\begin{aligned}&\lim_{n\to\infty} P\big(a_n(\tau_G - M(n)) \le u,\, b_n(\tau_G - M_u(n)) \le v\big)\\ &= \big(1 - e^{-(1-pF(\tau_G))u^\gamma}\big)\big(1 - e^{-pf(\tau_G)v^{1+\gamma}/(1+\gamma)}\big), \text{ for } u, v \ge 0. \end{aligned} \quad (5.11)$$

Further, this result remains true under the same assumptions when $p = 1$, and/or when $\tau_F = \infty$.

In any of Cases 1–3 we have $M(n) \xrightarrow{\mathrm{P}} \tau_H$ and $M_u(n) \xrightarrow{\mathrm{P}} \tau_J$ as $n \to \infty$.

Remarks. (i) Theorem 5.1 is proved in Maller et al. (2022). Possible choices of a_n and b_n are specified there. Note that in Case 2, F and G are required to have the same order of magnitude in neighbourhoods of their right extremes.

(ii) A common assumption is of an exponential distribution for lifetime survival: $F(t) = 1 - e^{-\lambda t}$, $\lambda > 0$, $t \geq 0$, and the uniform distribution for censoring, $G = U[A, B]$ (e.g., Goldman (1984, 1991)). These distributions for F and G constitute very good baseline reference distributions for assessing the practicality of theoretical results, and this situation, or a close approximation to it, is often the case in practice.

(iii) When $G = U[0, \tau_G]$, $\tau_G > 0$, we have $\overline{G}(\tau_G - z) = (z/\tau_G)\mathbf{1}_{\{0 \leq z \leq \tau_G\}}$. Thus G satisfies (5.7) with $a_G = 1/\tau_G$, $\gamma = 1$ and $L_G \equiv 1$, while $\tau_F = \infty$ when F is exponential (λ). Case 3 of Theorem 5.1 applies.

5.3. Asymptotics of extremes, right extreme of G infinite

In this subsection we assume F and G are in the domain of maximal attraction of the Gumbel and have infinite endpoints. In addition to $M_u(n)$ and $M(n)$ as previously defined, denote the *largest censored lifetime* by $M_c(n)$. Theorem 5.2 deals with the joint asymptotic distribution of $M_u(n)$ and $M_c(n)$. An additional Theorem 5.3 shows that the number of censored observations bigger than the largest uncensored lifetime is asymptotically geometric.

The analysis here is based on Maller and Resnick (2022). Throughout we assume both F and G are absolutely continuous and satisfy the von Mises condition (5.3) and an analogous condition for G. Then the product $\overline{F} \times \overline{G}$ also has the form of the tail of a von Mises distribution, as shown in Maller and Resnick (2022), namely, *If $\overline{F}$ and $\overline{G}$ are von Mises distribution tails satisfying* (5.3) *and the analogous condition for G with auxiliary function g, then $\overline{H} = \overline{F} \times \overline{G}$ is a von Mises distribution tail with auxiliary function $h := fg/(f+g)$.*

It follows that H is in the domain of attraction of the Gumbel, and we have

$$\lim_{t \to \infty} \frac{\overline{H}(t + xh(t))}{\overline{H}(t)} = e^{-x}, \ x \in \mathbb{R}. \tag{5.12}$$

The positive sequences $a(n)$ and $b(n)$ satisfying

$$\lim_{n \to \infty} n\overline{H}(b(n)) = 1 \quad \text{and} \quad a(n) = h(b(n)) \tag{5.13}$$

provide the correct centering and norming for maxima of samples drawn from H to converge in distribution to a Gumbel distribution; see Resnick (2008), p.40. Analogous sequences $a^G(n)$ and $b^G(n)$ are appropriate centering and norming sequences for maxima of samples from G.

$M_u(n)$ and $M_c(n)$ properly normalized have limit distributions which are products ((Maller and Resnick, 2022)) and therefore can reasonably be analyzed separately in large samples. The limits involve two independent Gumbel rvs, G_u and G_c which may depend on parameters ν_u, ν_c such that

$$P(G_u(\nu_u) \leq x) = \exp\{-\nu_u e^{-x}\}, \ x \in \mathbb{R},$$

and similarly for G_c. We add a third condition comparing the magnitudes of f and g:

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \kappa, \quad 0 \le \kappa \le \infty. \tag{5.14}$$

Theorem 5.2. *Assume* (5.3)*, the analogous condition for* G*, and* (5.14)*. Suppose* $0 < p \le 1$.
(a) Assume $p = 1$ *and* $a(n)$ *and* $b(n)$ *satisfy* (5.13).
(i) Suppose $\kappa \in (0, \infty)$*. Then, as* $n \to \infty$,

$$\Big(\frac{M_u(n) - b(n)}{a(n)}, \frac{M_c(n) - b(n)}{a(n)}\Big) \xrightarrow{\mathrm{D}} \Big(G_u\Big(\frac{1}{1+\kappa}\Big), G_c\Big(\frac{\kappa}{1+\kappa}\Big)\Big),$$

and consequently

$$\frac{M(n) - b(n)}{a(n)} \xrightarrow{\mathrm{D}} G_u\Big(\frac{1}{1+\kappa}\Big) \vee G_c\Big(\frac{\kappa}{1+\kappa}\Big).$$

(ii) Suppose $\kappa = 0$ *or* $\kappa = \infty$*. Then*

$$\frac{M(n) - b(n)}{a(n)} \xrightarrow{\mathrm{D}} G(1),$$

where $G(1)$ *is a Gompertz rv with cdf* $\exp\{-e^{-x}\}$, $x \in \mathbb{R}$.
(b) Suppose $\kappa \in [0, \infty)$ *and* $p < 1$*. Then*

$$\Big(\frac{M_u(n) - b(n)}{a(n)}, \frac{M_c(n) - b^G(n)}{a^G(n)}\Big) \xrightarrow{\mathrm{D}} \Big(G_u\Big(\frac{p}{1+\kappa}\Big), G_c(1-p)\Big).$$

A corollary to Theorem 5.2 gives an asymptotic distribution for the difference of $M(n)$ and $M_u(n)$ when $p = 1$.

Corollary 5.1. *Assume the same conditions as in Theorem 5.2, and suppose* $p = 1$*. When* $\kappa \in (0, \infty)$ *the normed difference* $D(n) := (M(n) - M_u(n))/a(n)$ *converges in distribution to the random variable* D *having cdf*

$$P[D \le x] = \frac{1}{1 + \kappa e^{-x}}, \ x \ge 0,$$

with mass $P[D = 0] = 1/(1+\kappa)$ *at 0. When* $\kappa = 0$, $D(n) \xrightarrow{\mathrm{P}} 0$*, and when* $\kappa = \infty$, $D(n) \xrightarrow{\mathrm{P}} \infty$.

Finally, we consider the number of censored observations that are bigger than the largest uncensored lifetime, denoted by $N_c^{>}(M_u(n))$.

Theorem 5.3. *Assume* (5.3)*, the analogous condition for* G*,* (5.14) *with* $0 < \kappa < \infty$*, and* $p = 1$*. Then* $N_c^{>}(M_u(n))$ *is asymptotically a geometric rv with success probability*

$$p_\kappa := \frac{\kappa}{1+\kappa}.$$

5.4. Asymptotic distribution of Q_n

In this section we give the large sample distribution of Q_n in situations both of insufficient (the main case of interest) and sufficient follow-up. The resulting formulae are used to illustrate some calculations for the power of the Q_n test. The proof (in Maller et al. (2023)) is an application of Theorem 5.1 and we impose similar conditions as there. Under these conditions, the asymptotic distribution of nQ_n is geometric when $\tau_G < \tau_F$ and that of Q_n is normal when $\tau_F < \tau_G$. We need the parameters

$$\nu^A := \frac{p\int_0^{\tau_F} \overline{G}(y)\,\mathrm{d}F(y)}{1-p\overline{G}(\tau_F)} \quad \text{and} \quad \nu^B := \frac{p\int_{2\tau_F-\tau_G}^{\tau_F} \overline{G}(y)\,\mathrm{d}F(y)}{1-p\overline{G}(\tau_F)}.$$

Theorem 5.4. Case 1: *Assume $0 < p \le 1$ and $\tau_G < \tau_F \le \infty$, and also*

$$\overline{G}(\tau_G - x) = a_G(1+o(1))x^\gamma, \text{ as } x \downarrow 0, \tag{5.15}$$

where a_G and γ are positive constants. In addition, assume F has a density f in a neighbourhood of τ_G which is positive and continuous at τ_G. Then

$$\lim_{n\to\infty} P\big(nQ_n = k\big) = \frac{1}{2^{\gamma+1}}\Big(1-\frac{1}{2^{\gamma+1}}\Big)^k, \ k = 0,1,2,\ldots, \tag{5.16}$$

so nQ_n is asymptotically geometric with parameter $1/2^{\gamma+1}$, equal to $1/4$ when $\gamma = 1$.

Case 2a: *Assume $0 < p < 1$, (5.15) holds, $\tau_F < \tau_G < 2\tau_F < \infty$, and in addition*

$$\overline{F}(\tau_F - x) = a_F(1+o(1))x^\beta, \text{ as } x \downarrow 0, \tag{5.17}$$

where a_F and β are positive constants, Then

$$\frac{\sqrt{n}(Q_n - \nu^B)}{\sqrt{\nu^B(1-\nu^B)}} \xrightarrow{\mathrm{D}} N(0,1), \text{ as } n \to \infty. \tag{5.18}$$

Case 2b: *Assume $0 < p < 1$, (5.15) and (5.17) hold, and $2\tau_F < \tau_G < \infty$. Then (5.18) holds provided ν^B is replaced by ν^A.*

Case 3: *Assume $p = 1$ and (5.15) and (5.17) hold with $a_G = a_F$ and $\gamma = \beta$.*

(a) If $\tau_F < \tau_G < 2\tau_F < \infty$, then (5.18) holds as stated;

(b) If $2\tau_F < \tau_G < \infty$, then (5.18) holds with ν^B replaced by ν^A.

Remarks. (i) Case 1 with $\tau_G < \tau_F$ is a situation of insufficient follow-up, and in it nQ_n has asymptotically a finite nondegenerate limit (a geometric rv). Hence in this situation $Q_n \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$, showing that the hypothesis of insufficient follow-up will be accepted in large samples with probability approaching 1 as $n \to \infty$ when it is true. When follow-up is sufficient, i.e, in Cases 2 and 3, Q_n is ultimately normally distributed around positive levels ν^A or ν^B in large samples, and, depending on sample size, the hypothesis of insufficient follow-up will be rejected, as it should be.

(ii) The conditions in Case 1 allow for a broad range of commonly used survival and censoring distributions. The density condition on F holds for survival distributions such as the exponential, Weibull, log-logistic, log-normal, gamma, and generalised gamma. The censoring distribution G is required to have a finite right extreme in Theorem 5.4 but this is inherently satisfied in the important Case 1 and will usually be the case in practice. (5.15) holds with $\gamma = 1$ if G has a finite positive lefthand derivative at τ_G. A simple but important case is when G is uniform on an interval $[0, \tau_G]$. More generally, a generalised Pareto distribution with finite endpoint τ_G has tail function of the form $\overline{G}(\tau_G - x) = 1 - e^{-a_G x^\gamma}$ for $0 < x < \tau_G$ (see Embrechts et al. (1997), p.152; set $\gamma = -1/\xi = \tau_G$ and $a_G = \tau_G^\gamma$ in their formula). So $\overline{G}$ satisfies (5.15), precisely. See Section 10 for further discussion on the right extremes. Note that there are no restrictions on F and G in Theorem 4.1.

(iii) The specific formulae for the distributions in (5.16) and (5.18) enable calculations of the power of the Q_n test. In view of (5.16), it is more convenient to use nQ_n than Q_n. The 95-th quantile $K_{0.95}$ of the asymptotic distribution of nQ_n from (5.16), assuming the hypothesis $H_0 : \tau_G < \tau_F$ (insufficient follow-up) is true, is $K_{0.95} = K_{0.95}(p, \tau_G) = \log(0.05)/\log(3/4) - 1 = 9.41$. Thus, under H_0, we have $P(nQ_n > K_{0.95}) \approx 0.05$, for large n. Then we successively increase τ_G above τ_F, hence in the region of the alternate hypothesis, and use (5.18) in the two cases to calculate the corresponding values of $P(nQ_n > K_{0.95})$.

Thus, using ν to denote ν^A or ν^B as appropriate, we take the function of τ_G defined by

$$P(\tau_G; \nu) := P\left(N(0,1) > \min\left(\frac{K_{0.95} - n\nu}{\sqrt{n\nu(1-\nu)}}, 1.58\right)\right) \tag{5.19}$$

as an approximation to the power of the test. Keep $\tau_G > \tau_F$, and, at first, $\tau_F < \tau_G < 2\tau_F$. As τ_G increases above τ_F, ν^B increases and $P(\tau_G; \nu^B)$ increases. When $\nu^B = K_{0.95}/n$ then $P(\tau_G; \nu^B)$ reaches 0.50, and once τ_G reaches $2\tau_F$ then $\nu^B = p\int_0^{\tau_F} \overline{G}(y)\,\mathrm{d}F(y)/(1 - p\overline{G}(\tau_F))$. For τ_G values greater than this ν^B is replaced in (5.19) by $\nu^A = p\int_0^{\tau_F} \overline{G}(y)\,\mathrm{d}F(y)/(1 - p\overline{G}(\tau_F))$ and we note that $\nu^A = \nu^B$ at the transition. For larger values of τ_G, $P(\tau_G : \nu^A)$ stays constant at a value which approaches 1 as $n \to \infty$.

Assume for illustration a sample size of $n = 100$, for G a Uniform $[0, \tau_G]$ distribution, and for F a unit exponential distribution truncated at a finite value $\tau_F = 5$. Since the probability in the tail of F above 5 is less than 0.01, this is effectively assuming a unit exponential distribution for susceptible lifetimes. A graph of $P(\tau_G : \nu)$ for these parameter values is in Fig. 9.

6. Adjusting for insufficient follow-up

For slowly proliferating cancers, such as early breast and thyroid cancers, cure rates are difficult to compute due to the large number of years required for follow-up. Tai et al. (2005) perused data in the 1973-1999 edition of the SEER (2019)

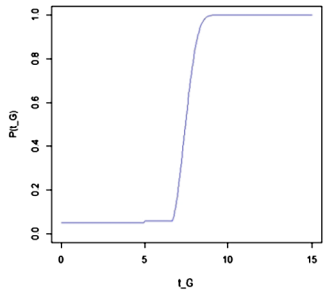


FIG 9. *Power as a function of* τ_G *for* Q_n*, with* $F \sim exp(1)$*, truncated at* $\tau_F = 5$*;* $G \sim [0, \tau_G]$*;* $n = 100$*.*

database and wrote: *The present commonly used five-year survival rates are not adequate to represent ... statistical cure ... the cancer-specific survival times of cancer patients who died of their disease from 42 cancer sites out of 49 sites were verified to follow different lognormal distributions. The threshold years* (i.e., leaving less than 2.25% uncovered) *validated for statistical cure varied ... from 2.6 years for pancreas cancer to 25.2 years for cancer of salivary gland. At the threshold year, the statistical cure rates estimated for 40 cancer sites were found to match the actuarial long-term survival rates estimated by the Kaplan-Meier method within six percentage points. For two cancer sites: breast and thyroid, the threshold years were so long that the cancer-specific survival rates could yet not be obtained because the SEER data do not provide sufficiently long follow-up.* They concluded that the minimum follow-up time required for a patient having a breast tumour with Grade II is around 26.3 years. Most patient records in the SEER data base are less extensive than that, indicating insufficient followup for many of them.

In general we may have sample data for which the KME is improper, so its right extreme value is less than 1, but it has not levelled enough to convince us that followup is sufficient. In such situations we would expect the nonparametric estimator $\widehat{p}_n$ to underestimate the true p.

In fact we have $\widehat{p}_n = \widehat{F}_n(M(n)) \xrightarrow{\text{P}} pF(\tau_H)$ as $n \to \infty$ (Maller and Zhou, 1996, Theorems 3.4 and 4.1), and when $\tau_G < \tau_F$, as we assume throughout this section, $F^*(\tau_H) = pF(\tau_G)$ is less than p, so the estimate $\widehat{p}_n$ of p is likely to be too low. We can attempt to adjust for the deficiency by adding a compensating component to $\widehat{p}_n$ to offset the bias.

This was done in Escobar-Bach and Van Keilegom (2019, 2023). They employed extrapolation techniques borrowed from extreme value theory, assuming that F belongs to the Fréchet maximum domain of attraction, to derive an estimator of the susceptible proportion which performs well for a broad range of censoring regimes and parameter values.

6.1. Adjustment in the Fréchet domain

Escobar-Bach and Van Keilegom (2019) rewrite (5.1) in the form[10]

$$\lim_{t\to\infty} \frac{\overline{F}\big((1+y\gamma)t\big)}{\overline{F}(t)} = (1+y\gamma)^{-\gamma}, \tag{6.1}$$

where $\gamma > 0$ and $y \in (0,1)$. Their adjusted estimator is defined by

$$\widehat{p}_y = \widehat{p}_n + \frac{\widehat{F}_n(M(n)) - \widehat{F}_n(yM(n))}{\widehat{y}_\gamma - 1} \tag{6.2}$$

with

$$\widehat{y}_\gamma := \frac{\widehat{F}_n(y^2M(n)) - \widehat{F}_n(yM(n))}{\widehat{F}_n(yM(n)) - \widehat{F}_n(M(n))}. \tag{6.3}$$

Under the assumption of insufficient follow-up, that is, $\tau_G < \tau_F = \infty$, they showed: *for any* $y \in (0,1)$, $\widehat{p}_y$ *is asymptotically normally distributed around a quantity* $p_y(\tau_G)$ *with a variance which depends on* y *and* τ_G, *and as* $\tau_G \to \infty$, $p_y(\tau_G) \to p$. An optimal choice of $y \in (0,1)$ is taken as the value for which $\widehat{p}_y$ is the closest to the average of a bootstrap experiment carried out using the data.

This method for the Fréchet domain was later extended by Escobar-Bach et al. (2022) to distributions in the domain of attraction of the Gumbel, as we discuss next.

6.2. Adjustment in the Gumbel domain

To carry the Escobar-Bach and Van Keilegom (2019) analysis over to the Gumbel situation an analogous bias adjustment is required. (5.1) has to be replaced with the more complicated (5.2), and the auxiliary function $f(t)$ in (5.4) has to be estimated. A modification of (6.2) which turns out to work well is to set

$$\widehat{p}_G(n,\varepsilon) = \widehat{F}_n(M(n)-\varepsilon) + \frac{\big(\widehat{F}_n(M(n)-\varepsilon/2) - \widehat{F}_n(M(n)-\varepsilon)\big)^2}{2\widehat{F}_n\left(M(n)-\varepsilon/2\right) - \widehat{F}_n(M(n)-\varepsilon) - \widehat{F}_n(M(n))}$$

(with $\widehat{p}_G(n,\varepsilon)$ replaced by $\widehat{p}_n$ if $\widehat{p}_G(n,\varepsilon) \le \widehat{p}_n$). Here ε is a free parameter which can be chosen to optimize the performance of the estimator. For consistent estimation of p by $\widehat{p}_G(n,\varepsilon)$, limits as $n \to \infty$ and $\varepsilon \downarrow 0$ are required, so ε needs to be chosen "small" in some data-dependent way. Escobar-Bach et al. (2022) use a quadratic errors criterion to choose ε in a neighbourhood where the estimator stabilises near its limiting value.

For their asymptotic analysis, Escobar-Bach et al. (2022) work within the von Mises class, assuming (5.5). Then, also assuming insufficient followup, the estimator $\widehat{p}_G(n,\varepsilon)$ is consistent in a certain sense for p and asymptotically normally distributed.

[10] We write γ where Escobar-Bach and Van Keilegom (2019) write $1/\gamma$.

Theorem 6.1 (**Consistency and Asymptotic Normality of $\widehat{p}_G(n,\varepsilon)$**). *Assume the iid censoring model holds with $\tau_G < \tau_F$, and F satisfies (5.5). Then for each $\varepsilon > 0$ the estimator $\widehat{p}_G(n,\varepsilon)$ converges in probability as $n \to \infty$ to a quantity $F(\tau_G) + C(\tau_G,\varepsilon)$, and $C(\tau_G,\varepsilon) \to C(\tau_G)$ as $\varepsilon \to 0$, where*

$$C(\tau_G) = -p\frac{F'(\tau_G)^2}{F''(\tau_G)};$$

and, further,

$$\lim_{\tau_G \uparrow \tau_F} F(\tau_G) + C(\tau_G) = p. \tag{6.4}$$

Assume in addition that $\lim_{n\to\infty} n\overline{G}(\tau_G - \delta/\sqrt{n}) = \infty$ for each $\delta > 0$ and that the integral in (1.4) is finite. Then, as $n \to \infty$, $\sqrt{n}\big(\widehat{p}_G(n,\varepsilon) - p\big)$ is asymptotically distributed as $N\big(0,\sigma^2(\varepsilon)\big)$ for each $\varepsilon > 0$, with a variance $\sigma^2(\varepsilon)$ which depends on ε. Further, $\sigma^2(\varepsilon)$ has a finite positive limit as $\varepsilon \to 0$.

Remarks. (i) As outlined in Subsection 5.3, (5.5) implies the von Mises condition, sufficient for F to belong to the Gumbel maximum domain of attraction. The condition $\lim_{n\to\infty} n\overline{G}(\tau_G - \delta/\sqrt{n}) = \infty$ is satisfied in most realistic situations; for example, if G is uniform on $[0,\tau_G]$. Theorem 4.2.3, p.82, of Gill (1980) is an important tool in the proof of Theorem 6.1.

(ii) Most of the common survival distributions are in the Gumbel domain of attraction, contained as submodels of the generalised F distribution discussed in Section 7. In particular, the generalised gamma distribution is a submodel of the generalised F for certain choices of parameters. Sub-models of the generalised gamma in the Gumbel domain of attraction include the exponential, gamma, Weibull, log-normal and Rayleigh distributions. The heavy-tailed distributions in the Fréchet domain of attraction are less common but include Pareto, log-gamma and log-logistic. Formulae for these distributions and the relations between them are in Section 7 and in more detail in Zhao (2023). Table 1 lists them and which domain of attraction, Fréchet or Gumbel, each is in.

There are lifetime distributions satisfying von Mises' condition that are not in the generalized F distribution classes, for example, the Gompertz distribution (but this is also discussed in Section 7).

Simulations in Escobar-Bach and Van Keilegom (2019) and Escobar-Bach et al. (2022) showed that both estimators $\widehat{p}_y$ and $\widehat{p}_G(n,\varepsilon)$ behaved well (were close to the true p and approximately normally distributed around it, for reasonable values of n and ε) for data with *insufficient* follow-up, and, further, $\widehat{p}_G(n,\varepsilon)$ even behaved well for data with *sufficient* follow-up, in the sense that very little adjustment to p was made in this situation, as should be the case.

7. Some parametric survival models for cure

This section gives an overview of a variety of survival distributions, with emphasis on those that can be fitted using R with a cured component allowed for, and consequently, when combined with the ideas outlined above, allow some very

sophisticated analyses of survival data. Examples of doing this, selecting between models, and fitting covariates, are in Zhao (2023). Relevant to Sections 5 and 6, we check which domains of maximal attraction the distributions are in (Table 1).

7.1. The generalised gamma distribution

The three-parameter *generalised gamma distribution* (GGD, Stacy (1962)) contains as sub-models the exponential, gamma, Weibull, log-normal, and Rayleigh distributions. Its density function is

$$f_{gg}(t) = \frac{\alpha}{\Gamma(\gamma)} t^{\alpha\gamma-1}\lambda^{\alpha\gamma}\exp\big(-(\lambda t)^{\alpha}\big),\ t>0, \tag{7.1}$$

with parameters $(\alpha, \lambda, \gamma)$, all positive. To counter computational problems, Prentice (1974) proposed the *log-gamma* and *extended generalised gamma distributions* (EGGD), obtained by the following reparameterisation of (7.1):

$$\mu_e := \frac{1}{\alpha}\log\gamma - \log\lambda, \quad \sigma_e := \frac{1}{\alpha\sqrt{\gamma}}, \quad \beta := \frac{1}{\sqrt{\gamma}}. \tag{7.2}$$

When T has the pdf in (7.1) and $Z = (\alpha\log(\lambda T) - \log\gamma)\sqrt{\gamma}$, the pdf of Z is

$$f_Z(z;\gamma) = \frac{\gamma^{\gamma-1/2}}{\Gamma(\gamma)}\exp\big(\sqrt{\gamma}z - \gamma\exp(z/\sqrt{\gamma})\big). \tag{7.3}$$

Using Stirling's approximation, we can see that this converges to the standard normal density as $\gamma\to\infty$. So the pdf of T can be written as $f_T(t) =$

$$\frac{|\beta|}{t\sigma_e\beta^{2\beta^{-2}}\Gamma(\beta^{-2})}\exp\left(\beta^{-2}\left(\beta\frac{\log t-\mu_e}{\sigma_e} - \exp\left(\beta\frac{\log t-\mu_e}{\sigma_e}\right)\right)\right), \tag{7.4}$$

where now the cases $\beta\le 0$ are allowed. In this formulation $\beta>0$ indicates that Z has the pdf $f_Z(z;\gamma)$ in (7.3), $\beta=0$ indicates that Z has a standard normal distribution, and $\beta<0$ indicates that $-Z$ has the pdf $f_Z(z;\gamma)$ in (7.3).

Lifetime distributions that are not sub-models of the generalised gamma include the log-logistic, Gompertz, inverse Gaussian, and the truncated (on the left of 0) normal distribution. The log-logistic is a sub-model of the generalised-F distribution, which is discussed next.

7.2. The generalised F distribution

Consider a location and scale model with the errors ε following the *log F-distribution*, i.e., being the logarithm of a random variable with Fisher's F-distribution $F(2s_1, 2s_2)$, having degrees of freedom $2s_1>0$ and $2s_2>0$. Thus $\exp(\varepsilon)\sim F(2s_1,2s_2)$. Introduce parameters μ and $\sigma>0$ and set

$$\log T^* = \mu + \sigma\varepsilon. \tag{7.5}$$

Then T^* has the four-parameter *generalised F-distribution* (GFD) with pdf $f_{GF}(t;\mu,\sigma,s_1,s_2) =$

$$\frac{\Gamma(s_1+s_2)}{t\sigma\Gamma(s_1)\Gamma(s_2)}\left(\frac{s_1}{s_2}\exp\left(\frac{\log t-\mu}{\sigma}\right)\right)^{s_1}\left(1+\frac{s_1}{s_2}\exp\left(\frac{\log t-\mu}{\sigma}\right)\right)^{-s_1-s_2}. \tag{7.6}$$

(The $\log F$ distribution itself is also known as the two-parameter GFD.) Referring to (7.2), the relationship between the parameters of the EGGD and those of the 4-parameter GFD is: $\mu_e = \mu$; $\sigma_e = \sigma/\sqrt{s_1}$; $\beta = 1/\sqrt{s_1}$. Of course s_1 is strictly positive, and $\beta > 0$ here. However, the EGGD with pdf in (7.4) can be viewed as a special case of the GFD (Peng et al. (1998)). The relations are as follows. When $s_1 \to \infty$, the GFD reduces to the EGGD with $\beta < 0$; when $s_2 \to \infty$, it reduces to the EGGD with $\beta > 0$; when both s_1 and s_2 approach infinity (in either order), it reduces to the log-normal distribution. Specifically, when $s_2 \to \infty$, we obtain from (7.6) the pdf (7.1) with parameters $\alpha = \sigma^{-1}$, $\lambda = s_1^{\sigma}e^{-\mu}$ and $\gamma = s_1$. Alternatively, defining $\sigma_e = \sigma/\sqrt{s_2}$ and letting $s_1 \to \infty$, (7.6) gives $\lim_{s_1\to\infty} f_{GF}(t) =$

$$\frac{s_2^{s_2-1/2}}{t\sigma_e\Gamma(s_2)}\exp\left(-s_2^{1/2}\frac{\log t-\mu}{\sigma_e}-s_2\exp\left(-s_2^{-1/2}\frac{\log t-\mu}{\sigma_e}\right)\right). \tag{7.7}$$

Letting $\mu_e = \mu$ and $\beta = -s_2^{-1/2}$, this is an EGGD with pdf (7.4).

A summary of the relationships between the GFD and the sub-models discussed so far is in Table 1. The table contains two other sub-models of the GFD, the log-logistic and members of the Burr families, discussed in Subsections 7.4 and 7.5. The notation is as in (7.6) and (7.7). The final column of the table indicates which domain of maximal attraction (DoA) the distribution is in.

TABLE 1
Generalised F submodels with their maximum domain of attraction (DoA); F=Fréchet, G=Gumbel.

Lifetime Distribution	s_1	s_2	μ	σ	DoA
Generalised F (s_1,s_2,μ,σ)	s_1	s_2	μ	σ	F
Generalised Gamma (α,λ,γ)	γ	∞	$\log(\gamma^{1/\alpha}/\lambda)$	α^{-1}	G
Weibull (α,λ)	1	∞	$-\log\lambda$	α^{-1}	G
Exponential (λ)	1	∞	$-\log\lambda$	1	G
Rayleigh (λ)	1	∞	$-\log\lambda$	0.5	G
Gamma (λ,γ)	γ	∞	$\log(\gamma/\lambda)$	1	G
Generalised Gamma $(\beta>0,\mu_e,\sigma_e)$	β^{-2}	∞	μ_e	$\sigma_e\beta^{-1}$	G
Log-Normal (μ_e,σ_e)	∞	∞	μ_e	$\sigma_e\sqrt{s_1}$	G
Generalised Gamma $(\beta<0,\mu_e,\sigma_e)$	∞	β^{-2}	μ_e	$\sigma_e\lvert\beta\rvert^{-1}$	G
Inverse Weibull (α_w,λ_w)	∞	1	$\log\lambda_w$	α_w^{-1}	F
Type III Gen. Log-Logistic (s,μ,σ)	s	s	μ	σ	F
Type II Gen. Log-Logistic (a,μ_1,σ_1)	1	a	$\log(e^{\mu_1}a^{-\sigma_1})$	σ_1	F
Log-Logistic (μ,σ)	1	1	μ	σ	F
(Log) Type XII Burr (k,c,μ_b,σ_b)	1	k	$\log(e^{\mu_b}k^{-\sigma_b})$	$c\sigma_b$	F
(Log) Type III Burr $(k,c,\mu_{b3},\sigma_{b3})$	k	1	$\log(e^{\mu_{b3}}k^{\sigma_{b3}}$	$c\sigma_{b3}$	F

7.3. Reparameterisation of the generalised F

Parameter estimation for the GFD can be unstable when the parameters are near the boundary points, s_1 or $s_2 \to \infty$. Prentice (1975) reparameterised the model, replacing s_1 and s_2 with P_F and β, where $P_F \geq 0$, $\beta \in \mathbb{R}$,

$$s_1 = 2\left(\beta^2 + 2P_F + \beta c_F\right)^{-1}, \quad s_2 = 2\left(\beta^2 + 2P_F - \beta c_F\right)^{-1}, \tag{7.8}$$

and $c_F = \left(\beta^2 + 2P_F\right)^{1/2}$. Recall that if $\varepsilon = (\log T^* - \mu)/\sigma$ follows a two-parameter GFD with parameters s_1 and s_2, then T^* has pdf of the form (7.6) with parameters s_1, s_2, μ, σ. After the reparameterisation, the convention is to redefine the two-parameter generalised F random variable as $\varepsilon = (\log T^* - \mu)/c_F^{-1}\sigma$, and hence $\log T^* = \mu + c_F^{-1}\sigma\varepsilon := \mu_e + \sigma_e$, where $\mu_e = \mu$ and $\sigma_e = \left(\beta^2 + 2P_F\right)^{-1/2}\sigma$.

This further reparameterisation of the scale parameter is consistent with the parameters in an EGGD. The relationships between the sub-models of the GFD and this reparameterised GFD are in Table 2.

TABLE 2
Reparameterised generalised F submodels.

Lifetime Distribution	β	P_F	μ_e	σ_e
Generalised F (β, P, μ, σ)	β	P_F	μ_e	σ_e
Generalised Gamma (β, μ_e, σ_e)	β	0	μ_e	σ_e
Weibull (α, λ)	1	0	$-\log\lambda$	α^{-1}
Exponential (λ)	1	0	$-\log\lambda$	1
Rayleigh (λ)	1	0	$-\log\lambda$	0.5
Gamma (λ, γ)	$\gamma^{-1/2}$	0	$\log(\gamma/\lambda)$	$\gamma^{1/2}$
Log-Normal (μ_e, σ_e)	0	0	μ_e	σ_e
Inverse Weibull (α_w, λ_w)	-1	0	$\log\lambda_w$	α_w^{-1}
Type III General Log-Logistic (s, μ, σ)	0	s^{-1}	μ	$\sqrt{\frac{\sigma^2 s}{2}}$
Type II General Log-Logistic (a, μ_1, σ_1)	$\frac{a-1}{\sqrt{a^2+a}}$	$\frac{2}{a+1}$	$\log(e^{\mu_1}a^{-\sigma_1})$	$\frac{a\sigma_1}{a+1}$
Log-Logistic (μ, σ)	0	1	μ	$\frac{\sigma}{\sqrt{2}}$
(Log) Type XII Burr (k, c, μ_b, σ_b)	$\frac{k-1}{\sqrt{k^2+k}}$	$\frac{2}{k+1}$	$\log(e^{\mu_b}k^{-\sigma_b})$	$\frac{ck\sigma_b}{k+1}$
(Log) Type III Burr $(k, c, \mu_{b3}, \sigma_{b3})$	$\frac{1-k}{\sqrt{k^2+k}}$	$\frac{2}{k+1}$	$\log(e^{\mu_{b3}}k^{\sigma_{b3}})$	$\frac{ck\sigma_{b3}}{k+1}$

7.4. The generalised logistic family

The *Type IV generalised logistic distribution* (GLD) contains Types I, II, III, defined by Johnson et al. (1995), as special cases. Johnson et al. (1995) and Nassar and Elmasry (2012) write its pdf as

$$f_{gl}(x) = \frac{\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_2)}\frac{e^{-s_1 x}}{(1+e^{-x})^{s_1+s_2}}, \; x > 0, \tag{7.9}$$

with $s_1, s_2 > 0$. The Type IV GLD is a special case of the two-parameter GFD. Specifically, the density function for $X := -\varepsilon - \log(s_1/s_2)$, where $\exp(\varepsilon) \sim$

$F(2s_1, 2s_2)$, is that in (7.9). Therefore, a GFD with parameters s_1, s_2, μ and σ is the distribution of $-X$, where X follows a Type IV generalised log-logistic distribution (GLLD) with parameters $s_1, s_2, \mu_1 = \mu - \sigma \log s_1 + \sigma \log s_2$ and $\sigma_1 = \sigma$.

When $s_1 = s_2 = s$, the Type IV generalisation (7.9) reduces to the Type III generalisation with pdf

$$f_{gl3}(x; s) = \frac{\Gamma(2s)}{\Gamma^2(s)} \frac{e^{-sx}}{(1 + e^{-x})^{2s}}, \ x > 0. \tag{7.10}$$

In that case, since (7.10) is an even function, $X = -(\log T - \mu)/\sigma$ follows the same Type III GLD as the random variable $-X = (\log T - \mu)/\sigma$. Thus a GFD with parameters μ, σ, s_1, s_2 reduces to the Type III GLD with parameters $\mu_1 = \mu, \sigma_1 = \sigma$ and s when $s_1 = s_2 = s$, whose pdf is $f_{gll3}(t; \mu, \sigma, s) =$

$$\frac{\Gamma(2s)}{t\sigma\Gamma^2(s)} \exp\left(-s\frac{\log t - \mu}{\sigma}\right)\left(1 + \exp\left(-\frac{\log t - \mu}{\sigma}\right)\right)^{-2s}. \tag{7.11}$$

Furthermore, when $s = 1$, (7.11) reduces to the standard LLD with pdf

$$f_{ll}(t) = \frac{1}{t\sigma} \exp\left(\frac{\log t - \mu}{\sigma}\right)\left(1 + \exp\left(\frac{\log t - \mu}{\sigma}\right)\right)^{-2}. \tag{7.12}$$

Setting $s_1 = 1$ in (7.9) we get the pdf

$$f_{gl}(x; s_1 = 1) = \frac{s_2 e^{-x}}{(1 + e^{-x})^{1+s_2}}. \tag{7.13}$$

which is the pdf of the Type I GLD with $a = s_2$. Thus, when $s_1 = 1$, the GFD with parameters s_1, s_2, μ and σ reduces to the Type II GLLD with parameters $a = s_2, \mu_1 = \mu + \sigma \log s_2$ and $\sigma_1 = \sigma$, whose pdf satisfies

$$f_{gll2}(t) = \frac{1}{t\sigma} \exp\left(\frac{\log t - \mu}{\sigma}\right)\left(1 + \frac{1}{s_2}\exp\left(\frac{\log t - \mu}{\sigma}\right)\right)^{-1-s_2}. \tag{7.14}$$

If we further let $s_2 \to \infty$, this converges to the Weibull pdf.

Gupta and Kundu (2010) noted that the Type I generalised logistic distribution can be regarded as a family of proportional reverse hazard distribution functions with the baseline distribution as the logistic distribution. In proportional reversed hazard models, instead of assuming the hazard functions are proportional, it is assumed that the "reversed hazard" functions, defined by "density/cdf", are proportional. However, like Type IV, the Type I generalised logistic distribution is not contained in the GFD family.

The Gumbel and Gompertz distributions are also closely related to the generalised log-logistic family. To see this, define $X_1 = Y - \log a$, with Y following a Type I GLD with parameter a. Let $a \to \infty$ in the pdf for X_1 to get

$$\lim_{a\to\infty} f_{X_1}(x_1) = \exp(-x_1 - \exp(-x_1)),$$

which is the kernel for the standard Gumbel distribution. It can be expanded with location and scale parameters as usual.

We note next that the Gompertz distribution is not a sub-model of the generalised F-distribution, but they are related as follows. Recall that if $\varepsilon = (\log T^* - \mu)/\sigma$ follows a two-parameter GFD then $X = -\varepsilon - \log(s_1/s_2)$ follows a two-parameter Type IV GLD. X reduces to the Type I GLD with parameter a when $s_1 = 1$ and $s_2 = a$, that is, $X = -\varepsilon + \log a$ has the distribution of Y in (7.13). Hence if we write

$$X_2 = \frac{\log(1/T^*) + \mu_2}{\sigma_2},$$

where $\mu_2 = (\mu + \sigma\eta + \sigma\log(\eta))$, $\sigma_2 = b\sigma$, and T^* follows a GFD with parameters s_1, s_2, μ and σ, then X_2 follows a Gompertz distribution with shape parameter η and scale parameter b when $s_1 = 1$ and $s_2 \to \infty$.

7.5. Burr distributions

Consider a random variable ε following a GFD with $s_1 = k > 0$ and $s_2 = 1$, and make the transform $Y_1 = (ke^{\varepsilon})^{1/c}$, $c > 0$. The pdf for Y_1 is

$$f_{B_1}(y) = kc\frac{y^{ck-1}}{(1+y^c)^{k+1}}, \tag{7.15}$$

which is the pdf of the *Burr Type III* distribution (Burr (1942)). The relationship between T^* in (7.5) and Y_1 is $\log T^* = \mu_{b3} + \sigma_{b3}\log Y_1$, where $\mu_{b3} = \mu - \sigma\log k$ and $\sigma_{b3} = c\sigma$. So the GFD reduces to the log of a Type III Burr distribution when $s_2 = 1$.

If instead ε follows a two-parameter GFD and we let $s_1 = 1$, $s_2 = k > 0$ and make the transformation $Y = (e^{\varepsilon}/k)^{\frac{1}{c}}$, then Y has pdf

$$f_Y(y) = kc\frac{y^{c-1}}{(1+y^c)^{k+1}},$$

which is the pdf of a *Type XII Burr* distribution. Let $\log T^* = \mu_b + \sigma_b \log Y$ where $\mu_b = \mu + \sigma\log k$ and $\sigma_b = c\sigma$. Then T^* has a four-parameter GFD with pdf

$$f_b(t) = \frac{1}{t\sigma}\exp\left(\frac{\log t - \mu}{\sigma}\right)\left(1 + \frac{1}{k}\exp\left(\frac{\log t - \mu}{\sigma}\right)\right)^{-k-1},$$

which is the pdf of the Type II GLLD (7.14) with $k = s_2$. Therefore the GFD reduces to the log of a Type XII Burr distribution, rather than to the Type XII Burr distribution, as sometimes claimed. For an application, see Ghitany et al. (2004) and associated papers.

7.6. Other distributions

We mentioned that the GGD reduces to the Weibull distribution when $\gamma = 1$, and hence the GFD reduces to the Weibull distribution when $s_1 = 1$ and $s_2 \to \infty$. On the other hand, when $s_1 \to \infty$, the GFD reduces to the EGGD with $\beta < 0$, with pdf in (7.7). If we then further take $s_2 = 1$, the distribution is, not, however, a Weibull distribution. We get

$$\lim_{s_1 \to \infty} f_{GF}(t; s_2 = 1) = t^{-\alpha_w - 1} \alpha_w \lambda_w^{\alpha_w} \exp\left(- \left(t^{-1} \lambda_w \right)^{\alpha_w} \right),$$

which is the pdf of an inverse Weibull with parameters $\alpha_w = \sigma^{-1}$ and $\lambda_w = e^{\mu}$.

For ε following a two-parameter GFD let $s_1 = 1$ and $s_2 = 1/\xi$, and set $X := \mu_p + e^{\sigma_p \varepsilon}$. The density function for X is:

$$f_X(x) = \frac{1}{\sigma_p} \left(1 + \frac{\xi(x - \mu_p)}{\sigma_p} \right)^{-\frac{1}{\xi} - 1},$$

which is the pdf of a generalised Pareto distribution (GPD) with positive parameter. The GPD with zero parameter ($\xi = \mu_p = 0$) is a limiting case, taken to be the exponential distribution, and hence is a GFD. However, the generalised F model cannot contain the GPD with negative parameter.

7.7. A Weibull model for Boag's data

Boag used a parametric approach, inferring the existence of cures in his population from his sample estimate of the proportion cured and its standard error, obtained from a lognormal mixture model fitted to the data by maximum likelihood. We observed in Section 1 that Boag in fact found an exponential mixture distribution to be a slightly better fit. Using the R package *flexsurvcure* (Amdahl (2020)) we fitted a Weibull mixture distribution to his data, thus producing an estimator of F^*. Fig. 10 shows that the Weibull is a good fit to the lifetime data (goodness of fit tests for parametric model fits are in Maller and Zhou (1996), Sect. 5.4, p.115, Geerdens et al. (2019) and Müller and Van Keilegom (2019)) and (with an estimated shape parameter of 0.88, not significantly different from 1) accords well with Boag's best fitting exponential model.

The sample also contains useful information about the censoring distribution. Fig. 10 shows the corresponding censoring KME for Boag's data, also with a Weibull distribution fitted. Censoring appears fairly uniform up to about 30 months but after that conforms better to the Weibull. Reasons for this could rest with the way the data was gathered.

We close this subsection by mentioning that papers continue to appear regularly with suggestions and evaluations of new models for lifetime data, often derived by adding to or modifying the parameters of existing models – too many to list here.

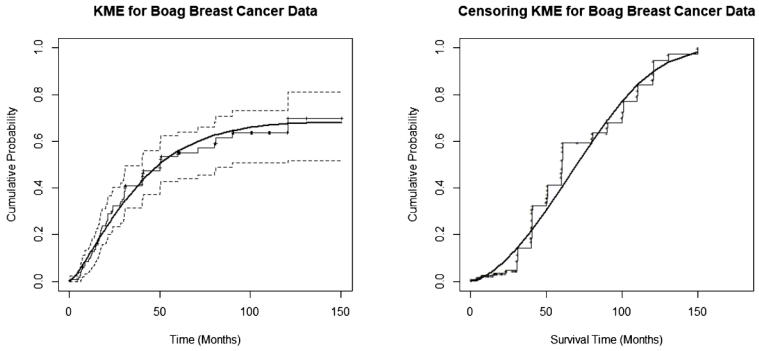


FIG 10. *KMEs for Boag Breast Cancer Data with Fitted Weibull Mixture Distributions. (a) Survival Distribution. (b) Censoring Distribution.*

8. The probability that an individual is cured

For diagnostic and prognostic purposes an important aspect of the cure model in survival analysis is the probability that an individual, having survived till a designated time t, is cured, that is, belongs to the cured component of the population. Given followup data on the individual, that is, knowing their current lifetime, we can estimate this nonparametrically or parametrically.

A formula for this probability is

$$p(t) := \frac{1-p}{1-F^*(t)}, \ t \geq 0, \tag{8.1}$$

and it can be estimated from sample data using the estimator

$$\widehat{p}_n(t) := \frac{1-\widehat{p}_n}{1-\widehat{F}_n(t)}, \ t \geq 0. \tag{8.2}$$

Recall that $\widehat{F}_n(0) = 0$ and we set $\widehat{F}_n(t) := \widehat{F}_n(M(n))$ for $t > M(n)$; see Section 1.2.

The formulae (8.1) and (8.2) are derived in Maller and Zhou (1996), Section 9.3, using a simple conditional probability argument. In this section we restrict ourselves to the case $p < 1$, since, in the alternative case, $p = 1$, there is no cured component in the population and hence no prospect of an individual's long-term cure; and we will only attempt to estimate $p(t)$ when $\widehat{F}_n(M(n)) < 1$, i.e., when the largest sample observation is censored. Recall from (1.3) that $\widehat{p}_n = \widehat{F}_n(M_u(n)) = \widehat{F}_n(M(n))$, so (8.2) assigns a value of 1 to $\widehat{p}_n(t)$ for t in $[M_u(n), M(n)]$; but as we show in the next theorem, $\widehat{p}_n(t)$ can only be guaranteed a consistent estimator of $p(t)$ when followup is sufficient.

8.1. The asymptotic distribution of $\widehat{p}_n(t)$

For the next theorem, recall the function $v(t)$ in (3.1).

Theorem 8.1 (Asymptotic distribution of $\widehat{p}_n(t)$)**.** *Assume the iid censoring model with* $0 < p < 1$ *and* (1.4) *holding. Then for each* $t \in [0, \tau_G]$, $\sqrt{n}\big(\widehat{p}_n(t) - p(t)\big)$ *is asymptotically normal with mean 0 and finite variance*

$$v_0(t) := \frac{(1-p)^2\big(v(\tau_G) - v(t)\big)}{(1 - F^*(t))^2}. \tag{8.3}$$

Remarks. (i) The result is trivial when $t = \tau_G$, because $\widehat{p}_n(\tau_G) = \widehat{p}_n(M(n)) = 1$ then, and of course the RHS of (8.3) is 0 then. But this does emphasise that an individual surviving to the maximum extent of followup is designated as cured with probability 1, *provided followup is sufficient*. Recall that finiteness of the integral in (1.4) implies the sufficient followup condition $\tau_F \leq \tau_G$. If follow-up is not sufficient, the probability of being cured will be overestimated, and an adjustment can be made as outlined in Section 6. It would be useful to incorporate this effect in a future study.

(ii) When $t = 0$, Theorem 8.1 tells us that $\sqrt{n}\big(\widehat{p}_n(0) - p(0)\big) = \sqrt{n}(\widehat{p}_n - p)$ converges in distribution to $N\big(0, (1-p)^2 v(\tau_G)\big)$ thus recovering (3.2).

(iii) To make the results practical, we need a sample estimate for the population quantity $v_0(t)$ in (8.3). We can obtain this from the consistent estimators of $v(t)$ and $v(\tau_G)$ in (3.3) and (3.4), together with use of the KME, $\widehat{F}_n(t)$, as a consistent estimator of $F^*(t)$, for each $t > 0$.

(iv) Jakobsen et al. (2020) define the *cure point* as the time at which the mortality risk in the sample reaches the same level as for the general population. This introduces the concept of *relative survival*, that is, where a comparison between observed and actuarial (assumed known) survival rates is made for deciding on "cure" status. Jakobsen et al. (2020) use the cure model to estimate the time at which the probability of cure as thus defined is sufficiently close to one, e.g., exceeding 95%, and illustrate with simulations.

(v) The EM algorithm has a natural application to the mixture cure model as a computational tool and also offers some theoretical insight. In this context $p(t)$ can be recognised as the posterior cure probability as widely used. See for example Larson and Dinse (1985), Taylor (1995) and Yu and Tiwari (2007).

8.2. The inverse problem

For the *inverse problem* we ask what length of life should a subject attain to achieve a given probability, a, say, of being cured? In other words, referring to (8.1), we seek the lifetime $T(a)$ such that

$$\frac{1-p}{1 - F^*(T(a))} = 1 - a, \tag{8.4}$$

which we expect will be estimated by the lifetime $T_n(a)$ satisfying, by (8.2),

$$\frac{1-\widehat{p}_n}{1 - \widehat{F}_n(T_n(a))} = 1 - a \in [1 - \widehat{p}_n, 1]. \tag{8.5}$$

We only evaluate (8.5) when $\widehat{p}_n \geq a$ as there is no time $T_n(a)$ for which (8.5) can be achieved when $\widehat{p}_n < a$. When $a = \widehat{p}_n$ we take $T_n(a) = 0$. We will show that for each $a < p$, $T_n(a)$ is a consistent estimator of $T(a)$, and is normally distributed around it.

We need some extra notation and appropriate assumptions for the inverse relationships. Define the left-continuous inverse functions to $\widehat{F}_n : [0, M(n)] \mapsto [0, \widehat{p}_n]$ and $F^* : [0, \tau_F] \mapsto [0, p]$ by

$$\begin{aligned} \widehat{F}_n^{\leftarrow}(u) &= \inf\{t > 0 : \widehat{F}_n(t) \geq u\};\ [0, \widehat{p}_n] \mapsto [0, M(n)] \\ F^{*,\leftarrow}(u) &= \inf\{t > 0 : F^*(t) \geq u\};\ [0, p] \mapsto [0, \tau_F], \end{aligned}$$

so, recalling that $\widehat{p}_n = \widehat{F}_n(M(n))$, we can write, for $a < p$,

$$T(a) = F^{*,\leftarrow}\Big(\frac{p-a}{1-a}\Big), \quad T_n(a) = \widehat{F}_n^{\leftarrow}\Big(\frac{\widehat{p}_n - a}{1-a}\Big). \tag{8.6}$$

Since $\widehat{p}_n \xrightarrow{\mathrm{P}} p$, $a < p$ implies $a < \widehat{p}_n$ on an event whose probability approaches 1 as $n \to \infty$. We restrict our analysis to this event, on which $T_n(a)$ is well defined and positive. The asymptotic normality proved in the next Theorem 8.2 means that 95% confidence intervals are accurate as claimed, in large samples.

Theorem 8.2 (Asymptotic distribution of $T_n(a)$)**.** *Assume the iid censoring model with $0 < p < 1$, Gill's integral (1.4) is finite, and assume in addition $F^* = pF$ has a positive continuous density f^* on $(0, \tau_F]$. Take $0 < a < p$, so that $T(a) > 0$. Then $\sqrt{n}\big(T_n(a) - T(a)\big)$ has a limiting normal distribution with mean 0 and variance*

$$\frac{(1-p)^2}{(1-a)^2\big(f^*(T(a))\big)^2}\big(v(\tau_G) - v(T(a))\big). \tag{8.7}$$

Remarks. Proofs of Theorems 8.1 and 8.2 are in the Supplementary Material to the paper. To estimate the denominator in (8.7) we can substitute $T_n(a)$ for $T(a)$ and use (3.1) and (3.4) to calculate $v(T_n(a))$ and $v(\tau_G)$. We also need to estimate the derivative $f^*(T(a))$. A similar problem arises when estimating the variance of a sample quantile; see, for example, Brookmeyer and Crowley (1982). It can be handled by using a locally smoothed estimate of F^*, or even a fitted distribution if a good parametric fit can be obtained.

8.3. Estimating the probability of cure for Boag's data

In Fig. 10 we showed the KME of the survival (lifetime) distribution with 95% confidence intervals and a Weibull mixture distribution fitted to the Boag (1949) data. In Fig. 11 we plot the function in (8.2) together with 95% confidence intervals obtained from Theorem 8.1. Rather than using the nonparametric estimates in (8.2) we could use parametric estimates from a fitted model. Fig. 11 shows the curve obtained in this way using the Weibull model fitted in Subsection 7.7.

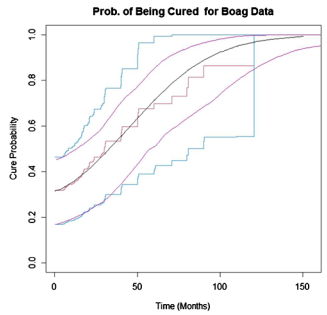


FIG 11. *Boag breast cancer data: probability of being cured with 95% confidence intervals. The step function is the estimate* (8.2)*; the smooth function is derived from the fitted Weibull.*

It can be thought of as a smoothed version of $\widehat{p}_n(t)$. Confidence intervals for it were obtained by resampling.

To conclude this section we quote from Engels et al. (2021). In a very large study, they also used a Weibull parametric version in the formula (8.1) to estimate probabilities of cure for a sample of over 10.5 million general population cancer cases, with 17 different types of cancer, in data from the US Transplant Cancer Match Study. Their aim was *to determine whether statistical models that predict a patient's probability of being cured of their cancer could inform evaluation of patients with cancer for solid organ transplantation.* Their study concluded that *it is reasonable to offer transplantation to patients who can be predicted to have a high likelihood of being cured of their cancer.* They noted the value of the technology especially to *evaluate individual patients.* We note again the importance of assessing sufficiency of followup in studies like this.

9. Some other issues

Most of our discussion has concerned the one-sample case and continuous survival time distributions. Here we consider briefly some more general situations.

9.1. Many-sample cases, and covariates

In practice, we usually have information on one or more groups (treatment groups, or stratifying variables, etc.), and/or covariates, and want to examine the effects of these. Much of Maller and Zhou (1996) is concerned with methods for handling this, and the recent book by Peng and Yu (2021) focuses on general mixture cure models with covariates. See also López-Cheda et al. (2019), Chen et al. (2023) and López-Cheda et al. (2023).

A wide variety of models which allow for treatment effects (or, generally, categorical variables) as well as continuous-valued covariates can be fitted routinely with the package R (R Core Team (2018)); see Cai et al. (2012), Jackson (2016), Amdahl (2020) and López-Cheda et al. (2021). These cover a class of generalised F models and, as a submodel, an extended generalised gamma model,

which between them include as submodels most of the usual survival distributions. Methods of distinguishing between them in a data analysis with the cure model are illustrated in Zhao (2023). Lambert (2007) gives a Stata program for fitting both mixture and nonmixture cure models, also enabling the modeling of relative survival.

When covariates are included in a model, special issues arise relating to identifiability, especially regarding sufficiency of followup, as mentioned in Section 4.4. For categorical variables, testing for sufficient followup can be done separately for each level, and continuous variables can be discretised into convenient categories and handled similarly.

9.2. Tied failure times and grouped survival data

Tied (exactly equal) failure times can occur when the assumed failure distribution F^* has a jump at one or more points, or, more commonly, when observed failure data is grouped for convenience into a smaller number of categories. The assumption throughout this paper has been that the failure distribution is continuous, but many of the theoretical and practical results continue to hold without this. Formula (1.2) for the KME remains true if the convention that an uncensored observation tied with a censored one is indexed before the censored one; see Eq. (1.7), Section 1.2, of Maller and Zhou (1996). Many of the analyses in that book (especially, concerning properties of the KME in Chapter 3) remain valid with this convention.

Most parametric models in use such as the distributions discussed in Section 7 are continuous, so this is not an issue in practice, but when observations occur in grouped form or are grouped for convenience special methods have been developed. Yu et al. (2004a) and Yu and Tiwari (2007) derive a version of the cure mixture model for grouped survival data which also takes into account relative survival (see Remark (iv) in Subsection 8.1). Their methods are applied to analyze some colorectal cancer relative survival data from the 1973–1999 edition of the SEER (2019) database using a Weibull mixture model. Cancho et al. (2020) extend proportional hazards frailty models to allow a discrete distribution for the frailty variable. In their model an individual having zero frailty can be interpreted as being immune or cured.

10. Discussion

Kaplan-Meier plots provide very strong intuition. See for example the evocative plots in Powles et al. (2021), who are expressly concerned with *determin(ing) which patients ... are cured after surgery.* A KME which levels off or plateaus at its right hand end with total mass less than 1 because the largest or a number of the largest lifetimes are censored is often the most striking feature of a Kaplan-Meier plot in the medical literature, and this we know may indicate the presence of cured or immune individuals in the population. An awareness of the importance of analysing such data properly, and the means for doing so, have grown substantially in the last few decades or so.

There are a number of aspects to consider. The use of the Cox proportional hazards model is near-ubiquitous in the analysis of survival data these days, but hazards are unlikely to be proportional if cured individuals are present. (Hazards for *susceptible* individuals may be proportional; Kuk and Chen (1992) have a method for dealing with this.) So standard analyses of survival data may be quite misleading if the possible presence of cures is overlooked – and this is apart from the valuable extra information that a cure model analysis can give.[11] On the other hand, allowing for the possible presence of cures when none are in fact present can do little or no harm, apart from the possibility of over-fitting which may lead to some bias in small data sets.

In another context, in Liu et al. (2018), the times of occurrence of four endpoints (overall survival, disease-specific survival, disease-free interval, or progression-free interval) for 11,160 patients across 33 cancer types were obtained from follow-up data files, with a view to making recommendations to clinicians regarding their patient's status. They used Q_n and a method of Shen (2000) to assess and measure followup.

Understanding how Q_n depends on the sample properties of censored survival time data, and the formulae for the exact and asymptotic distributions of Q_n we have obtained, opens the way to its more general use in the analysis of survival data with immune or cured individuals. We note that under $H_0 : \tau_G < \tau_F$, the hypothesis of insufficient follow-up, with some reasonable side conditions, the asymptotic distribution of Q_n is completely non-parametric (cf. (5.16)).

Future directions of research could usefully include issues of sufficient follow-up in competing risks analysis, and in multivariate survival analysis with cured individuals. For the latter, see Yau and Ng (2001), Chatterjee and Shih (2001, 2003), Peng et al. (2007), Yu and Peng (2008), Niu and Peng (2013), Coelho-Barros et al. (2016), Niu et al. (2018), Tawiah et al. (2020b), Oliveira et al. (2022). (Think of a study on peoples' eyes, where one or both eyes may be affected by a degenerative disease.) Other recent applications of the cure model are in Law et al. (2002) (an extension of cure models to incorporate a longitudinal disease progression marker), Zhang and Shao (2018) (assessments of prognostic utilities of biomarkers (e.g. primary tumor thickness and ulceration status of melanoma) for predicting survival of uncured patients) and Lakhal-Chaieb et al. (2020) (assessing the significance of the genetic variant in logistic and survival regressions simultaneously).

It is natural also to base tests for sufficient follow-up on the length of the interval $(M_u(n), M(n)]$, or the number of censored survival times larger than the largest uncensored survival time, or some combination or variant of these. So we might use the difference between the extremes, $M(n)-M_u(n)$, or a standardised version of this such as $R_n = 1 - M_u(n)/M(n)$, which is in $(0, 1)$. Formulae for their distributions assuming the iid censoring model are in Section 2.2. These variables measure the absolute or relative length of the level stretch of the KME

[11] Hsu et al. (2021) go so far as to suggest that some of the published survival data for immunotherapies should be re-analyzed for potential misinterpretation. They provide a method to convert inappropriate Cox hazard ratios to appropriate cure model treatment-effect estimates.

rather than a proportion of observations, as Q_n does. At present their properties remain to be investigated in detail. We note that they, like Q_n, are very sensitive to the occurrence of one or a few failures in the righthand end of the KME. This is a robustness issue to be addressed as in any statistical analysis. A test for outliers in the iid model is in Maller and Zhou (1994).

Alternative methods of assessing sufficient followup are in Shen (2000), Klebanov and Yakovlev (2007) and Xie et al. (2023). A more extensive investigation is warranted.

11. Conclusions

Here we give a summary of the discussion with main points highlighted.

• It's very common in survival analysis to encounter a KME which has levelled off at a value less than 1. This may indicate the presence of immune or cured individuals in the population — but not always — even in the absence of cures, it's possible for the right extreme of the KME to be less than 1 just by chance.

• A significance test is available for the hypothesis $H_0 : p = 1$ when a well-fitting parametric model has been found for the data (keeping in mind the one-sided nature of the test).

• A nonparametric test for $H_0 : p = 1$ is available too, using $\widehat{p}_n$, but at present we have to rely on simulated, tabulated, percentage points for its distribution. Finding a general finite sample or asymptotic distribution for $\widehat{p}_n$ when $p = 1$ remains a challenge.

• An important point is whether the KME has levelled off *sufficiently* at its right endpoint. The Q_n statistic has been developed to measure and test for this. We now understand its finite sample and asymptotic properties quite well. The "sufficient followup" condition $\tau_F \leq \tau_G$ is necessary for the convergence of the integral in (1.4) and plays an important role in many of the large-sample results in Gill (1980) and in the literature.

• We've confined our discussion mainly to the single-sample case but methods involving covariates are well developed. A wide variety of parametric models can be fitted routinely with R which makes the methodology conveniently available for practical use.

• We've also confined our discussion to medical data and survival analysis. But the methodology applies to many other kinds of time-to-event data. A wide variety of examples can be found in an internet search. Maller and Zhou (1996) use criminological data (time to re-arrest of a released prisoner, etc.) as well as medical data for illustration and many other examples can be found in the literature.

• Ignoring the possible presence of cured, immune or long-term survivors in a population not only risks losing valuable information but can result in bias and misleading conclusions. If their presence is allowed for but found not to be significant, no harm is done.

• The mixture cure model can be regarded as a special case of a competing risks analysis where death or failure of an individual may be due to a number

of possible causes. The important issue of sufficient followup is clearly relevant in this context, but has not been addressed at all, so far.

Acknowledgments

We are very grateful to two referees who read the paper carefully and gave constructive suggestions which helped us to improve it.

References

W.R. Allen. Letter to the editor – a note on conditional probability of failure when hazards are proportional. *Operations Research*, 11:658–659, 1963.

J. Amdahl. flexsurvcure: Flexible parametric mixture and non-mixture cure models for time-to-event data. *https://cran.r-project.org/web/packages/flexsurvcure/flexsurvcure.pdf*, 2020.

M. Amica and I. Van Keilegom. Cure models in survival analysis. *Ann. Rev. Statist. Appl.*, 5:311–342, 2018. MR3774750

P. Armitage. The comparison of survival curves. *J. Roy. Statist. Soc. Ser. A.*, 122:279–300, 1959.

N. Balakrishnan. Editorial. *Statist. Meth. Med. Res.*, 1999:26, 2017. MR3712215

N. Balakrishnan and S. Barui. Destructive cure models with proportional hazards lifetimes and associated likelihood inference. *Comm. Statist.: Case Studies, Data Analysis and Application*, 9:16–50, 2023.

R. Beran. Nonparametric regression with randomly censored survival data. *Technical Report, Univ. California*, 1981.

J. Berkson and P. R. Gage. Survival curve for cancer patients following treatmen. *J. Amer. Statist. Assoc.*, 47:501–505, 1952.

J.W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc. B (Method.)*, 11:15–53, 1949.

N. Breslow and J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, 2:437–453, 1974. MR0458674

R. Brookmeyer and J. Crowley. A confidence interval for the median survival time. *Biometrics*, 38:29–41, 1982.

I. W. Burr. Cumulative frequency functions. *Ann. Math. Statist.*, 13:215–232, 1942. MR0006644

C. Cai, Y. Zou, Y. Peng, and J. Zhang. smcure: An R-package for estimating semiparametric mixture cure models. *Comp. Meth. Prog. Biomed.*, 108, 2012. MR3259956

M.E. Cairns, K.P. Asante, S. Owusu-Agyei, D. Chandramohan, B.M. Greenwood, and P.J. Milligan. Analysis of partial and complete protection in malaria cohort studies. *Malaria J.*, 12:355, 2013.

V.G. Cancho, M.A.C. Macera, A.K. Suzuki, F. Louzada, and K.E.C. Zavaleta. A new long-term survival model with dispersion induced by discrete frailty. *Lifetime Data Analysis*, 26:221–244, 2020. MR4079665

M.N. Chang. Exact distribution of the Kaplan-Meier estimator under the proportional hazards model. *Statist & Probab. Letters*, 28:153–157, 1996. MR1394668

N. Chatterjee and J.H. Shih. A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics*, 57:779–786, 2001. MR1863448

N. Chatterjee and J.H. Shih. On use of bivariate survival models with cure fraction. *Biometrics*, 59:1184–1185, 2003. MR2019824

C-M. Chen, H-J. Chen, and Y. Peng. Mean residual life cure models for right-censored data with and without length-biased sampling. *Biometrical Journal*, 65, 2023. MR4603524

Y.Y. Chen, M. Hollander, and N.A. Langberg. Small-sample results for the Kaplan-Meier estimator. *J. Amer. Statist. Assoc.*, 77:141–144, 1982. MR0648036

P. Cheng and G.D. Lin. Maximum likelihood estimation of a survival function under the Koziol–Green proportional hazards model. *Statist. Probab. Lett*, 5: 75–80, 1987. MR0873939

K.C. Choi and X. Zhou. Large sample properties of mixture models with covariates for competing risks. *J. Mult. Anal.*, 82:331–366, 2002. MR1921390

E. Coelho-Barros, J.A. Achcar, and J. Mazucheli. Bivariate Weibull distributions derived from copula functions in the presence of cure fraction and censored data. *J. Data Science*, 14:295–316, 2016.

D.R. Cox. The analysis of exponentially distributed life-times with two types of failure. *J. Roy. Statist. Soc. Ser. B.*, 21:411–421, 1959. MR0114280

V. Damuzzo, L. Agnoletto, L. Leonardi, M. Chiumente, D. Mengato, and A. Messori. Analysis of survival curves: Statistical methods accounting for the presence of long-term survivors. *Frontiers in Oncology*, 9:1–6, 2019.

L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer-Verlag, New York, 2006. MR2234156

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, 1997. MR1458613

E.A. Engels, G. Haber, A. Hart, C.F. Lynch, J. Li, K.S. Pawlish, B. Qiao, K.J. Yu, and R.M. Pfeiffer. Predicted cure and survival among transplant recipients with a previous cancer diagnosis. *J. Clinic. Oncol.*, 39:4039–4048, 2021.

M. Escobar-Bach and I. Van Keilegom. Non-parametric cure rate estimation under insufficient follow-up using extremes. *J. Roy. Statist. Soc. Ser. B (Methodological)*, 81:861–880, 2019. MR4025400

M. Escobar-Bach and I. Van Keilegom. Nonparametric estimation of conditional cure models for heavy-tailed distributions and under insufficient follow-up. *Computational Statistics & Data Analysis*, 2023. MR4555680

M. Escobar-Bach, R.A. Maller, I. Van Keilegom, and M. Zhao. Estimation of the cure rate for distributions in the Gumbel maximum domain of attraction under insufficient follow-up. *Biometrika*, 109:243–256, 2022. MR4374652

V.T. Farewell. A model for a binary variable with time censored observations. *Biometrika*, 64:43–46, 1977a.

V.T. Farewell. The combined effect of breast cancer risk factors. *Cancer*, 40: 931–936, 1977b.

V.T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046, 1982.

V.T. Farewell. Mixture models in survival analysis: are they worth the risk? *Canad. J. Statist.*, 3:257–262, 1986. MR0859638

F. Fenner, D.A. Henderson, I. Arita, Z. Jezek, and Ladnyi. *Smallpox and Its Eradication*. World Health Organization, Geneva, 1988.

M.J. Frank. On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$. *Aequationes Mathematicae*, 19:194–226, 1979. MR0556722

C. Geerdens, P. Janssen, and I. Van Keilegom. Goodness-of-fit test for a parametric survival function with cure fraction. *Test*, 29:768–792, 2019. MR4140783

M.E. Ghitany, R.A. Maller, and X. Zhou. Estimating the proportion of immunes in censored samples: A simulation study. *Statistics in Medicine*, 14:39–49, 1995.

M.E. Ghitany, F. Al-Awadhi, and S.A. Al-Awadhi. Modeling the presence of long-term survivors using generalized Burr XII mixture model. *Advances and Applications in Statistics*, 4, 2004. MR2082168

R.D. Gill. *Censoring and Stochastic Integrals*. Math. Centre Tracts, 124, Amsterdam: Math. Centrum, 1980. MR0596815

A. Goldman. Survivorship analysis when cure is a possibility: A Monte Carlo study. *Statistics in Medicine*, 3:153–163, 1984.

A. Goldman. The cure model and time confounded risk in the analysis of survival and other timed events. *J. Clinic. Epidem.*, 44:1327–1340, 1991.

M. Greenwood and J.O. Irwin. The biostatistics of senility. *Human Biology*, 11 (1):1–23, 1939.

C. Gupta, J. Cobre, A. Polpo, and D. Sinha. Semiparametric Bayesian estimation of quantile function for breast cancer survival data with cured fraction. *Biometrical Journal*, 58:1164–1167, 2016. MR3545654

R. D. Gupta and D. Kundu. Generalized logistic distributions. *J. Appl. Statist. Sci.*, 51, 2010. MR2808605

J. Halpern and B.W. Jr. Brown. Cure rate models: power of the logrank and generalized wilcoxon tests. *Statistics in Medicine*, 6:483–489, 1987.

L. Haybittle. A two-parameter model for the survival curve of treated cancer patients. *J. Amer. Statist. Assoc.*, 60:16–26, 1965.

Chih-Yuan Hsu, Pei-Ying Lin, and Yu Shyr. Development and evaluation of a method to correct misinterpretation of clinical trial results with long-term survival. *JAMA Oncol*, 7:1041–1044, 2021.

C.H. Jackson. flexsurv: A platform for parametric survival modeling in R. *J. Statist. Software*, 70:1–33, 2016.

L.H. Jakobsen, T.M. Andersson, J.L. Biccler, Poulsen L.O., M.T. Severinsen, T.C. El-Galaly, and Bogsted M. On estimating the time to statistical cure. *BMC Med Res Methodol.*, 20:71–, 2020.

J-H Jeong and J. Fine. Direct parametric inference for the cumulative incidence function. *J. Roy. Statist. Soc. Ser. C (Applied Statistics)*, 55:187–200, 2006.

MR2226544
N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Volume 2*, volume 289. John Wiley and Sons, 1995. MR1299979
J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 2003. MR1924807
O. Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2021. MR4226142
E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958. MR0093867
S.N.U.A. Kirmani and J.Y. Dauxois. Testing the Koziol–Green model against monotone conditional odds for censoring. *Statist. Prob. Lett.*, 66:327–334, 2004. MR2045477
L.B. Klebanov and A.Y. Yakovlev. A new approach to testing for sufficient follow-up in cure-rate analysis. *J. Statist. Plan. Inf.*, 137:3557–3569, 2007. MR2363277
M.V. Koutras and F.S. Milienos. A flexible family of transformation cure rate models. *Statistics in Medicine*, 36:2559–2575, 2017. MR3660151
R. Kovar, I. Mala, and F. Habarta. Dependent censoring in survival regression models. *12th Int. Days of Statist. & Economics, Prague, Sept. 6-8*, 2018.
J.A. Koziol and S.B. Green. A Crameŕ-von Mises statistic for randomly censored data. *Biometrika*, 63:465–474, 1976. MR0448695
A.Y.C. Kuk and C. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika 7*, 79:531–541, 1992. MR2400249
S.W. Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, 35:139–156, 1979.
S.W Lagakos and J.S. Williams. Models for censored survival analysis: a cone class of variable-sum models. *Biometrika*, 65:181–189, 1978. MR0655435
L. Lakhal-Chaieb, J. Simard, and S. Bull. Sequence kernel association test for survival outcomes in the presence of a non-susceptible fraction. *Biostatistics*, 21:518–530, 2020. MR4120337
P.C. Lambert. Modeling of the cure fraction in survival studies. *The Stata Journal*, 7:351–375, 2007.
M.G. Larson and G.E. Dinse. A mixture model for the regression analysis of competing risk data. *Appl. Statist.*, 34:201–211, 1985. MR0827668
E.M. Laska and M.J. Meisner. Nonparametric estimation and testing in a cure model. *Biometrics*, 48:1223–1234, 1992.
N.J. Law, J.M.G. Taylor, and H. Sandler. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, 3:547–563, 2002.
J. Leão, M. Bourguignon, D.I. Gallardo, R. Rocha, and V. Tomazella. A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. *Statist. in Med.*, 39:3272–3284, 2020. MR4157839
S.X. Lee, S.K. Ng, and G.J. McLachlan. Finite mixture models in biostatistics. *In: Handbook of Statistics: Disease Modelling and Public Health, Part A, A.S.R. Rao, S. Pyne, and C.R. Rao (Eds.). Amsterdam: Elsevier*, 36:75–102, 2017. MR3838244

S.X. Lee, G.J. McLachlan, and K.L. Leemaqz. Multi-node EM algorithm for finite mixture models. *Statist. Anal. Data Mining: The ASA Data Science Journal*, 14:297–304, 2021. MR4325677

M. Lemdani and O. Pons. Estimation and tests in finite mixture models for censored survival data. *Statistics*, 29:363–388, 1997. MR1474946

M. Lemdani and O. Pons. Estimation and tests in long-term survival mixture models. Special issue on mixtures. *Comp. Stat. Data Anal.*, 41:465–479, 2003. MR1968066

K.M. Leung, R.M. Elashoff, and A.A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18:83–104, 1997.

C-S. Li, J. Sy, and J.M.G. Taylor. Identifability of cure models. *Statistics & Probability Letters*, 54:389–395, 2001. MR1861384

J. Liu, T. Lichtenberg, K.A. Hoadley, L.M. Poisson, A.J. Lazar, A.D. Cherniack, A.J. Kovatich, C.C. Benz, D.A. Levine, A.V. Lee, L. Omberg, D.M. Wolf, C.D. Shriver, V. Thorsson, and H. Hu. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173:400–416, 2018.

A. López-Cheda, M.A. Jácome, I. Van Keilegom, and R. Cao. Nonparametric covariate hypothesis tests for the cure rate in mixture cure models. *Statistics in Medicine*, 39:2291–2307, 2019. MR4119732

A. López-Cheda, M.A. Jácome, and I. López-de Ullibarri. npcure: An R package for nonparametric inference in mixture cure models. *The R Journal*, 13(1): 21–41, 2021. URL https://doi.org/10.32614/RJ-2021-027.

A. López-Cheda, Y. Peng, and M.A. Jácome. Nonparametric estimation in mixture cure models with covariates. *TEST*, 32:467–495, 2023. MR4621154

R.A. Maller and S.I Resnick. Extremes of censored and uncensored lifetimes in survival data. *Extremes*, 25:1–31, 2022. MR4417409

R.A. Maller and S. Zhou. The probability that the largest observation is censored. *Journal of Applied Probability*, 30:602–615, 1993. MR1232738

R.A. Maller and X. Zhou. Testing for sufficient followup and outliers in survival data. *J. Amer. Statist. Assoc.*, 89:1499–1506, 1994. MR1310239

R.A. Maller and X. Zhou. *Survival Analysis with Long-Term Survivors*. Wiley, Chichester, 1996. MR1453117

R.A. Maller and X. Zhou. Analysis of parametric models for competing risks. *Statistica Sinica*, 12:725–750, 2002. MR1929961

R.A. Maller, S.I. Resnick, and S. Shemehsavar. Splitting the sample at the largest uncensored observation. *Bernoulli*, 28:2234–2259, 2022. MR4474542

R.A. Maller, S.I. Resnick, and S. Shemehsavar. Finite sample and asymptotic distributions of a statistic for sufficient follow-up in cure models. *Canad. J. Statistics, to appear*, 2023.

G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley NY, 2nd edition, 2008. MR2392878

G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley NY, 2nd edition, 2000. MR1789474

G.J. McLachlan, S.X. Lee, and S.I. Rathnayake. Finite mixture models. *Ann. Rev. Statist. Appl.*, 6:355–378, 2019. MR3939525

F.S. Milienos. On a reparameterization of a flexible family of cure models. *Statistics in Medicine*, 41:4091–4111, 2022. MR4476478

M. Morbiduccia, A. Nardi, and C. Rossia. Classification of "cured" individuals in survival analysis: the mixture approach to the diagnostic–prognostic problem. *Comp. Stat. Data Anal.*, 41:515–529, 2003. MR1968067

U. Müller and I. Van Keilegom. Goodness-of-fit tests for the cure rate in a mixture cure model. *Biometrika*, 106:211–217, 2019. MR3912392

E. Musta, V. Patilea, and I. Van Keilegom. A presmoothing approach for estimation in semiparametric mixture cure models. *arXiv:2008.05338*, 2021. MR4474559

M. Nassar and A Elmasry. A study of generalized logistic distributions. *J. Egypt. Math. Soc.*, 20:126–133, 2012. MR3011590

R.B. Nelsen. *An Introduction to Copulas*. Springer, New York, 2006. MR2197664

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945–966, 1972.

Y. Niu and Y. Peng. A semiparametric marginal mixture cure model for clustered survival data. *Statistics in Medicine*, 32:2364–2373, 2013. MR3067389

Y. Niu, L. Song, Y. Liu, and Y. Peng. Modeling clustered long-term survivors using marginal mixture cure model. *Biometrical Journal*, 60:780–796, 2018. MR3830958

Y. Nui, X. Wang, and Y. Peng. geecure: An R-package for marginal proportional hazards mixture cure models. *Comp. Meth. Prog. Biomed.*, 161:115–124, 2018.

R.P Oliveira, M.V.O Peres, E.Z. Martinez, and J.O. Achcar. A new cure rate regression framework for bivariate data based on the Chen distribution. *Statistical Methods in Medical Research*, 31:2442–2455, 2022.

M. Parsa and I. Van Keilegom. Accelerated failure time vs Cox proportional hazards mixture cure models: David vs Goliath? *Stat. Papers*, 64:835–855, 2023.

V. Patilea and I. Van Keilegom. A general approach for cure models in survival analysis. *The Annals of Statistics*, 48:2323–2346, 2020.

Y. Peng. Fitting semiparametric cure models. *Comput. Statist. & Data Analysis*, 41:481–490, 2003.

Y. Peng and K.C. Carriere. An empirical comparison of parametric and semiparametric cure models. *Biometrical Journal*, 44:1002–1014, 2002.

Y. Peng and K.B.G. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56:237–243, 2000.

Y. Peng and J.M.G. Taylor. Cure models. *In: Klein, J., van Houwelingen, H., Ibrahim, J.G., and Scheike, T.H., Eds: Handbook of Survival Analysis, Ch. 6. Chapman & Hall, Boca Raton, FL, USA.*, pages 113–134, 2014.

Y. Peng and B. Yu. *Cure Models: Methods, Applications, and Implementation.* Chapman & Hall, 2021.

Y. Peng, K.B.G. Dear, and J.W. Denham. A generalized F-mixture model for cure rate estimation. *Statistics in Medicine*, 17:813–830, 1998.

Y. Peng, J.M.G. Taylor, and B. Yu. A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Anal.*, 25:1–25, 2007.

A.V. Peterson. Bounds for a joint distribution function with fixed subdistribu-

tion functions: application to competing risks. *Proc. Natl. Acad. Sci. USA*, 73:11–13, 1976.

S.J. Pocock, S.M. Gore, and G. Kerr. Long-term survival analysis: the curability of breast cancer. *Statistics in Medicine*, 1:93–104, 1982.

T. Powles, Z.J. Assaf, N. Davarpanah, R. Banchereau, B.E. Szabados, K.C. Yuen, P. Grivas, M. Hussain, S. Oudard, J.E. Gschwend, P. Albers, D. Castellano, H. Nishiyama, S. Daneshmand, S. Sharma, B.G. Zimmermann, H. Sethi, A. Aleshin, M. Perdicchio, J. Zhang, D.S. Shames, V. Degaonkar, X. Shen, C. Carter, C. Bais, J. Bellmunt, and S. Mariathasan. ctDNA guiding adjuvant immunotherapy in urothelial carcinoma. *Nature*, 595:432–437, 2021.

R. L. Prentice. A log gamma model and its maximum likelihood estimation. *Biometrika*, 61:539–544, 1974.

R.L. Prentice. Discrimination among some parametric models. *Biometrika*, 62: 607–614, 1975.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statist. Comp., Vienna, 2018. URL https://www.R-project.org/.

S.I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer, New York, 2008. Reprint of the 1987 original.

K. Rufibach, L. Grinsted, J. Li, H.J. Weber, C. Zheng, and J. Zhou. Quantification of follow-up time in oncology clinical trials with a time-to-event endpoint: Asking the right questions. *Pharmaceutical Statistics*, 22:671–691, 2023.

L. Rutquist and A. Wallgren. Is breast cancer a curable disease? *Cancer*, 53: 1793–1800, 1984.

L. Rutquist and A. Wallgren. Long-term survival of 458 young breast cancer patients. *Cancer*, 55:658–665, 1985.

W.C. Safari, I. López-de Ullibarri, and M.A. Jácome. Nonparametric kernel estimation of the probability of cure in a mixture cure model when the cure status is partially observed. *Statist. Methods in Medical Res.*, 2022.

W.C. Safari, I. López-de Ullibarri, and M.A. Jácome. Latency function estimation under the mixture cure model when the cure status is available. *Lifetime Data Analysis*, 2023.

G. Salvadori, C. DeMichele, N.T. Kottegoda, and R. Rosso. *Extremes in Nature: An Approach using Copulas*. Springer Nature, 2007.

DCCPS SEER, National Cancer Institute. Surveillance research program, released April 2019, based on the November 2018 submission. 2019.

Pao-sheng Shen. Testing for sufficient follow-up in survival data. *Statist. Prob. Letters*, 49:313–322, 2000.

A. Sklar. Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

R. Sposto, H.N. Sather, and S.A. Baker. A comparison of tests of the difference in proportions of patients who are cured. *Biometrics*, 48:87–99, 1992.

E. W. Stacy. A generalization of the gamma distribution. *Ann. Math. Statist.*, 33:1187–1192, 1962.

C.A. Struthers and V.T. Farewell. A mixture model for time to aids data with left truncation and an uncertain origin. *Biometrika*, 76:814–817, 1989.

J. Sy and J.M.G. Taylor. Estimation in a Cox proportional hazards cure model. *Biometrics*, 56:227–236, 2000.

P. Tai, E. Yu, G. Cserni, G. Vlastos, M. Royce, I. Kunkler, and V. Vinh-Hung. Minimum follow-up time required for the estimation of statistical cure of cancer patients: verification using data from 42 cancer sites in the SEER database. *BMC Cancer*, 5:48, 2005.

R. Tawiah, G.J. McLachlan, and S.K. Ng. A bivariate joint frailty model with mixture framework for survival analysis of recurrent events with dependent censoring and cure fraction. *Biometrics*, 76:753–756, 2020a.

R. Tawiah, G.J. McLachlan, and S.K. Ng. Mixture cure models with time-varying and multilevel frailties for recurrent event data. *Statistical Methods in Medical Research*, 29:1368–1385, 2020b.

J.M.G. Taylor. Semi-parametric estimation in failure time mixture models. *Biometrics*, 51:899–907, 1995.

A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proc. Nat. Acad. Sci.*, 72:20–22, 1975.

H.T.V. Vu, R.A. Maller, and X. Zhou. Asymptotic properties of a class of mixture models for failure data: The interior and boundary cases. *Ann. Institut. Statist. Math.*, 50:627–653, 1998.

P. Wang and S. Pal. A two-way flexible generalized gamma transformation cure rate model. *Statistics in Medicine*, 2022.

P. Xie, M. Escobar-Bach, and I. Van Keilegom. Testing for sufficient follow-up in censored survival data by using extremes. *Preprint*, 2023.

K.K.W. Yau and A.S.K. Ng. Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma. *Statistics in Medicine*, 20:1591–1607, 2001.

G. Yin and J. Ibrahim. Cure rate models: A unified approach. *Canad. J. Statist.*, 33:559–570, 2005.

B. Yu and Y. Peng. Mixture cure models for multivariate survival data. *Comput. Statist. & Data Analysis*, 52:1524–1532, 2008.

B. Yu and R.C. Tiwari. Application of EM algorithm to mixture cure model for grouped relative survival data. *Journal of Data Science*, 5:41–51, 2007.

B. Yu, R.C. Tiwari, K.A Cronin, and E.J. Feuer. Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, 23: 1733–1747, 2004a.

M. Yu, Ngayee J., N.J. Law, J.M.G. Taylor, and H.M. Sandler. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14:853–862, 2004b.

X.Q. Yu, R. DeAngelis, T.M.L. Andersson, P.C. Lambert, D.L. O'Connell, and P.W. Dickman. Estimating the proportion cured of cancer some practical advice for users. *Cancer Epidemiology*, 37:836–842, 2013.

Y. Zhang and Y. Shao. Concordance measure and discriminatory accuracy in transformation cure models. *Biostatistics*, 19:14–26, 2018.

M. Zhao. *Topics on Survival Analysis with Long-term Survivors*. PhD thesis, ANU, 2023.