Accelerated failure time vs Cox proportional hazards mixture cure models: David vs Goliath?

Motahareh Parsa [1] and Ingrid Van Keilegom [1]

[1] ORSTAT, KU Leuven, Belgium

Address for correspondence: Ingrid Van Keilegom, ORSTAT, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium.

E-mail: ingrid.vankeilegom@kuleuven.be.

Phone: (+32) 16 32 87 44.

Fax: .

Abstract: A mixture cure model relies on a model for the cure probability and a model for the survival function of the uncured subjects. For the latter, one often uses a Cox proportional hazards model. We show the identifiability of this model under weak assumptions. The model assumes that the cure threshold is the same for all values of the covariates, which might be unrealistic in certain situations. An alternative mixture cure model is the accelerated failure time (AFT) model. We also show the identifiability of this model under minimal assumptions. The cure threshold in this model depends on the covariates, which often leads to a better fit of the data. This is especially true when the follow-up period is insufficient for certain values of

the covariates. We study these two models via simulations both when the follow-up is sufficient and when it is insufficient. Moreover, the two models are applied to data coming from a breast cancer clinical trial. We show that the AFT and the Cox model both fit the data well in the region of sufficient follow-up, but differ drastically outside that region.

Key words: Accelerated failure time model; Cox proportional hazards model; Identifiability; Insufficient follow-up; Mixture cure models.

1 Introduction

When studying the time until a certain event of interest occurs, it often happens that a certain proportion of the subjects under study never experience this event. This happens e.g. when the event of interest is the recurrence of a disease, the start of a new job after a period of unemployment, or the birth of a child. In that case, the event of interest will not happen when a patient is cured of his/her disease, when someone never finds a job, or when someone never gets a child, respectively. The probability of such an event is called the cure proportion. Subjects who experience the event of interest are called susceptible or uncured, the others are called cured. Cure models are survival models that take this cure proportion into account, often depending on a vector of covariates or prognostic factors. Book-length reviews of cure models are Maller and Zhou (1996) and Peng and Yu (2021).

A popular cure model is the so-called mixture cure model, which relies on a model for the cure probability and a model for the survival function of the uncured subjects,

conditional on a set of covariates. For the cure probability (called the incidence) it is common practice to use a logistic model, as this is easy to estimate and often provides a good fit to the data. For the survival function of the uncured subjects (called the latency), a variety of models can be used, ranging from fully parametric (see e.g. Berkson and Gage, 1952; Farewell, 1982, 1986), to semiparametric (see Patilea and Van Keilegom, 2020, among others) to completely nonparametric models (see Xu and Peng, 2014; López-Cheda et al., 2017). An overview of possible cure models can be found in Amico and Van Keilegom (2018). However, although a broad spectrum of models for the latency is available, a popular choice for the latency is a semiparametric Cox proportional hazards (Cox, 1972) model. See e.g. Kuk and Chen (1992), Peng and Dear (2000), Sy and Taylor (2000), Lu and Ying (2004), Fang et al. (2005) and Lu (2008), among others, for papers who studied the Cox mixture cure model. The popularity of the model is thanks to its counterpart in the case without cure fraction.

Despite the many advantages of this model, like its easy interpretation and the availability of software, it is well known that the Cox model also suffers from some important drawbacks, both in the case with and without cure fraction. The most important one is the assumption of proportional hazards, which might be wrong in practice, leading then to serious bias and loss of power when estimating or making inference about the effect of a given prognostic factor on mortality (see Abrahamowicz et al., 1996). One needs to check the proportionality assumption when fitting a Cox model (see Zhang, 2002). In Yu et al. (2004) parametric cure models and semi-parametric proportional hazards (PH) cure models are applied to population-based grouped survival data to investigate the sensitivity of the parameter estimates.

An alternative to using a Cox model for the susceptible subjects is to use a semi-

parametric accelerated failure time (AFT) model (see e.g. Kalbfleisch and Prentice, 2011; Cox and Oakes, 1984). In the context with cure fraction, this model has not been studied a lot in the literature so far. We refer to Li and Taylor (2002), Zhang and Peng (2007), and Lu (2010). The latter paper proposed a kernel-based maximum likelihood estimation method for the AFT mixture cure model. The AFT model specifies a direct relation between the logarithm of the survival time of the susceptible subjects and the explanatory variables, like in any multiple linear regression model. Also, the AFT model does not need the assumption of proportional hazards, and the interpretation of the results is clearer and easier, since it models directly the effect of explanatory variables on the survival. In addition, we can estimate quantities like the mean, the mean residual lifetime and the median lifetime (or in general any quantile) directly and straightforwardly. Orbe et al. (2002) proposed a method for a censored linear regression model and compared its performance with the ones based on PH and AFT parametric models. Also, we refer to Patel et al. (2006) for more discussion on the comparison between the Cox and the AFT model in the absence of a cure fraction.

The contributions of this paper are as follows. We will start by developing sufficient conditions under which the Cox and AFT mixture cure models are identifiable. These identifiability assumptions show that under the Cox model all conditional survival functions of the susceptible subjects have the same (finite) support, regardless of the values of the prognostic factors. This is a strong model assumption, since in many practical situations covariates do not only contain important information about the cure proportion but also about the event time of subjects under study, meaning that covariates corresponding to good prognosis tend to correspond to survival functions with longer support than those corresponding to bad prognosis. The Cox mixture

cure model, unfortunately, does not allow for this differentiation. On the other hand, we will show that the AFT mixture cure model is identified if the error distribution in the model has finite support. This will entail that the support of conditional survival functions under the AFT model depends on the values of the covariates, which is more realistic in many real-life situations.

A further consequence of these identifiability results is concerned with the case where the follow-up is insufficient for some regions of the covariate space. This means that not all susceptible subjects have experienced their event by the time the study ends. In that case, the Cox model will lead to biased estimators of the cure proportion in these regions (the cure proportions will be estimated too large), whereas under the AFT model we will show that provided there is a region of the covariate space where the follow-up is sufficient, tail information coming from the regions with sufficient follow-up can be transferred to regions with insufficient follow-up, leading henceforth to unbiased estimators of the cure proportions. This is an important property of the AFT model since the estimation of the cure proportions is often of major importance in practice.

The paper is organized as follows. In Section 2 the mixture cure model is formally introduced and some notations and definitions are given. Section 3 provides the definitions of the Cox and AFT mixture cure models and the identifiability conditions of these models are studied. Section 4 focuses on the consequences of applying these models and on the misleading results that can be caused by using a Cox mixture cure model. Section 5 presents the results of a simulation study in which the cure fraction is estimated via a Cox and an AFT model, both when the follow-up is sufficient and when it is insufficient. In Section 6 both models are employed to estimate the

survival curve for a breast cancer data set coming from the SEER database containing information about the stage of the cancer. Finally, Section 7 summarizes the results obtained in this paper and discusses some ideas for future research. The Appendix contains the proofs of the identifiability results.

2 The mixture cure model

In the mixture cure model the survival function $S(t|x,z) = P(T > t \mid X = x, Z = z)$ of the survival time T given a set of real-valued covariates $(X,Z) = (x,z)$ is given by

$$S(t|x,z) = 1 - p(z) + p(z)S_u(t|x), \qquad t \geq 0, \tag{2.1}$$

where $p(z) = P(B = 1 \mid Z = z)$ is the conditional probability of being uncured, $B = I(T < \infty)$ denotes the uncure status, and $S_u(t|x) = P(T > t \mid B = 1, X = x)$ is the conditional survival function for the uncured subjects. Here, the vectors x and z can contain (totally or partially) the same covariates, but they can also be completely different.

We suppose throughout this paper that the uncure probability $p(z)$ is given by

$$p(z) = \frac{\exp(\gamma^t z)}{1 + \exp(\gamma^t z)}, \tag{2.2}$$

where the vector γ contains an intercept, so the first element of the vector z is 1. Note that other parametric models for p are also possible, as long as the parameters in the model are uniquely identified.

The survival time T is subject to random right censoring, *i.e.* instead of observing T we observe the couple (Y, Δ), where $Y = \min(T, C)$ is the observed survival time, $\Delta =$

$I(T \leq C)$ is the censoring indicator, and C is the censoring time. As often, we assume that T and C are independent given the covariates $(X^t, Z^t)^t$. Let $(Y_i, \Delta_i, X_i, Z_i)$, $i = 1, \ldots, n$, be i.i.d. realizations of (Y, Δ, X, Z). Under this data generating process, the likelihood of an observation (y, δ, x, z) is given by

$$L(y, \delta, x, z) = \{p(z) f_u(y|x)\}^{\delta} \{1 - p(z) + p(z) S_u(y|x)\}^{1-\delta}, \tag{2.3}$$

where $f_u(t|x) = -(d/dt) S_u(t|x)$, since an uncensored observation ($\delta = 1$) is always uncured, whereas for a censored observation ($\delta = 0$) we do not know the cure status.

Identifiability is an important aspect of any statistical model, and especially for a mixture cure model, since, as we will see below, additional conditions are necessary to make our model identifiable. A model is identifiable if different values of the parameters generate different probability distributions of the observable variables, so of Y, Δ, X, and Z in our case. This probability distribution is given by $L(y, \delta, x, z)$.

Hanin and Huang (2014) studied the identifiability of mixture cure models, but they did not use the above commonly used definition of identifiability. Instead, they say that a mixture cure model is identifiable if different values of the parameters generate different distributions of T for given X and Z over the observation period. Moreover, they are interested in the identifiability of the function $S_u(\cdot|x, z)$ and not in the identifiability of the components of this function. Therefore, in the next section, we will consider the identifiability of the Cox proportional hazards and the AFT (accelerated failure time) mixture cure models in full detail.

3 Cox and AFT mixture cure models

The identifiability results will be valid under the following basic assumptions:

(A) (i) For all z, $0 < p(z) < 1$.

(ii) The matrices $\mathrm{Var}(X)$ and $\mathrm{Var}(Z)$ are positive definite.

Additional assumptions will be required, which we specify separately for the Cox and the AFT mixture cure model.

3.1 Cox mixture cure model

Consider the logistic mixture cure model (2.1)-(2.2). For the latency, a Cox model is considered with survival function

$$S_u(t|x) = S_0(t)^{\exp(\beta^t x)}, \tag{3.1}$$

where $S_0(t) = P(T > t \mid B = 1, X = 0)$ is the baseline conditional survival function, which is completely unspecified, and β is a vector of parameters associated with X that does not include an intercept.

We will denote the likelihood given in (2.3) by $L_{Cox}(\gamma, \beta, S_0 \mid y, \delta, x, z)$ to make clear which parameters we are considering. The identifiability result is derived under the following additional set of assumptions:

(B) (i) The function S_0 has support $[0, \tau]$ for some $\tau < \infty$.

(ii) $P(C > \tau \mid X, Z) > 0$ for almost all X and Z.

Note that condition (B)(ii) implies that τ (called the cure threshold) has to be finite, and that (B)(i) shows that $T > \tau$ if and only if $T = \infty$, and that

$$\text{Supp}(T \mid X, Z, T < \infty) = [0, \tau]$$

(where $\text{Supp}(\cdot)$ denotes the support of a variable). Hence, the support of the finite values of T for given X and Z does not depend on X and Z, and in particular, there is no differentiation in the support according to values of X and Z that correspond to good or bad prognosis. This might be unrealistic in certain medical applications for instance, where patients with a good prognosis but who die eventually, tend to live longer than patients who have a bad prognosis. Examples of such situations can be found for instance in cancer studies, where aggressive tumors lead to faster death than slowly growing tumor, or where different stages of the cancer influence not only the survival chances but also the time until patients die.

Note that assumption (B)(ii) is equivalent to imposing that $\tau_{C|X,Z} > \tau$ for almost all X and Z, where $\tau_{C|X,Z}$ is the right endpoint of the support of C for given X and Z. We do not exclude that it is possible to refine the proof of Proposition 1 below in such a way that (B)(ii) can be weakened to $\tau_{C|X,Z} \geq \tau$ (although, if this would be possible, it would definitely not be an easy refinement). This would then allow for the case where $\tau_{C|X,Z} = \tau = \infty$, and hence the Cox model would have infinite support in that case. Note however that in practice $\tau_{C|X,Z}$ is always finite, as studies can only take place for a finite length of time. Hence, in practice, τ would still need to be finite in that case.

We are now ready to state the identifiability result for the Cox mixture cure model.

Proposition 1 *Under (A) and (B), the model given by (2.1), (2.2) and (3.1) is*

identified, in the sense that if there exist parameters $\tilde{\gamma}, \tilde{\beta}, \tilde{S}_0$ *such that*

$$L_{Cox}(\gamma, \beta, S_0 \mid y, \delta, x, z) = L_{Cox}(\tilde{\gamma}, \tilde{\beta}, \tilde{S}_0 \mid y, \delta, x, z)$$

for all realizations (y, δ, x, z) *of* (Y, Δ, X, Z)*, then* $(\gamma, \beta, S_0) = (\tilde{\gamma}, \tilde{\beta}, \tilde{S}_0)$*, where* (γ, β, S_0) *are the true parameters of the model.*

The estimation of the Cox mixture cure model has been studied in many papers, which we mentioned already in the introduction. Here, we do not wish to go deeper into the inferential properties of the Cox model, since our focus is on the identifiability, the behavior under insufficient follow-up and the comparison with the AFT model.

3.2 AFT mixture cure model

Consider the logistic mixture cure model given by (2.1)-(2.2), and write $T = T^*B + \infty(1-B)$, so T^* is the survival time of the susceptible subjects. Instead of assuming a Cox model for the latency, we will now consider a semiparametric AFT model of the form

$$\log T^* = \beta_0 + \beta^t X + \epsilon, \tag{3.2}$$

where the error ϵ is independent of X and Z and its distribution is unspecified, and β is a vector of parameters associated with X. Equivalently, we can define the AFT model by specifying the survival function

$$S_u(t|x) = S_0(t \exp(-\beta^t x)), \tag{3.3}$$

where $S_0(t) = P(\exp(\beta_0 + \epsilon) > t)$ is the error survival function, which corresponds to the conditional survival function for $X = 0$.

As before, we will denote the likelihood given in (2.3) by $L_{AFT}(\gamma, \beta, S_0 \mid y, \delta, x, z)$ to indicate which parameters we are considering. Our identifiability result is derived under the following additional set of assumptions:

(C) (i) The variable $\exp(\epsilon)$ has support $[0, \tau_0]$ for some $\tau_0 < \infty$.

(ii) $P(C > \tau_0 \exp(\beta^t X) \mid X, Z) > 0$ for all $(X, Z) \in S = S_X \times S_Z$, where S_X and S_Z are such that $P(X \in S_X, Z \in S_Z) > 0$, $\mathrm{Var}(X|X \in S_X) > 0$ and $\mathrm{Var}(Z|Z \in S_Z) > 0$.

Note that assumption (B)(ii) is a requirement on almost all (X, Z), whereas assumption (C)(ii) only needs to hold true on a set of positive probability, which can be a serious relaxation in practice. Also note that assumption (C)(i) shows that

$$\mathrm{Supp}(T \mid X, Z, T < \infty) = [0, \tau_0 \exp(\beta^t X)],$$

and hence, contrary to the Cox mixture cure model, the support does depend on the covariate vector X. Large values of the acceleration factor $\exp(\beta^t X)$ slow down the process, leading hence to longer support. As explained in the previous subsection, in real-life situations the support of the finite lifetime T^* often depends on the values of the covariates or prognostic factors.

We are now ready to state the identifiability result for the AFT mixture cure model.

Proposition 2 *Under (A) and (C), the model given by (2.1), (2.2) and (3.3) is identified, in the sense that if there exist parameters $\tilde{\gamma}, \tilde{\beta}, \tilde{S}_0$ such that*

$$L_{AFT}(\gamma, \beta, S_0 \mid y, \delta, x, z) = L_{AFT}(\tilde{\gamma}, \tilde{\beta}, \tilde{S}_0 \mid y, \delta, x, z)$$

for all realizations (y, δ, x, z) of (Y, Δ, X, Z), then $(\gamma, \beta, S_0) = (\tilde{\gamma}, \tilde{\beta}, \tilde{S}_0)$, where (γ, β, S_0) are the true parameters of the model.

Although much less popular in practice than the Cox mixture cure model, the estimation of the AFT mixture cure model has been studied in a few papers so far. We refer to the introduction section for the references. In the next section, we discuss what the identifiability results in Propositions 1 and 2 imply in practice.

4 Practical consequences

From the previous sections, we know that under the semiparametric AFT model the upper bound of the support of T^* given X, denoted by $\tau_{T^*|X}$, equals $\tau_0 \exp(\beta^t X)$, so it depends on the value of X, whereas under the Cox model $\tau_{T^*|X}$ equals τ, and hence is constant. This has two major consequences.

First, in practice, it seems much more likely that the support does depend on the value of the covariates. Some individuals tend to have faster deterioration of their illness than others, depending on the characteristics (covariates) of the patients. Hence, in many situations, a support that depends on the covariates seems more realistic than a fixed support.

Second, suppose we are indeed in a situation where $\tau_{T^*|X}$ depends on X, and consider the case where the follow-up is insufficient for some (but not all) values of X, *i.e.* the upper bound of the support of C, say τ_C, is smaller than $\tau_{T^*|X}$ for some X (for simplicity we suppose that the support of C does not depend on X, but this is by no means essential). In this case, when fitting a Cox model all individuals larger than τ_C will be considered as cured, which is wrong, since there are still uncured individuals beyond the time τ_C.

On the other hand, under the AFT model, one would first transform T to $\epsilon = \log T - \beta^t X$. Since ϵ is independent of X, the support of $\exp(\epsilon)$ (which equals $[0, \tau_0]$) can be correctly identified if there is a region of the covariate space with positive mass for which $\tau_0 \exp(\beta^t X) \leq \tau_C$, *i.e.* for which $\tau_{T^*|X} \leq \tau_C$. Typically, this region contains values of the covariates that correspond to bad prognosis. This identifiability of τ_0 can then be used for all covariates, also for those for which $\tau_{T^*|X} > \tau_C$, since ϵ has the same distribution for all X.

So to summarize, under the Cox model, apart from the fact that in practice the support of T^* often depends on X, there is an additional problem occurring in the case of insufficient follow-up. If $\tau_C < \tau_{T^*|X}$ for some X, then some susceptible subjects will be incorrectly classified as cured. Under the AFT model, this problem occurs for none of the X values, as long as there is a subspace R of the support of X for which $P(X \in R) > 0$ and $\tau_{T^*|X} \leq \tau_C$ for all $X \in R$ (*i.e.* as long as there are some values of the covariates for which the follow-up is sufficient). Additionally, note that the aforementioned consequences are derived according to the properties of the AFT and the Cox model for the latency part in the mixture cure model and have nothing to do with the incidence part of the model.

We illustrate the above situation with a simple example, that is visualized in Figure 1. In this example, the support of the censoring time does not depend on the covariate, which we take one-dimensional for simplicity. Moreover, we divide the range of the covariate into two intervals, one corresponding to the values of the covariate for which the follow-up is sufficient (region A in the figure) and the other one corresponding to insufficient follow-up (region B). That the follow-up is sufficient in region A can be seen from the fact that the support of $\log T^*$ given X, which is equal to $(-\infty, \tau_{\log T^*|X}]$,

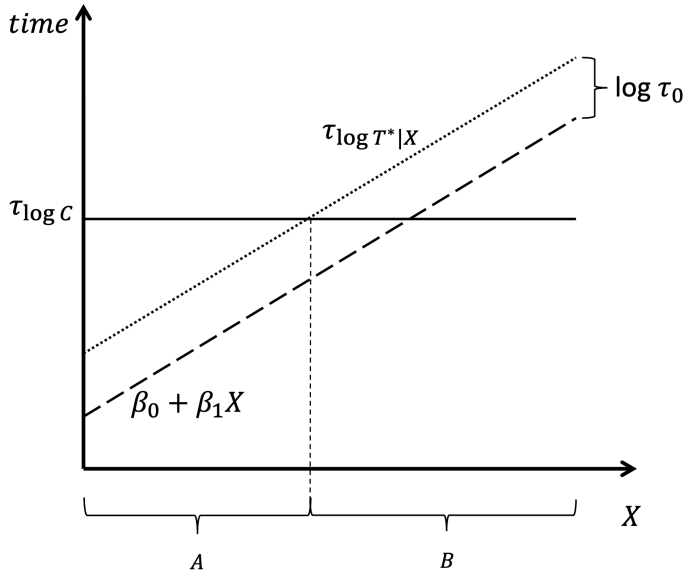

Figure 1: Graphical representation of the problem of insufficient follow-up.

is included in the support $(-\infty, \tau_{\log C}]$ of $\log C$. Since ϵ is living on $(-\infty, \log \tau_0]$, the distribution of ϵ can be identified thanks to interval A where the follow-up is sufficient. This identification then implies the identification of the distribution of T^* given X for all X, also those in interval B.

Hence, the AFT mixture cure model allows identifying the survival function further in the right tail than the Cox model, leading to better chances to identify the cure fraction. This is a very important feature in the context of cure models since cure models should only be used if one can correctly identify the cure fraction.

On the other hand, it is clear that if in reality the support of T^* does not depend on X, and if the follow-up is sufficient, a Cox model might give a better fit than an AFT model. So, we do not want to advocate the use of AFT mixture cure models in *all* circumstances. We only want to warn against blindly choosing a Cox model, not only in the absence of a cure fraction but even more when a cure fraction is present,

and the above discussion explains why it is advisable to be careful in practice.

Finally, note that the above comparison between Cox and AFT can be generalized to other mixture cure models. The assumption of a constant cure threshold is also made in other mixture cure models (see e.g. Taylor, 1995), whereas the special feature of the AFT model also shows up in any other location-scale model for the uncured individuals (of parametric, semiparametric or nonparametric nature), see e.g. Chown et al. (2020), or in the accelerated hazards model (see e.g. Zhang et al., 2011), which assumes that $S_u(t|x) = S_0(t \exp(-\beta^t x))^{\exp(\beta^t x)}$.

The next section provides a simulation study to analyze the effect of choosing the AFT or Cox model for the latency part of the mixture cure model when there is sufficient or insufficient follow-up for some values of the covariates.

5 Simulation studies

The main objective of this section is to investigate the behavior of the Cox and AFT mixture cure model in situations where the follow-up period is either sufficient or insufficient. We will do this by calculating the bias of the cure fraction estimated by means of the Cox or the AFT model, when the data are generated from either model. Here, we follow the estimation method implemented in the R-package *'smcure'* introduced by Cai et al. (2012), in which the EM algorithm is used to estimate the parameters of interest in the Cox and AFT mixture cure model.

We generate 500 data sets of size $n = 300$. We assume to have one covariate in the latency and incidence part which is the same in both parts, and is generated from a

uniform distribution on $[0,1]$. The incidence part follows a logistic model given by (2.2), with $\gamma_0 = 1$ and $\gamma_1 = -0.5$, producing an average cure fraction of 32 percent. In the latency part we consider $\beta_0 = 0$ and $\beta_1 = 2$. The other simulation settings are as follows.

Setting (I): Cox model

We suppose that in the uncured sub-population the survival time T satisfies a Cox model with cumulative baseline hazard given by

$$H_0(t) = -\log\left(\frac{e^{-\lambda t} - e^{-\lambda\tau}}{1 - e^{-\lambda\tau}}\right) \tag{5.1}$$

for $0 \leq t \leq \tau$, which is the cumulative hazard of an exponential distribution that is truncated at τ. This choice is motivated by the fact that if τ tends to infinity, the model becomes an AFT model, since (5.1) approaches the cumulative hazard λt of an exponential variable with parameter λ in that case. We take $\lambda = 2$. To generate data from this Cox mixture model, we first generate $X = Z \sim Un[0,1]$. Next, we generate a uniform variable U on $[0,1]$. If $U > p(Z)$ then define $T = \infty$. If $U \leq p(Z)$ then $U/p(Z) \sim Un[0,1]$. Following Bender et al. (2005) we can now generate the survival time T by

$$T = H_0^{-1}[-\log(U/p(Z))\exp(\beta_1 X)],$$

where $H_0^{-1}(s) = -\lambda^{-1}\log[e^{-s}(1-e^{-\lambda\tau})+e^{-\lambda\tau}]$ for $0 \leq s < \infty$. We consider $\tau = 1, 2, 4$, which corresponds to the situation where we truncate 14%, 2% and less than 0.01% of the data from the exponential distribution. The censoring time C is generated from a uniform distribution on $[0, c]$ for $c = 0.75\tau, \tau$ and 1.25τ. Hence, the follow-up is sufficient if $c \geq \tau$ and insufficient otherwise.

Setting (II): AFT model

We now suppose that the uncured individuals follow an AFT model and that the survival function S_0 of $\exp(\epsilon)$ is given by a truncated exponential distribution, given by

$$S_0(t) = \frac{e^{-\lambda t} - e^{-\lambda \tau_0}}{1 - e^{-\lambda \tau_0}} \tag{5.2}$$

for $0 \leq t \leq \tau_0$. In this way we make sure that the support of $\exp(\epsilon)$ is $[0, \tau_0]$, and at the same time we know that when τ_0 approaches infinity, the AFT model is also a Cox model, since we approach an (untruncated) exponential distribution in that case (and since for any Weibull distribution we have at the same time a Cox and an AFT model). We work with $\lambda = 2$ and $\tau_0 = 1, 2, 4$, which corresponds, as under Setting (I), to the situation where we truncate 14%, 2% and less than 0.01% of the data from the exponential distribution. If there would be no truncation at all (i.e. if $\tau_0 = \infty$), then the model is equal to the model under Setting (I) with $\tau = \infty$. To generate data from this AFT mixture model, we first generate $X = Z \sim Un[0, 1]$. Next, we generate a uniform variable U on $[0, 1]$. If $U > p(Z)$ then define $T = \infty$. If $U \leq p(Z)$ then $U/p(Z) \sim Un[0, 1]$ and we can generate T by inverting the formula in (5.2):

$$T = -\exp(\beta_1 X)\lambda^{-1} \log\left[1 - \frac{U}{p(Z)}(1 - e^{-\lambda \tau_0})\right].$$

The censoring time C is generated from a uniform distribution on $[0, d]$. Hence, the follow-up is sufficient for a given value of X if $d > \tau_0 \exp(\beta_1 X)$, i.e. if $X < \log d - \log \tau_0$ since $\beta_1 = 1$. We work with $\log d - \log \tau_0 = 0, 0.5$ and 1, corresponding to the situation where the follow-up is insufficient for all the data, for half of the data (namely those for which $X \in [0.5, 1]$), and for no data at all, since X follows a uniform distribution on $[0, 1]$.

In Table 1 we report the bias of the average cure fraction $1 - E(p(Z))$ estimated using both the Cox model and the AFT model when the data are generated under Settings

	Setting (I)				Setting (II)			
τ	c	% RC	Bias AFT	Bias PH	d	% RC	Bias AFT	Bias PH
	0.75	79	0.0492	0.2753	1	75	0.0945	0.2500
1	1	72	-0.1259	0.0869	1.65	65	0.0093	0.1352
	1.25	64	-0.1905	-0.0014	2.72	56	-0.0095	0.0544
	1.5	71	0.0088	0.1908	2	66	0.0141	0.1490
2	2	65	-0.0928	0.0584	3.30	57	-0.0157	0.0727
	2.25	58	-0.1557	-0.0021	5.44	49	-0.0054	0.0253
	3	59	-0.0177	0.0900	4	55	-0.0171	0.0552
4	4	55	-0.0266	0.0370	6.59	47	-0.0071	0.0193
	5	50	-0.0655	-0.0012	10.87	42	-0.0020	0.0042

Table 1: Bias of the estimated cure fraction under the AFT and Cox (PH) mixture cure model for different values of τ and percentages of right censoring (% RC). The survival times in the uncured sub-population are generated from a Cox proportional hazards model under Setting (I) (left), and from an AFT model under Setting (II) (right).

(I) and (II). The value of τ allows seeing the effect of misspecification of the model, since for $\tau = 1$ the misspecification is rather serious, whereas for $\tau = 4$ the models under Settings (I) and (II) are very close to an AFT and a Cox model with exponential error or baseline. On the other hand, the value of c and d allows evaluating the effect of insufficient follow-up on the estimation of the cure rate. The larger the value of these parameters, the more the follow-up is sufficient. The results show that under Setting (I) the bias of the estimated cure fraction from a Cox model is not always

close to zero, especially when the follow-up is insufficient (i.e. when $c = 0.75\tau$). On the other hand, under Setting (II) the bias obtained from an AFT model is always close to zero, also when the follow-up is (partially) insufficient. This is in line with the principle of transfer of tail information explained in Section 4. We also observe that under a misspecified model there is always some bias present. The latter bias is however smaller when an AFT model is used to estimate the cure fraction coming from a Cox mixture model, than when a Cox model is used to estimate the cure fraction coming from an AFT mixture model. In some cases the AFT model even leads to a smaller bias for the estimated cure fraction than when using the correct Cox model. This is especially the case when $c = 0.75\tau$ in the Cox model, i.e. when the follow-up is insufficient.

An application of the Cox and AFT mixture cure model on a cancer data set is presented in the next section. We will use this data set to further demonstrate the aforementioned identifiability issues in the case of insufficient follow-up.

6 Application to cancer data

In this section, we study the survival times of breast cancer patients with tumor Grade II from the 1988-2016 SEER database. The complete data set contains 178,505 patients, but we take a random sample of size $n = 30,000$ to reduce computational complexity. The data set contains information on cancer staging ranging from I to IV. The Kaplan and Meier (1958) curves for the four stages, shown in Figure 2, suggest that the follow-up period is sufficient for stage IV, but not for the three other stages, since for stage IV the Kaplan-Meier plot contains a long plateau without any events.

This is confirmed by a test developed by Maller and Zhou (1996) and carried out in Escobar-Bach et al. (2021). Hence, it will be interesting to compare the fits of the Cox and the AFT mixture cure models for these data. Some descriptive statistics for the data set are given in Table 2.

Cancer stage	Sample size	% RC
Stage I	15,624	95.44
Stage II	10,185	89.15
Stage III	3,248	73.09
Stage IV	943	43.37
All	30,000	89.25

Table 2: The sample size and the percentage of right censoring (% RC) for the sample out of the SEER breast cancer data set, and for each cancer stage separately.

As was discussed in the previous section, we expect that under the AFT model the estimated survival function for stages I to III will have a longer tail than under the Cox model. We will fit the models by creating three dummy variables for the four stages, and by utilizing the R package *smcure* introduced in Cai et al. (2012), which includes both the estimation of the AFT and the Cox mixture cure model.

Figure 2 shows the estimated survival functions for the four stages based on the AFT model, the Cox model and the Kaplan-Meier estimator. Obviously, the curves that decay the fastest (slowest) correspond to stage IV (stage I). The end of the study period, which is 347 months, is also indicated. The figure shows that before the end of the study period the three estimators are very similar, suggesting that both a Cox and an AFT model can be used for fitting the survival curves. However, there are

major differences between AFT on one hand and Cox and Kaplan-Meier on the other hand when looking at the behavior after the end of the study period. By construction, the estimators using Cox modeling and Kaplan-Meier remain constant, whereas the estimators obtained from the AFT model continue to go down for the stages where the follow-up is insufficient.

Of course, we can never be sure whether the AFT model is correctly extrapolating the survival curves beyond the end of the study, as we do not know the true curves. Therefore, we add a fourth estimator to our analysis, which is the estimator proposed by Escobar-Bach et al. (2021). The latter estimator is based on a correction of the Kaplan-Meier estimator using extrapolation techniques from extreme value theory. Hence, this estimator is constructed in a very different way than under the AFT model.

Table 3 provides the estimators of the cure fractions for the four stages of cancer according to the AFT model, the Cox model, the Kaplan-Meier estimator, and the estimator proposed by Escobar-Bach et al. (2021). We see that for stage IV there is little difference between the four estimators, which is to be expected since the follow-up is sufficient in that case. On the other hand, for the three other stages, the AFT model and the estimator of Escobar-Bach et al. (2021) provide very similar results, whereas there is a substantial difference between these two estimators on one hand, and those obtained under the Cox and Kaplan-Meier model on the other hand. This can be explained by the fact that the former two estimators extrapolate information to the tails. Under the AFT model, this extrapolation is done by transferring tail information from the region of sufficient follow-up (stage IV) to the regions of insufficient follow-up (stages I-III), while for the estimator of Escobar-Bach et al. (2021) the

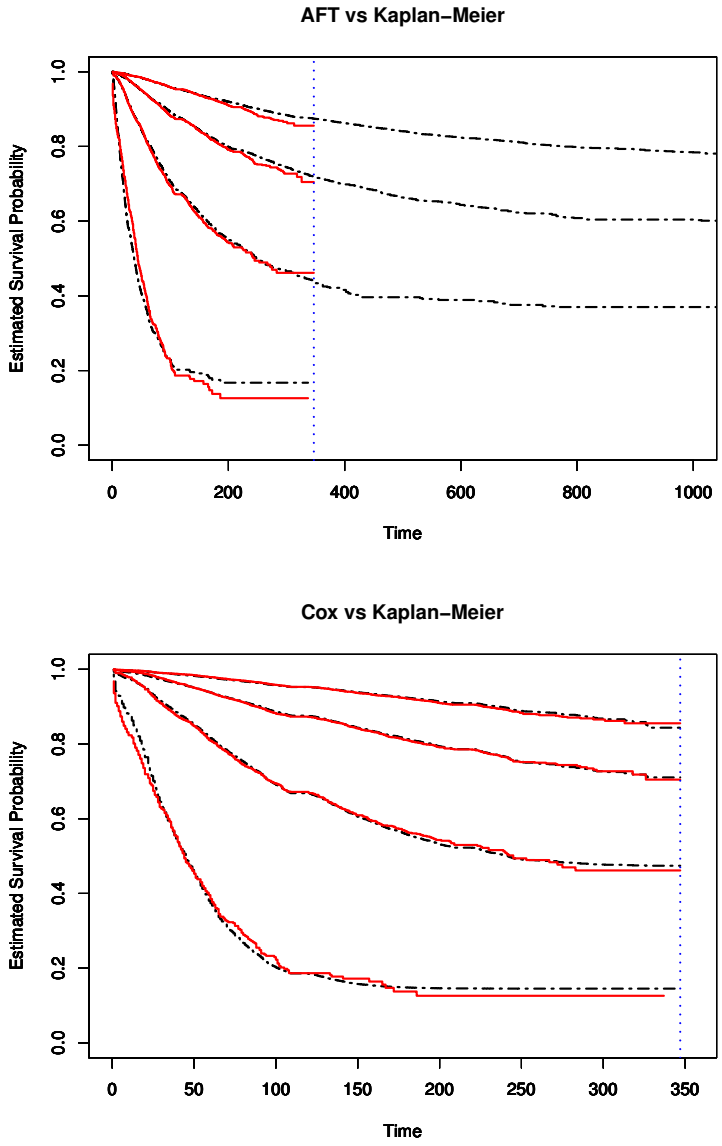

Figure 2: Estimated survival functions for the four stages of the breast cancer data using the AFT model (-·-·-) and the Kaplan-Meier estimator (—) (top); and using the Cox model (-·-·-) and the Kaplan-Meier estimator (—) (bottom). The vertical dotted line corresponds to the end of the study (347 months).

extrapolation takes place using nonparametric extreme value theory. On the other hand, the estimator based on the Cox mixture cure model and the Kaplan-Meier

estimator stay constant after the last observation, and therefore they overestimate the cure fraction as the follow-up is likely to be insufficient. The closeness between the estimators under the AFT model and those using the procedure in Escobar-Bach et al. (2021) suggests that the AFT model might be a good fit to the data.

Cancer stage	AFT	Cox	KM	EB
Stage I	0.766	0.844	0.856	0.778
Stage II	0.587	0.711	0.705	0.579
Stage III	0.370	0.474	0.462	0.382
Stage IV	0.167	0.145	0.126	0.122

Table 3: The estimation of the cure fraction at each cancer stage using the AFT model, the Cox model, the Kaplan-Meier (KM) estimator, and the estimator of Escobar-Bach et al. (2021) (EB).

7 Summary

In this paper, we studied the AFT and Cox mixture cure models. We developed conditions under which these models are identifiable, and we discussed the advantages and drawbacks of these models in practice. An important advantage of the AFT model is that it allows us to extrapolate the survival curves beyond the end of the study period. This is important in situations where the follow-up is insufficient, as it often leads to more realistic estimators of the cure proportions. Another important advantage is the fact that under the AFT model, the support of the conditional survival function depends on the value of the covariates, whereas for the Cox model

it is independent of the covariates. The latter situation is found not to be very realistic in many applications.

A possible project for future research would be the development of extensions of the AFT mixture cure model that enjoy the same important advantage of the transfer of tail information as the AFT model. These extensions can go in different directions, but a crucial property of any extension is that the error term in the model is independent of the covariates. Otherwise, the extrapolation beyond the end of the study period cannot take place. A possible promising extension is the following model

$$\Lambda(T^*) = m(X) + \sigma(X)\epsilon,$$

where ϵ and X are independent, the distribution of ϵ is unknown, Λ is a transformation of the response (can be parametric or nonparametric), and m and σ are appropriate regression and scale functions, which can again range from fully parametric to completely nonparametric. A variety of estimation procedures can be developed for this model. Note that the AFT model is a special case, by letting $\Lambda(\cdot) = \log(\cdot)$, $m(X) = \beta^t X$ and $\sigma(X) \equiv 1$.

A further promising road of research consists of developing goodness-of-fit tests for the Cox and AFT mixture cure models, to ease the choice between these two models in practice. We plan to study this in another paper. To the best of our knowledge, such a test has not been developed so far. A possible test could exist in extending the test developed by Geerdens et al. (2020). In the latter paper, the authors develop a test for the parametric form of the survival function of the susceptible subjects in the absence of covariates. An extension of this test to semiparametric survival functions that depend on covariates (e.g. under a Cox or an AFT model) would be a very useful research direction.

Acknowledgement

The authors acknowledge support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409).

Appendix: Proofs of main results

In this Appendix, we provide the proofs of Propositions 1 and 2.

Proof of Proposition 1. We need to show that if

$$\begin{aligned} &\Big\{p(z)f_u(y|x)\Big\}^{\delta}\ \Big\{1-p(z)+p(z)S_u(y|x)\Big\}^{1-\delta} \\ =\ &\Big\{\tilde{p}(z)\tilde{f}_u(y|x)\Big\}^{\delta}\ \Big\{1-\tilde{p}(z)+\tilde{p}(z)\tilde{S}_u(y|x)\Big\}^{1-\delta} \end{aligned} \tag{7.1}$$

for all realizations (y,δ,x,z) of (Y,Δ,X,Z), then $\gamma=\tilde{\gamma}, \beta=\tilde{\beta}$ and $S_0\equiv\tilde{S}_0$. Here, $p(z)=\exp(\gamma^t z)/[1+\exp(\gamma^t z)]$, $\tilde{p}(z)=\exp(\tilde{\gamma}^t z)/[1+\exp(\tilde{\gamma}^t z)]$, $S_u(y|x)=S_0(y)^{\exp(\beta^t x)}$, $\tilde{S}_u(y|x)=\tilde{S}_0(y)^{\exp(\tilde{\beta}^t x)}$, and $f_u(y|x)$ and $\tilde{f}_u(y|x)$ are the corresponding probability density functions.

First, consider $y>\tau$ and $\delta=0$. Note that

$$\begin{aligned} P(Y>\tau,\Delta=0\mid X,Z) &= P(C>\tau, C\le T\mid X,Z) \\ &= P(C>\tau, T=\infty\mid X,Z) \\ &= P(C>\tau\mid X,Z)\ (1-p(Z))>0, \end{aligned}$$

thanks to (A)(i), (B)(i) and (B)(ii). Hence, $(y>\tau,\delta=0,x,z)$ is a possible realization of (Y,Δ,X,Z). Equation (7.1) reduces in this case to $1-p(z)=1-\tilde{p}(z)$. It follows that $\gamma\equiv\tilde{\gamma}$, since $\mathrm{Var}(Z)$ is positive definite by assumption (A)(ii).

Next, consider any $0 < y \leq \tau$. Then,

$$\begin{aligned} P(Y \leq y, \Delta = 1 \mid X, Z) &= P(T \leq y, T \leq C \mid X, Z) \\ &= \int_0^y P(C > t \mid X, Z)\, dP(T \leq t \mid X, Z). \end{aligned} \tag{7.2}$$

Hence, the corresponding sub-density equals $p(Z)P(C > y \mid X, Z)f_u(y|X)$. Since the support of $f_u(\cdot|X)$ is $[0, \tau]$ and since $p(Z)P(C > \tau \mid X, Z) > 0$ thanks to assumptions (A)(i), (B)(i) and (B)(ii), it follows that the support of Y when $\Delta = 1$ and given X and Z is also $[0, \tau]$. For any $0 < y \leq \tau$ and $\delta = 1$, the likelihood contribution in (7.1) is such that $p(Z)f_u(y|X) = \tilde{p}(Z)\tilde{f}_u(y|X)$. Since $p \equiv \tilde{p} > 0$, it follows that $f_u(y|X) = \tilde{f}_u(y|X)$. Since this is true for all $0 < y \leq \tau$, it follows that $S_u(y|X) = \tilde{S}_u(y|X)$. Hence,

$$S_0(y)^{\exp(\beta^t X)} = \tilde{S}_0(y)^{\exp(\tilde{\beta}^t X)}.$$

Taking a logarithmic transformation at both sides we get

$$\exp(\beta^t X) \log S_0(y) = \exp(\tilde{\beta}^t X) \log \tilde{S}_0(y),$$

or equivalently, $(\beta - \tilde{\beta})^t X = \log\left[\log \tilde{S}_0(y) / \log S_0(y)\right]$ for all $0 < y \leq \tau$. It follows that

$$(\beta - \tilde{\beta})^t \mathrm{Var}(X)(\beta - \tilde{\beta}) = 0.$$

Since $\mathrm{Var}(X)$ is positive definite by assumption (A)(ii), this is only possible if $\beta = \tilde{\beta}$, which implies that $S_0 \equiv \tilde{S}_0$. □

Proof of Proposition 2. We need to show that if

$$\begin{aligned} &\Big\{p(z)f_u(y|x)\Big\}^{\delta} \Big\{1 - p(z) + p(z)S_u(y|x)\Big\}^{1-\delta} \\ = \;&\Big\{\tilde{p}(z)\tilde{f}_u(y|x)\Big\}^{\delta} \Big\{1 - \tilde{p}(z) + \tilde{p}(z)\tilde{S}_u(y|x)\Big\}^{1-\delta} \end{aligned} \tag{7.3}$$

for all realizations (y, δ, x, z) of (Y, Δ, X, Z), then $\gamma = \tilde{\gamma}, \beta = \tilde{\beta}$ and $S_0 \equiv \tilde{S}_0$. Here, $p(z) = \exp(\gamma^t z)/[1 + \exp(\gamma^t z)]$, $\tilde{p}(z) = \exp(\tilde{\gamma}^t z)/[1 + \exp(\tilde{\gamma}^t z)]$, $S_u(y|x) = S_0(y \exp(-\beta^t x))$, $\tilde{S}_u(y|x) = \tilde{S}_0(y \exp(-\tilde{\beta}^t x))$, and $f_u(y|x)$ and $\tilde{f}_u(y|x)$ are the corresponding probability density functions.

First, consider any $0 < y \leq \tau_0 \exp(\beta^t x)$ with $(x, z) \in S$. Then, we know from (7.2) that the conditional sub-density of the uncensored Y-values equals $p(z)P(C > y \mid X = x, Z = z)f_u(y|x) = p(z)P(C > y \mid X = x, Z = z)f_0(y \exp(-\beta^t x)) \exp(-\beta^t x)$. Since the support of $f_0(\cdot)$ is $[0, \tau_0]$ and since $p(z)P(C > \tau_0 \exp(\beta^t x) \mid X = x, Z = z) > 0$ thanks to assumptions (A)(i), (C)(i) and (C)(ii), it follows that the support of Y when $\Delta = 1$, $X = x$ and $Z = z$ is $[0, \tau_0 \exp(\beta^t x)]$. Hence, $\beta^t x = \tilde{\beta}^t x$, which is only possible if $\beta = \tilde{\beta}$ since $\mathrm{Var}(X|X \in S_X) > 0$.

Next, consider $y > \tau_0 \exp(\beta^t x)$ and $\delta = 0$ with $(x, z) \in S$. Note that

$$\begin{aligned}
& P(Y > \tau_0 \exp(\beta^t x), \Delta = 0 \mid X = x, Z = z) \\
= \;& P(C > \tau_0 \exp(\beta^t x), C \leq T \mid X = x, Z = z) \\
= \;& P(C > \tau_0 \exp(\beta^t x), T = \infty \mid X = x, Z = z) \\
= \;& P(C > \tau_0 \exp(\beta^t x) \mid X = x, Z = z)\,(1 - p(z)) > 0,
\end{aligned}$$

thanks to (A)(i), (C)(i) and (C)(ii). Hence, $(y > \tau_0 \exp(\beta^t x), \delta = 0, x, z)$ is a possible realization of (Y, Δ, X, Z). Equation (7.3) reduces in this case to $1 - p(z) = 1 - \tilde{p}(z)$. It follows that $\gamma = \tilde{\gamma}$ since $\mathrm{Var}(Z|Z \in S_Z) > 0$.

Finally, for any $0 < y \leq \tau_0 \exp(\beta^t x)$ and $\delta = 1$, the likelihood contribution in (7.3) is such that $p(z)f_u(y|x) = \tilde{p}(z)\tilde{f}_u(y|x)$. Since $p(z) = \tilde{p}(z) > 0$, it follows that $f_u(y|x) = \tilde{f}_u(y|x)$. Since this is true for all $0 < y \leq \tau_0 \exp(\beta^t x)$, it follows that $S_u(y|x) = \tilde{S}_u(y|x)$ for all $0 < y \leq \tau_0 \exp(\beta^t x)$, and hence $S_0 \equiv \tilde{S}_0$. □

References

Abrahamowicz, M., MacKenzie, T., and Esdaile, J. M. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, **91**(436), 1432–1439.

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application*, **5**, 311–342.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, **24**(11), 1713–1723.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, **108**(3), 1255–1260.

Chown, J., Heuchenne, C., and Van Keilegom, I. (2020). The nonparametric location-scale mixture cure model. *TEST*, **29**, 1008–1028.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society - Series B*, **34**(2), 187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, New York.

Escobar-Bach, M., Maller, R., Van Keilegom, I., and Zhao, M. (2021). Estimation of the cure rate for distributions in the Gumbel maximum domain of attraction under insufficient follow-up. *Biometrika (to appear)*.

Fang, H.-B., Li, G., and Sun, J. (2005). Maximum likelihood estimation in a semi-parametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, **32**(1), 59–75.

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

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**(3), 257–262.

Geerdens, C., Janssen, P., and Van Keilegom, I. (2020). Goodness-of-fit tests for univariate mixture cure rate models. *TEST*, **29**, 768–792.

Hanin, L. and Huang, L.-S. (2014). Identifiability of cure models revisited. *Journal of Multivariate Analysis*, **130**, 261–274.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.

Kuk, A. Y. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**(3), 531–541.

Li, C.-S. and Taylor, J. M. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, **21**(21), 3235–3247.

López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, **105**, 144–165.

Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics*, **60**(3), 545–574.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, **20**, 661–674.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, **91**(2), 331–343.

Maller, R. A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Orbe, J., Ferreira, E., and Núñez-Antón, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in Medicine*, **21** (22), 3493–3510.

Patel, K., Kay, R., and Rowell, L. (2006). Comparing proportional hazards and accelerated failure time models: an application in influenza. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, **5**(3), 213–224.

Patilea, V. and Van Keilegom, I. (2020). A general approach for cure models in survival analysis. *Annals of Statistics*, **48**(4), 2323–2346.

Peng, Y. and Yu, B. (2021). *Cure Models: Methods, Applications, and Implementation (1st ed.)*. CRC Press.

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**(1), 237–243.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56**(1), 227–236.

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

Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, **42**(1), 1–17.

Yu, B., Tiwari, R. C., Cronin, K. A., and Feuer, E. J. (2004). Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, **23** (11), 1733–1747.

Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, **26**(16), 3157–3171.

Zhang, J., Peng, Y., and Zhao, O. (2011). A new semiparametric estimation method for accelerated hazard model. *Biometrics*, **67**(4), 1352–1360.

Zhang, M.-J. (2002). Cox proportional hazards regression models for survival data in cancer research. In *Beam C. (eds) Biostatistical Applications in Cancer Research. Cancer Treatment and Research*, volume 113, pages 59–70. Springer, Boston.