Published in final edited form as:
Stat Med. 2023 February 10; 42(3): 209–227. doi:10.1002/sim.9610.

RECeUS: Ratio estimation of censored uncured subjects, a different approach for assessing cure model appropriateness in studies with long-term survivors

Subodh Selukar[1], **Megan Othus**[2]

[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee,

[2]Department of Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Washington,

Abstract

The need to model a cure fraction, the proportion of a cohort not susceptible to the event of interest, arises in a variety of contexts including tumor relapse in oncology. Existing methodology assumes that follow-up is long enough for all uncured subjects to have experienced the event of interest at the time of analysis, and researchers have demonstrated that fitting cure models without sufficient follow-up leads to bias. Few statistical methods exist to evaluate sufficient follow-up, and they can exhibit poor performance and lead users to falsely conclude sufficient follow-up, leading to bias, or to falsely claim insufficient follow-up, possibly leading to additional, costly data collection. We propose a new quantitative statistic (RECeUS) to evaluate whether cure models may be appropriate to apply to censored data. Specifically, we propose that the estimated proportion of censored uncured subjects in a study can be used to evaluate cure model appropriateness. We evaluated the performance of RECeUS against existing methods via simulation and with two data examples, and we observe that RECeUS displays superior performance. In simulated and real-world settings, RECeUS correctly identifies both situations in which data appear appropriate for cure modeling and when data seem inappropriate for fitting cure models.

Keywords

cure models; oncology; sufficient follow-up time; survival analysis

1 | INTRODUCTION

1.1 | Sufficient follow-up time in cure models

Researchers have been interested in estimating the fraction of patients cured of cancer for over 70 years. Early studies reported the five-year survival rate as an assessment of cure,

Correspondence: Subodh Selukar, Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA. subodh.selukar@stjude.org.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

but Boag[1] and Berkson and Gage[2] argued against this measure and introduced models that instead analyzed the cure fraction explicitly as the proportion of a cohort not susceptible or "cured" of the event of interest. Since then, cure model literature has grown with many new methods and extensions. Amico and Van Keilegom[3] describe that a key assumption in all cure models is that there is sufficient follow-up to identify model parameters.

To motivate the importance of this sufficient follow-up time, we examine the 3-arm phase 2/3 study, S1117 (clinicialtrials.gov identifier: NCT01522976[4]), which investigated combination therapy vs single-agent azacitidine for newly diagnosed myelodysplastic syndrome (MDS) patients. Anderson[5] had previously reported that approximately 40% of patients with MDS may be cured by allogeneic bone marrow transplantation. While the combination arms in S1117 failed to show sufficient efficacy to proceed to phase 3, prior research such as Anderson's paper motivated the clinical leadership to want to evaluate whether some subjects may have been cured of their disease.

Based on the data first released by the data safety monitoring committee, clinical investigators believed that results for overall survival (Figure 1) indicated a plateau at the right tail or that a nonzero fraction of subjects may be long-term survivors of the disease. However, there is heavy censoring before the plateau, so it seems there may not be adequate follow-up to fit a cure model.

1.2 | Existing methods to study cure model appropriateness

As per Amico and Van Keilegom,[3] the mixture cure framework is a popular and well-studied area in the broader cure models literature. This framework considers the population to be a mixture of cured subjects who will not experience the outcome of interest and subjects susceptible to the outcome. As a result, the cumulative distribution function (cdf) of the event times T can be written as $F(t) = (1 - \pi)F_0(t)$, where π is the cure fraction in the population and $F_0(t)$ is the cdf of the event times for uncured subjects. The existing methods for studying cure model appropriateness were developed within this framework.

In their 1996 book, Maller & Zhou outline a procedure for testing cure model appropriateness as (1) testing for the presence of immunes, or cured subjects (the true proportion of cured subjects $\pi = 0$ vs $\pi > 0$), then (2) testing for sufficient follow-up if there is reason to believe a cure fraction exists to generate the data.[6]

For the first component, testing for the presence of immunes, they present a nonparametric test for the presence of immunes, the $\hat{p}_n$ test, by comparing the value of the right tail of the Kaplan-Meier estimate based on a sample of size n against simulated critical values based on the rate of censoring and sample size.[6] They also describe a parametric version employing the deviance statistic, d_n from a given parametric model estimate on n samples, which they had proposed in an earlier paper from 1995.[7]

Maller and Zhou[8] also were among the first to study the second component, tests for sufficient follow-up time, and they focused on the supports of the censoring distribution, with cdf $G(t)$, and the event time distribution for uncured subjects. In particular, they examined $\tau_{F_0} = \min_t\{t : F_0(t) = 1\}$, the earliest time that the event time cdf for uncured subjects reaches 1, and τ_G, the analogous quantity for the censoring distribution. They identified a criterion $\tau_{F_0} < \tau_G$ as a necessary condition for valid assessment of a cure fraction. In words, this states that the longest event times for uncured subjects cannot be unobservable due to censoring.

They proposed a hypothesis test to quantify this approach, called the $\hat{\alpha}_n$ test. The test quantifies the difference between the largest event time (an estimate of τ_{F_0}) and the largest censored time (an estimate of τ_G) among n observations with the null hypothesis of $\tau_{F_0} \geq \tau_G$.

Two studies remarked on poor control of type-1 error by the $\hat{\alpha}_n$ test and each developed a different method to address the issue: one by Maller and Zhou[6] themselves, the q_n test, and one by Shen,[9] the $\tilde{\alpha}_n$ test. Both modify the test statistic of $\hat{\alpha}_n$ to improve upon it, but all continue to test the same null hypothesis $\tau_{F_0} \geq \tau_G$ for n observations. Shen's article studies these properties via simulation under a truncated Weibull distribution and demonstrates the improvements by q_n and $\tilde{\alpha}_n$ expected by theory.

However, in practice, many studies will necessarily have finite follow-up due to cost, and they may not have long enough follow-up to allow for the tail of the uncured subjects' event time distribution to be observed before being censored. This means the premise of testing based on these right tail quantities may itself be unrealistic in many research settings, and this has not been rigorously studied for these methods.

Fortunately, Yu et al[10] describe how cure fraction and median survival estimates improve as follow-up time increases without necessarily reaching the point of all failures being observed. Simulations summarized in the appendix expand on this to show that in Weibull mixture cure models, estimation with low mean-squared error and nominal confidence interval coverage can be achieved with 1% uncured remaining or longer follow-up, depending on the setting. Taken together, this motivates a different approach to quantifying sufficient follow-up time.

This paper aims to provide an approach for evaluating the appropriateness of a cure model under the setting where a nonzero fraction of uncured subjects remain at the analysis time. We describe our proposed method in Section 2.1. Section 2.2 defines the possible classification errors in claiming cure model appropriateness with this approach. We provide asymptotic properties for the proposed statistic estimated by maximum likelihood in Section 2.3 and then suggest approaches for addressing sensitivity to model misspecification in Section 2.4. We assess finite sample performance in Section 3. Then we conclude with two data examples in Section 4 and a discussion, Section 5.

2 | A NOVEL METHOD FOR ASSESSING CURE MODEL APPROPRIATENESS

2.1 | Targeting the proportion of uncured remaining

Throughout the rest of the paper, in light of the mixture cure framework, we consider the true event times T to have survival function

$$S(t; \pi, \theta) = \pi + (1 - \pi)S_{uc}(t; \theta) \tag{1}$$

with π the unknown cure fraction and a survival function $S_{uc}(t)$ belonging to common parametric families with unknown parameter vector $\theta \in \Theta \subseteq \Re^p$, $p < \infty$ (we may suppress the dependence of S and S_{uc} on π and θ in the notation for clarity). We also consider independent, uniformly sampled accrual times $A \sim \text{Unif}(0, a)$ (a known) and a known administrative censoring time τ (with $a < \tau$). These are intended to parallel the real-world context of clinical trials that often accrue until a prespecified time a and analyze at a prespecified time τ.

Based on Yu et al[10] and summaries in the appendix, we observe that estimation and inference improve as the proportion of uncured remaining decreases but may also depend on the underlying cure fraction. This motivates that the proportion of uncured remaining can be used to quantify the sufficiency of follow-up. Based on this, we propose that the quantity

$$r = \frac{S_{uc}(\tau)}{S(\tau)} = \frac{S_{uc}(\tau)}{\pi + (1 - \pi)S_{uc}(\tau)} \tag{2}$$

be used to quantify sufficient follow-up time in a mixture cure setting. We can then use the following estimator (where we replace the population quantities with suitable estimates indexed by a sample of size n):

$$\hat{r}_n = \frac{\hat{S}_{n,uc}(\tau)}{\hat{S}_n(\tau)} = \frac{\hat{S}_{n,uc}(\tau)}{\hat{\pi}_n + (1 - \hat{\pi}_n)\hat{S}_{n,uc}(\tau)}\,. \tag{3}$$

Heuristically, this ratio statistic quantifies the estimated proportion of uncured remaining at the administrative censoring time and standardizes it to the estimated overall proportion remaining and censored at τ. This standardization incorporates the cure fraction and censoring pattern in the data (as opposed to simply targeting the proportion uncured alone).

We propose to use this statistic in a method we call RECeUS (Ratio Estimation of Censored Uncured Subjects, pronounced “ree-sus”) to assess cure model appropriateness as follows:

1. Estimate the quantities $S(\tau)$, $S_{uc}(\tau)$, and π.
2. If $\hat{\pi}_n$ is small, then either a cure model is likely not a valid model or follow-up is likely insufficient for valid results.
3. If $\hat{\pi}_n$ is away from 0, then estimate r - small values of $\hat{r}_n$ represent sufficiency of follow-up time.

This procedure simultaneously addresses both (1) evaluating the presence of immunes and also (2) evaluating the sufficiency of follow-up.

This method requires specifying thresholds for $\hat{\pi}_n$ and $\hat{r}_n$. In this article, we summarize properties of RECeUS when using a threshold for $\hat{\pi}_n$ of 2.5% and a threshold for $\hat{r}_n$ of 5%: with this choice, the RECeUS method claims a cure model is appropriate when both $\hat{\pi}_n > 0.025$ and $\hat{r}_n < 0.05$, and, if one or both conditions fail, then RECeUS concludes a cure model is not appropriate. We discuss the choice of these thresholds and the impact of varying these thresholds in Sections 3 and 5 and in the Appendix D.2.

2.2 | Defining possible errors when concluding cure model appropriateness

In addition to motivating a statistic for quantifying the sufficiency of follow-up, the results of Yu et al[10] and the appendix also motivate reframing the possible errors when concluding cure model appropriateness.

In previous literature, errors have been defined in the context of hypothesis testing using type-1 and −2 errors. In tests for the presence of immunes, the null hypothesis is $\boldsymbol{\pi} = 0$, no immunes present, with the alternative of $\boldsymbol{\pi} > 0$. This leads to type-1 errors as being incorrect conclusions of the presence of immunes and type-2 errors being incorrectly failing to detect the presence of immunes. When testing for sufficient follow-up, the methods assume the presence of immunes and assess the null hypothesis $\tau_{F_0} \geq \tau_G$ against the alternative $\tau_{F_0} < \tau_G$. Type-1 errors occur when follow-up is declared sufficient when uncured subjects remain at risk, and type-2 errors occur when follow-up is not declared sufficient but no uncured subjects remain at risk.

To highlight the impact of follow-up time on cure model appropriateness, we suggest instead defining errors explicitly in terms of follow-up time in addition to event time generating parameters.

We summarize the rates at which a particular method correctly or incorrectly concludes a cure model is appropriate under a given setting. We focus on two particular settings of interest: (1) a true cure fraction $\boldsymbol{\pi} = 0$ with any amount of follow-up and (2) a nonzero cure fraction and long follow-up. In the first setting, a claim of a cure model being appropriate is incorrect, while in the second setting, the claim is correct. A third scenario also exists with a nonzero cure fraction and short or intermediate follow-up. We discuss this further in Sections 3.4 and 5. We parameterize the length of follow-up by the fraction of uncured subjects remaining at the end of the study because this notation holds across different event time generating distributions.

These scenarios differ from errors in hypothesis testing because they rely on the latent status of subjects at the end of the study in addition to data generating parameters. We further distinguish the errors by adopting the following notation. First, define the random variable $G = \mathbf{1}\{\text{Declare cure model is appropriate}\}$. Next, let $\boldsymbol{\pi} \in [0, 1]$ be the underlying cure fraction, and $u = S_{uc}(\boldsymbol{\tau}) \in [0, 1]$ be the fraction of uncured subjects remaining (have not experienced the event of interest) at the end of the study (note that uncured subjects would not be identifiable in practice as cure is considered a latent status). Then define

$$\gamma_\pi(u) \equiv Pr(G = 1; u, \pi). \tag{4}$$

A useful method for distinguishing when a cure model is and is not appropriate should have the following properties:

1. $\gamma_\pi(u) \rightarrow 1$ when $\pi > 0$ and $u \rightarrow 0$; and
2. $\gamma_\pi(u) \rightarrow 0$ when $\pi = 0$ with any value $u \in [0, 1]$

The first property states that the method increases its probability to 1 in reaching the correct conclusion that the cure model is appropriate when a cure fraction exists and the proportion of uncured subjects remaining at the end of the study decreases to 0.

An important consideration is the time at which we declare we have follow-up sufficient for practical purposes, or the largest value u for which we would practically employ a cure model when a nonzero cure fraction exists. Because u is a latent quantity, we must make this decision based on external knowledge. In their textbook, Maller & Zhou suggest $u = 0.2\%$ or smaller numbers.[6] In this article, based on our simulation studies, we summarize results using $u = 0.1\%$. This translates to expecting rates of $\gamma_\pi(0.001)$ approaching 1. We discuss our results for different u in Sections 3.4 and 5.

The second property states that the method decreases to 0 in its probability in reaching the incorrect conclusion that the cure model is appropriate when no cure fraction exists to generate the data at any length of follow-up. This translates to expecting rates of $\gamma_0(u)$ approaching 0 for any u.

Closed form expressions for $\gamma_\pi(u)$ are challenging to obtain, so, in this article, we compute the Monte Carlo estimates $\gamma_\pi(u)^*$ via simulation in Section 3. For each method, we calculate $\gamma_\pi(u)^*$ as the proportion of simulations generated with cure fraction π and follow-up corresponding to u that conclude a cure model is appropriate using that method.

2.3 | Asymptotic properties when estimating via maximum likelihood

For the estimator $\hat{r}_n$, we require estimation of $S(\tau)$. In this section, we briefly describe asymptotic properties when we fit a parametric mixture cure model for $S(\tau)$ via maximum likelihood. The asymptotic properties of $\hat{r}_n$ follow directly from the well-known properties of maximum likelihood estimation. We discuss sensitivity to parametric model misspecification in Section 2.4.

We accrue each subject $i = 1, \ldots, n$ at an accrual time $A_i \sim \text{Unif}(0, a)$, and we follow these subjects until the administrative censoring time τ. We record an indicator for censoring (Δ_i) along with an observed time as the minimum of either their elapsed study time (T_i) or the time from accrual until administrative censoring time $(\tau - A_i)$. In other words, we observe a sample of n independent and identically distributed pairs $\{(Y_i, \Delta_i) : i = 1, \ldots, n\}$ with $Y_i =$

$\min(T_i, \boldsymbol{\tau} - A_i)$ and $\Delta_i = \mathbf{1}\{T_i \leq \boldsymbol{\tau} - A_i\}$. Define $F(t; \boldsymbol{\pi}, \Theta) = Pr(T \leq t)$ with derivative $f(t; \boldsymbol{\pi}, \Theta)$ based on the parametric model for the event times.

Because of the assumed independence between the event times and accrual, the log-likelihood is proportional to

$$l_n(\pi, \theta; Y, \Delta) \propto \sum_{i=1}^{n} \Delta_i \log f(Y_i; \pi, \theta) + (1 - \Delta_i)\log(1 - F(Y_i; \pi, \theta)) .$$

We first estimate the model parameters

$$\eta_n = (\pi_n, \theta_n) = \underset{\pi \in [0,1] \times \theta \in \Theta}{\arg\max} \; l_n(\pi, \theta; Y, \Delta) .$$

Now, define

$$\ell\,(\eta; Y, \Delta) = \Delta \log f(Y; \pi, \theta) + (1 - \Delta) \log(1 - F(Y; \pi, \theta))$$

with a second derivative ($(p + 1) \times (p + 1)$ Hessian matrix) $\ddot{\ell} = \left[\frac{\partial^2}{\partial\eta_i \partial\eta_j} \ell\,(\eta; Y, \Delta) \right]$ with $i, j = 1, \ldots, p + 1$.

Then we have, with S_C and f_C the survival and probability density functions of $C = \boldsymbol{\tau} - A$,

$$\mathbb{E}[\,\ddot{\ell}\,(\eta; Y, \Delta)] = \int_0^{\tau} \ddot{\ell}\,(\eta; Y, \Delta = 1) S_C(y) g(y)\, dy + \int_{\tau - a}^{\tau} \ddot{\ell}\,(\eta; Y, \Delta = 0)(1 - F(y)) f_C(y)\, dy . \tag{5}$$

Based on this setup, we can apply the standard theory of maximum likelihood estimation and the Delta method to establish that $\hat{r}_n$ is consistent for $r = \frac{S_{uc}(\tau)}{S(\tau)}$ and asymptotically normal with variance given by $D\mathscr{I}^{-1}D^T$, where $\mathscr{I} \equiv \mathscr{I}(\eta) = -\mathbb{E}[\,\ddot{\ell}\,(\eta; Y, \Delta)]$ and $D \equiv D(\eta_0) = \left[\frac{\partial}{\partial\eta_i} r(\eta) \Big|_{\eta = \eta_0} \right]$ with $i = 1, \ldots, p + 1$.[11] See Appendix C for technical details.

2.4 | Addressing sensitivity to model misspecification: RECeUS-AIC

Parametric models can be sensitive to model misspecification. As a result, it is common practice in this area to perform model selection when employing parametric models. Researchers can visually inspect various model fits by comparing the fitted survival curve against the Kaplan-Meier estimate, and it has also been suggested to study the profile likelihood function of the cure fraction parameter.[10]

We suggest addressing sensitivity to model misspecification in RECeUS by employing AIC.[12] First, select among a class of models with AIC, then use the RECeUS method with the best-fitting model. In this article, we select from a class of models that includes the noncure and mixture-cure versions of the Exponential, Weibull, Gamma, and Log-Logistic models. Choosing any noncure model immediately leads to a conclusion that a cure model is not appropriate. This class reflects an array of commonly-used models that can capture a wide variety of behavior in the data.

3 | COMPARING RECEUS-AIC WITH EXISTING METHODS VIA SIMULATION

3.1 | Existing methods and data generation

In this section, we evaluate the claim from Section 2.1 that the RECeUS-AIC method adequately assesses both (1) testing for the presence of immunes and (2) testing for sufficient follow-up.

We first compare RECeUS-AIC against existing methods for testing for the presence of immunes when the true cure fraction $\pi = 0$, and then, with $\pi > 0$, we compare against methods developed for testing sufficient follow-up. We provide a table briefly summarizing all of the available methods, Table 1. We include the formulas for each statistic in Appendix A. In short, $\hat{p}_n$ is the Kaplan-Meier estimate evaluated at the last observation time, d_n is a traditional deviance difference statistic for the null hypothesis of no cure fraction, and each of the methods for sufficient follow-up quantify the difference between the largest observation and the largest event time. We do not study the $\hat{\alpha}_n$ test because of the improvements to the procedure by the q_n and $\tilde{\alpha}_n$ tests.

For the q_n test, we choose $B = 6$ (corresponding to follow-up of 0.2% uncured remaining from their setup) to most closely match our use of 0.1% uncured remaining as indicative of sufficient follow-up. We use the Kaplan-Meier estimate of $\hat{p}_n$ rounded to the first significant digit and round the sample size down to the nearest available level in order to use the published tables. We use the 5% significance level for all existing methods.

For the RECeUS-AIC procedure, we follow the procedure as outlined in Section 2.1 and include AIC for model selection as indicated in Section 2.4. We present results employing a threshold $\hat{r}_n < 0.05$ and screening with $\hat{\pi}_n > 0.025$. If either criterion is not satisfied, the procedure indicates the data are not adequate for fitting a cure model - due to insufficient follow-up and/or very small or no cure fraction. Larger thresholds for $\hat{r}_n$ or smaller thresholds for $\hat{\pi}_n$ lead to more frequently concluding a cure model is appropriate - both when this is a correct and an incorrect decision. In the Appendix D.2, we study the impact of varying these thresholds.

We present the properties of the RECeUS-AIC procedure in sample sizes of 100, 250, 500, and 1000 to evaluate the procedure in real-world sample sizes against existing methods.

We generate data from Exponential, Gamma, Weibull and Log-Logistic mixture cure distributions with varying cure fractions (ranging from 0 to 0.8). We additionally vary the administrative censoring time (τ) to represent (approximately) the 75th, 90th, 95th, 99th, and 99.9th percentiles of the uncured distribution. Using the notation of Section 2.2, this corresponds to using $\pi \in [0, 0.8]$ and $u \in \{0.25, 0.1, 0.05, 0.01, 0.001\}$.

To better represent clinical trial analyses occurring at prespecified times, follow-up times corresponding to the percentiles are rounded to the nearest quarter to generate the data. For example, for the Weibull(2, 1) distribution, we use administrative censoring times of 1.25, 1.5, 1.75, 2.25, and 2.75.

For a given distribution, we fix the accrual end at $a = \tau_{0.75}/2$ (half of the 75th percentile of the given uncured distribution) and vary τ as we indicate above. With a fixed a, increasing τ indicates longer study follow-up.

In each setting (distribution with cure fraction π and follow-up corresponding to u), we compute the Monte Carlo estimate $\gamma_\pi(u)^*$ for a given method using 10 000 simulations. We provide the results from the Weibull(2, 1) mixture cure distribution, but patterns are similar across distributions. (One exception is the behavior of q_n with long follow-up and nonzero cure fraction generated by the Exponential and Log-Logistic mixture cure distributions, which we elaborate on below.) In the Appendix D.1, we provide results using an Exponential(1) mixture cure distribution that aim to replicate the data generation process used in Maller & Zhou (1996).

3.2 | Testing for the presence of immunes

Recall that we desire $\gamma_0(u)$ close to 0 for all u, a low probability of incorrectly claiming a cure model is appropriate when a cure fraction does not exist to generate the data. We illustrate the properties of the methods under this setting in Figure 2.

The rate at which the RECeUS-AIC procedure incorrectly claims a cure model is appropriate when $\pi = 0$ decreases with both sample size and u. It has $\gamma_0(u)^*$ as high as 12.5% with $n = 100$ and $u = 0.25$, but this decreases to 1.1% or less by $n = 1000$ for all u.

The properties of $\hat{p}_n$ seem to improve with decreasing u, but they do not generally improve with increasing sample size. The rates of concluding cure model appropriateness can be as small as 0.5% with long follow-up, but it can be as large as 99.2% with shorter follow-up.

3.3 | Testing for sufficient follow-up

In this article, we use $u = 0.1\%$ as follow-up sufficient for practical cure modeling. This translates to desiring $\gamma_\pi(0.001)$ close to 1 for $\pi > 0$, or a high probability of correctly claiming a cure model is appropriate when 0.1% uncured subjects are remaining and the true cure fraction is nonzero. In Figure 3, we present $\gamma_\pi(0.001)^*$ with varying $\pi \in [0.1, 0.8]$ for each procedure.

As sample size increases, the RECeUS-AIC procedure more frequently reaches the correct conclusion that a cure model is appropriate with a nonzero cure fraction and 0.1% uncured remaining ($u = 0.001$). The RECeUS-AIC procedure has $\gamma_\pi(0.001)^* > 81\%$ at $n = 100$ and reaches $\gamma_\pi(0.001)^* > 99\%$ by $n = 1000$ for all $\pi > 0$. In small samples, we observe that rates are highest with intermediate cure fractions 30%−60%. Smaller cure fractions ($\pi \leq$ 10%) require more information to conclude cure model appropriateness, while larger cure fractions ($\pi \geq$ 80%) necessarily suffer from higher rates of censoring when there is a fixed amount of follow-up. Both are addressed by increasing either sample size or follow-up time.

The $\tilde{\alpha}_n$ test has lower $\gamma_\pi(0.001)^*$, with a maximum of 31.1%, and these rates do not improve with sample size. The q_n test has rates of $\gamma_\pi(0.001)^*$ of 2% or less for all sample sizes in this Weibull(2, 1) setting. However, for the Log-Logistic(2, 1) mixture distribution (results not presented), we find that q_n can achieve larger rates of $\gamma_\pi(0.001)^*$. In the Log-Logistic(2, 1) setting, we observe rates in the range of 12.7%–80.0%, but these continue to be uniformly smaller than the corresponding rates of the RECeUS-AIC. The q_n test also has improved rates in the Exponential(1) setting, which we discuss in Section 5 and Appendix D.1. Sensitivity to the data generating distribution arises from the difference between the Exponential(1) distribution's tail used to generate the published critical values compared to other distributions.

3.4 | Cure model appropriateness with intermediate follow-up and nonzero cure fraction using RECeUS-AIC

Interpreting the results for 25% down to 1% uncured remaining ($u \in [0.01, 0.25]$) is less straightforward. A cure fraction does exist, but these follow-up times may not represent adequate follow-up time to identify a cure model in practice, and, to our knowledge, no definitive resource exists to establish properties of cure models in the setting with intermediate follow-up. In this section, we report the properties of RECeUS-AIC in this setting.

We tabulate $\gamma_\pi(u)^*$ in Table 2 for RECeUS-AIC across the varying $\pi \in [0, 0.8]$ and $u \in \{0.25, 0.1, 0.05, 0.01, 0.001\}$ when using the thresholds $\hat{\pi}_n < 0.025$ and $\hat{r}_n > 0.05$. In addition to $\gamma_\pi(u)^*$, the table includes the "True Ratio" r for additional context: if r is small, then we expect $\gamma_\pi(u)$ to be large and the opposite if r is large.

With cure fractions of 30% or larger, the probability of concluding a cure model is appropriate increases with sample size when 1% uncured remain ($u = 0.01$). However, with a cure fraction of 10%, this does not hold: this reflects that longer follow-up may be needed to identify smaller cure fractions.

Across all settings with nonzero cure fraction, $\gamma_\pi(u)^*$ decreases to 0 with sample size when 5% or more of the uncured remain ($u \geq 0.05$). This follows from the use of 0.05 as the threshold for $\hat{r}_n$, which implicitly expresses that a cure model is appropriate with at most 5% uncured remaining and even less with smaller cure fractions.

4 | REVISITING THE MOTIVATING DATA EXAMPLE

SWOG Cancer Research Network, a US NCI-funded clinical trials cooperative group, provides a rare setting for studying cure models because SWOG continues to follow patients after the primary analysis. Therefore, we can directly evaluate whether tests of sufficient follow-up are concordant with results after additional years of follow-up with data examples.

The data in Figure 1 represent trial data from 2014. To assess cure model appropriateness using the current paradigm, we first test for the presence of immunes. Using the 2014 data, we calculate $\hat{p}_n = 0.574$ and a censoring rate of 71.4%. We compare 0.574 with the published critical value 0.8776 based on $n = 250$ (rounding down from $n = 276$ in the data), $\mu =$

1 (rounding from the approximation $\mu \approx (1/0.714) - 1$) given in the text[6]) and the 5th percentile. We see that $\hat{p}_n$ concludes that evidence exists for the presence of immunes.

We can then proceed to test for sufficient follow-up. The q_n and $\tilde{\alpha}_n$ tests do not conclude that evidence exists for sufficient follow-up. We compare $q_n = 0.029$ with the published critical value 0.5050 based on $n = 200$ (rounding down from $n = 276$ in the data), $B = 6$ and the 95th percentile. We calculate $\tilde{\alpha}_n = 0.134$.

Alternatively, we can employ RECeUS-AIC. For the class given in Section 2.4, AIC selects the noncure Log-Logistic model, and we immediately conclude a cure model is not appropriate for these data. Below, we evaluate whether these decisions are consistent with results using longer follow-up.

Figure 4 represents a scenario with 3 additional years of follow-up for the trial S1117, and the data do not seem sufficient to support cure model analysis: the tail of the survival probability estimate continues to decrease after 2014. Thus, summaries from 2014 data that require accurate estimation of the right tail would be biased.

We see that all methods agree with that conclusion of insufficient follow-up in 2014.

We can also assess whether a cure model may be appropriate with the extended follow-up data. The $\hat{p}_n$ test concludes that there are immunes comparing $\hat{p}_n = 0.817$ against the critical value 0.9772 with $n = 250$, $\mu = 3$ (from 27.5% censoring and the approximation $\mu \approx (1/0.275) - 1) \approx 3$) and the 5th percentile.

If we proceed to test for sufficient follow-up, we see that both the q_n and $\tilde{\alpha}_n$ tests conclude that follow-up is not sufficient. We compare $q_n = 0.004$ against 0.6200 with $n = 200$, $B = 6$ and the 95th percentile, and we calculate $\tilde{\alpha}_n = 0.367$.

With the extended data, the RECeUS-AIC method selects the mixture-cure Log-Logistic model with $\hat{\pi}_n = 0.11 > 0.025$ and $\hat{r}_n = 0.55 > 0.05$, so it does not conclude a cure model is appropriate.

The existing paradigm and the RECeUS approach both reach the conclusion that a cure model would be inappropriate for analyzing the extended follow-up data.

As a second example, we examine trial S0106 (clinicialtrials.gov identifier: NCT00085709[13]). The trial randomized acute myeloid leukemia (AML) patients to either standard therapy or the combination of standard therapy and the drug mylotarg. As with S1117, the promises of allogeneic transplant inducing cure in AML patients in the past prompted investigators to explore cure modeling in S0106.

First, we assess whether we can validly assume the presence of immunes using the data released in 2011. The $\hat{p}_n$ test claims that evidence exists that immunes are present. We calculate $\hat{p}_n = 0.546$ and a proportion censored of 58.0%, and compare to the published critical value 0.9386, corresponding to $n = 500$ (rounding down from $n = 600$ in the data), 5th percentile and $\mu = 1$ (using the approximation $\mu \approx (1/0.58) - 1) \approx 1$).

In tests for sufficient follow-up, we see the $\tilde{\alpha}_n$ and q_n tests do not conclude sufficient follow-up. We compute $q_n = 0.010$ and compare this with the published critical value 0.1500, corresponding to $n = 600$, $p = 0.6$ (rounding from $\hat{p}_n = 0.546$), $B = 6$ and the 95th percentile. And we have $\tilde{\alpha}_n = 0.368$.

For the RECeUS-AIC method, we select a Weibull mixture-cure model, and we calculate $\hat{\pi}_n = 0.426 > 0.025$ and then $\hat{r}_n = 0.032$. With $\hat{r}_n < 0.05$, we have evidence to conclude a cure model is appropriate.

For this trial, SWOG also has additional follow-up in Figure 5. We see that even with extended follow-up, a plateau exists at a similar level in the tail of the survival function estimate. This supports the conclusion that follow-up at the trial’s data release in 2011 does seem sufficient and a cure model analysis would be appropriate. Further, we see that the fitted curve for the best-fitting mixture-cure model, the Weibull mixture-cure model, seems to fit the extended follow-up well.

In this situation, the extended follow-up confirms the conclusions reached by RECeUS calculated based on 2011 data but not the conclusions of the q_n and $\tilde{\alpha}_n$ approaches.

These data examples illustrate other advantages of using the RECeUS approach in assessing sufficient follow-up over other methods: the approach agrees with the results after additional years of follow-up in both examples, while the other methods may provide contradictory results.

5 | DISCUSSION

Existing tests for assessing sufficient follow-up were developed with the premise that a cure model is inappropriate if any uncured subjects remain at the end of the study. Because of this, existing tests are not calibrated for the real-world setting in which some proportion of uncured subjects do exist at the end of the study: as discussed in earlier work,[6,9] the $\hat{\alpha}_n$ test frequently concludes sufficient follow-up for a cure model when inappropriate, and we see in our results that the $\tilde{\alpha}_n$ and q_n tests rarely conclude sufficient follow-up even with a nonzero cure fraction and long follow-up.

From Yu et al[10] and simulations summarized in the Appendix B, evidence exists to motivate a different approach. We suggest quantifying cure model appropriateness by studying the proportion of uncured subjects remaining in the study (or a standardized version, r), and we illustrate that the proposed procedure RECeUS assesses cure model appropriateness by simultaneously evaluating (1) the presence of immunes and (2) the sufficiency of follow-up.

We demonstrate that, when implemented via maximum likelihood estimation, the estimator $\hat{r}_n$ is consistent and asymptotically normal under standard regularity conditions. In finite samples, the RECeUS procedure using AIC model selection, RECeUS-AIC, frequently comes to the correct conclusion of cure model appropriateness with nonzero cure fraction and long follow-up. When the cure fraction is zero, the method has a low rate for incorrectly concluding a cure model is appropriate.

The existing methods for assessing cure model appropriateness were developed to test for either the presence of immunes or sufficient follow-up. The results presented here indicate that the RECeUS-AIC procedure outperforms these methods within their respective task. In their textbook, Maller & Zhou suggest that test statistics that use more information in the right tail may be preferable for evaluating sufficient follow-up.[6] This seems evident in these results, where we see an advantage to the parametric structure of RECeUS-AIC compared to existing methods, which are fully nonparametric.

However, we acknowledge several limitations to our simulation results. To employ the $\hat{p}_n$ and q_n tests, we had to make choices regarding rounding of censoring rates, cure fractions and sample sizes and make a choice of adequate follow-up time to use the published critical value tables. In addition, we did not assess the d_n test for the presence of immunes for this article because the subjectivity involved in evaluating the goodness of fit is challenging to implement within a simulation study. Next, for both $\hat{p}_n$ and q_n, we can identify specific methodological aspects that may limit these simulation results. For $\hat{p}_n$, we observe particular sensitivity to shorter follow-up times. The authors for $\hat{p}_n$ noted that results of $\hat{p}_n$ may need to be supplemented by evaluation of sufficient follow-up, so it may be a limitation of this study to evaluate $\hat{p}_n$ alone. In a similar vein, the authors recommend the decision of sufficient follow-up using q_n may benefit from an assessment of outliers. And, while q_n is constructed as a fully nonparametric statistic, it relies on simulated critical value tables because of the challenge in obtaining exact (or even large-sample) critical values.[6] (This is also true for $\hat{p}_n$.) We find here that this existing approach to using q_n is sensitive to model specification: the available critical value tables are generated from an Exponential distribution, which is misspecified for the results in the main text, but we see improvement when assessing results derived from an Exponential distribution (Appendix D.1). An alternative approach to employ these tests could involve generating critical values reflective of the available data for each new dataset, and this suggestion or other future work may improve the properties of these methods.

When evaluating the properties of methods for sufficient follow-up via simulation, we choose follow-up corresponding to $u = 0.1\%$ uncured remaining, and this is consistent with guidance from Maller & Zhou (1996), where they suggest $u = 0.2\%$ ($B = 6$ in their notation) as reasonably sufficient. Using this threshold, we conclude that RECeUS-AIC has desirable high rates of concluding cure model appropriateness when $\pi > 0$ and $u = 0.1\%$, but we observe lower rates for $\tilde{\alpha}_n$ and q_n. However, to our knowledge, no resource exists to definitively claim that any specific level is an appropriate level in general. This is related to why the statistics $\tilde{\alpha}_n$ and q_n conclude a cure model is appropriate at rates lower than RECeUS-AIC with $u = 0.1\%$: by construction, these methods require longer follow-up to frequently conclude a cure model is appropriate. This underscores the need for continued research on the impact of varying follow-up time on cure model analysis, and results in our

Appendix B point to the possibility that a single universal threshold for sufficient follow-up may not be appropriate.

Unlike existing methods, the proposed method is not a formal hypothesis testing procedure, and it cannot make guarantees for error rates that accompany methods developed for formal hypothesis testing. However, RECeUS-AIC when using a threshold for $\hat{\pi}_n$ of 2.5% and a threshold for $\hat{r}_n$ of 5% exhibits desirable behavior across the range of settings under study, which we believe reflect a broad range of real-world biomedical settings.

In this article, we propose the use of RECeUS with AIC as a tool to minimize sensitivity to model misspecification. This implicitly requires that the class of models used for selection be sufficiently rich to effectively model the data. In our experience, a class including Exponential, Weibull, Gamma and Log-Logistic models suffices for a variety of data applications, especially in biomedical research, and we observe strong performance when using RECeUS-AIC with this class. However, researchers for specific applications may desire additional flexibility than provided by this class. One suggestion is to include more flexible parametric models such as the Generalized Gamma or piecewise Exponential models within the class for AIC model selection. In future research, we will also study an alternative approach by estimating $S(\tau)$ with semiparametric methods, as the RECeUS method readily permits flexible model specification for $S(\tau)$.

Subject area knowledge should guide the decision of cure model appropriateness. Researchers will frequently anticipate a clinically meaningful cure fraction and only wish to apply a cure model if evidence exists to suggest this level of cure has been achieved. In addition, different areas may accept different fractions of uncured remaining when modeling cure. For example, researchers expecting events to occur slowly over a long time may accept modeling a cure fraction with larger values of u than those expecting all events to occur rapidly. The RECeUS method can accommodate both of these aspects by modifying the thresholds for $\hat{\pi}_n$ and $\hat{r}_n$. We provide additional details for modifying the thresholds in the Appendix D.2. (While we do not study $u < 0.1\%$ explicitly, the rate of concluding cure model appropriateness for RECeUS-AIC is high based on the results for $u = 0.1\%$.) For all choices of u and a cure fraction $\pi = 0$, RECeUS-AIC protects against the incorrect conclusion of a cure model being appropriate. The rate of concluding cure model appropriateness with a nonzero cure fraction at choices of u depends on the threshold chosen for $\hat{r}_n$ (we demonstrate high rates of concluding cure model appropriateness when using the criterion $\hat{r}_n < 0.05$ for u as large as 0.1%).

We believe that this novel method offers a simpler procedure for assessing cure model appropriateness and outperforms existing methods. We hope that the practicality and efficacy leads to the widespread use of RECeUS in scientific practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the helpful feedback from referees and funding from grant number T32CA09168 from the National Cancer Institute (NCI), NIH.

Funding information

National Cancer Institute, Grant/Award Number: T32CA09168

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

APPENDIX A.: FORMULAS FOR EXISTING METHODS

A.1 Testing for the presence of immunes

Let $\widetilde{S}_{KM,n}(t)$ be the Kaplan-Meier estimate of the survival function at time t and $t_{(n)}$ be the largest observation in the available data with n observations. Then $\hat{p}_n = \widetilde{S}_{KM,n}\left(t_{(n)}\right)$[6]

Let $l_n(\theta)$ be the log-likelihood for a particular (cure) model at θ with sample size n, $\tilde{\theta}_n$ be the maximizer of the log-likelihood and $\tilde{\theta}_{H_0}$ be the maximizer under the null hypothesis (no cure fraction, $\pi = 0$). Then d_n is the traditional deviance difference statistic $d_n = 2\left(l_n\left(\tilde{\theta}_n\right) - l_n\left(\tilde{\theta}_{H_0}\right)\right)$.[6]

A.2 Testing for sufficient follow-up

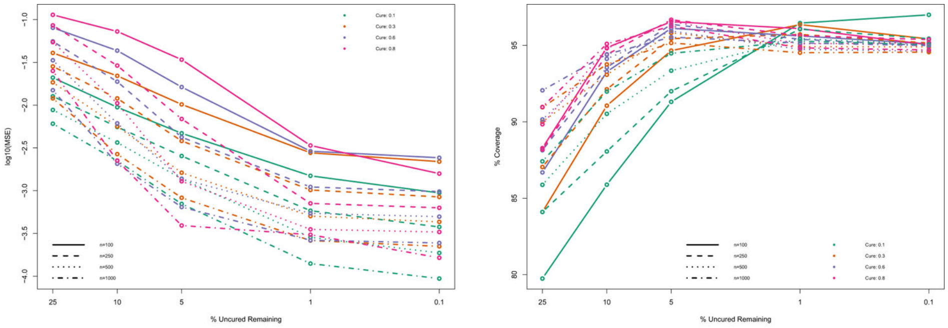
All of the existing methods $\hat{\alpha}_n$, $\tilde{\alpha}_n$, and q_n compare the largest event time and the largest observation in the data of n observations and have similar construction. We continue to let $t_{(n)}$ be the largest observation, and let $t^*_{(n)}$ be the largest event time.

Now let $N_n^{(1)}$ be the number of event times in the interval $I^{(1)} = \left(2t^*_{(n)} - t_{(n)}, t^*_{(n)}\right]$. Then $q_n = \frac{N_n^{(1)}}{n}$[6] and $\hat{\alpha}_n = \left(1 - \frac{N_n^{(1)}}{n}\right)^n$.[8]

The statistic $\tilde{\alpha}_n$ uses a different interval. Let $\widehat{w} = \left(t_{(n)} - t^*_{(n)}\right)/t_{(n)}$ and $\hat{\tau}_g = \widehat{w}t^*_{(n)} + (1 - \widehat{w})t_{(n)}$, and $N_n^{(2)}$ is the number of event times in interval $I^{(2)} = \left[\hat{\tau}_g t^*_{(n)}/t_{(n)}, t^*_{(n)}\right]$. Then $\tilde{\alpha}_n = \left(1 - \frac{N_n^{(2)}}{n}\right)^n$.[9]

APPENDIX B.: ESTIMATION AND INFERENCE OF MIXTURE CURE MODEL PARAMETERS IMPROVE WITH LONGER FOLLOW-UP

This section expands on the results of Yu et al[10] and assesses the hypothesis that estimation and inference need not be limited to the setting in which we believe no uncured subjects remain at the end of the study, or $\tau_{F_0} < \tau_G$. We perform simulations to study the mean-squared error and coverage of 95% confidence intervals when fitting mixture cure models as follow-up time increases. (Note that here emphasis lies in the patterns across follow-up times.) These results are drawn from the simulation study described in Section 3.



(A) The mean-squared error for the cure fraction estimate in Weibull(2,1) mixture cure distributions decreases with longer follow-up times and with sample size.

(B) The coverage of confidence intervals for the cure fraction estimate in Weibull(2,1) mixture cure distributions approach nominal levels with longer follow-up times and with sample size.

FIGURE B1.

Results to assess the properties of cure fraction estimation for Weibull(2, 1) Mixture cure distributions with longer follow-up time

We fit correctly-specified mixture cure models via maximum-likelihood estimation and compute the mean-squared error and coverage of Wald-based 95% confidence intervals for each model parameter estimate.

Researchers often focus on estimation and inference on the cure proportion, so we provide results of this parameter from a Weibull(2, 1) mixture cure generating distribution in this section. Figures B1A, B demonstrate clear improvements in estimation and inference over time and with increased sample size. (We do notice that at small sample sizes of $n = 100$, confidence intervals may be larger than needed with conservative coverage.)

We emphasize that a similar trend of improvement over time remains across other distribution settings and across other model parameters. The shape and magnitude of improvement may differ, but mean-squared error decreases and confidence interval coverage approaches the nominal rate.

In addition to supporting the claim that researchers can compute statistics with nominal coverage and low mean-squared error without $\tau_{F_0} < \tau_G$ holding, figures such as these can also quantify the length of time most helpful in a trade-off between follow-up time and statistical validity (as measured by mean-squared error and coverage). For the Weibull model, it seems that mean-squared error and coverage improve dramatically from 25% uncured remaining to 5% uncured remaining. However, with fewer than 1% uncured remaining, additional follow-up seems to show diminishing gains. As such, we might reasonably consider 1% or lower uncured remaining to be strong evidence for sufficient follow-up, and we might be skeptical of the follow-up when many more than 5% of uncured remain in the study.

We believe that these results may support a shift for the discussion of “sufficient” follow-up time being a single number for a given study to being a range of times depending on a researcher’s decision on such a trade-off.

APPENDIX C.: TECHNICAL DETAILS FOR ASYMPTOTIC PROPERTIES

To derive the consistency and asymptotic normality of $\hat{r}_n$, we invoke van der Vaart[11] Theorems 5.41 and 5.42. (Similar theorems have also been published elsewhere.)

Theorem 5.41 establishes asymptotic normality of maximum likelihood estimators (in the view of maximum likelihood estimation as Z-estimation with the score equations) and 5.42 ensures consistent roots exist in this setting. Define $\psi_\eta(x) \equiv \dot{l}_n(\eta) = \frac{d}{\eta_i} l(\eta; y, \delta)$ and $m_\eta(x) \equiv l(\eta; y, \delta)$ (with $X = (Y, \Delta)$).

The regularity conditions needed for consistency and asymptotic normality of maximum likelihood estimation are as follows.

R1. The parameter of interest η_0 is a local maximum of $\mathbb{E} m_\eta(Y, \Delta)$ and $0 = \mathbb{E}\psi_{\eta_0}(Y, \Delta)$.

R2. The score equations $\psi_\eta(y, \delta)$ are twice continuously differentiable for every (y, δ).

R3. The second moment $\mathbb{E}\|\psi_{\eta_0}(Y, \Delta)\|^2 < \infty$.

R4. The matrix $\mathbb{E}\dot{\psi}_{\eta_0}(Y, \Delta)$ exists and is nonsingular.

R5. The second-order derivatives of the score equations are dominated by a fixed integrable function $\ddot{\psi}(y, \delta)$ for every η in a neighborhood of η_0.

R6. $S(t)$ is uniformly bounded away from 0 when $t \in [0, \tau]$.

By van der Vaart,[11] these conditions are sufficient in the setting of maximum likelihood estimation from common parametric families but are likely stricter than necessary. The Exponential, Gamma, Weibull and Log-Logistic families used with Uniform censoring in this report, as popular choices in applied survival analysis, clearly meet these conditions due to their smoothness. The final condition holds in real-world conditions with finite follow-up.

In addition, for consistency and asymptotic normality of $\hat{r}_n$, we do require $\eta \mapsto r(\eta)$ be continuous and differentiable at η_0. But the listed conditions are frequently sufficient for this because $r(\eta)$ is a smooth transformation of η in these parametric models.

These theorems do not preclude the existence of other, inconsistent roots. We have not experienced this in practice, but if needed to address multiple roots, van der Vaart[11] suggests considering a preliminary consistent estimate and then choose the closest root. In particular, we recommend the root closest to $\tilde{\pi}_n = \tilde{S}_{KM,n}(\tau)$ (where $\tilde{S}_{KM,n}(\tau)$ is the estimate of survival probability at the end of the study by Kaplan-Meier) because this is a simple, consistent nonparametric estimator of π_0.[6]

APPENDIX D.: EXTENDED SIMULATION RESULTS

D.1 Simulation results with an Exponential(1) mixture cure distribution

The main text Section 3 for this article presents results using a Weibull(2, 1) mixture cure distribution. In this section, we provide results that replicate the data generation process used by Maller & Zhou (1996) in order to evaluate the existing methods when they are most correctly specified. The authors simulated event times from an Exponential(1) distribution mixture cure distribution and generated censoring times from a Uniform[0, B] distribution with $B \in \{2, 4, 6, 8, 10\}$.

We verify the validity of our replication by comparing our simulated critical values for $\hat{p}_n$ and q_n against the published critical values.

For q_n, our simulated critical values generally agree to the second significant digit. (For a small number of values in some settings, there was a difference of +∕−0.01 at the second significant digit.) For $\hat{p}_n$, our values agree to the second significant digit for many settings, but, for the setting with $B = 2$, some values differ even to the first significant digit. Our results have systematically smaller critical values compared to the published tables when $B = 2$. However, we continue to use the published critical values for presenting properties of the $\hat{p}_n$ test, so this means that our results for $\hat{p}_n$ with short follow-up, $B = 2$, will have large rates of concluding a cure model is appropriate when $\pi = 0$.

The difference in setup changes the fraction of uncured remaining (u) generating these results. Instead of decreasing from 25% to 0.1%, we have 13.5% uncured remaining down to 0.005%. Analogously to the main text, we present results of the existing methods at different u when $\pi = 0$ (Figure D1) and at different π with 0.005% uncured remaining. (We choose 0.005% in order to assess the methods in the longest available follow-up.)

We see that the conclusions of the main text continue to hold in testing for the presence of immunes, but we see improvement in all of the methods in the Exponential(1) setup. For testing sufficient follow-up, we see all methods improve, and q_n improves dramatically: in the Weibull case, rates of $\gamma_\pi(0.001)$* are uniformly small, but with the Exponential setup, we see much larger rates and improvement with sample size.

D.2 Assessing different thresholds for use with RECeUS

We propose the RECeUS procedure in Section 2.1 using 0.025 as a threshold for $\hat{\pi}_n$ and 0.05 as a threshold for $\hat{r}_n$. The choice of threshold impacts the rates of concluding cure model appropriateness and represents a tradeoff between low $\gamma_0(u)$ and high $\gamma_\pi(0.001)$ (for $\pi > 0$). Our proposed combination seeks to provide a balance between these, and in this section, we describe the impact of varying these thresholds. We use the Weibull(2, 1) mixture cure distribution setup as given in Section 3.

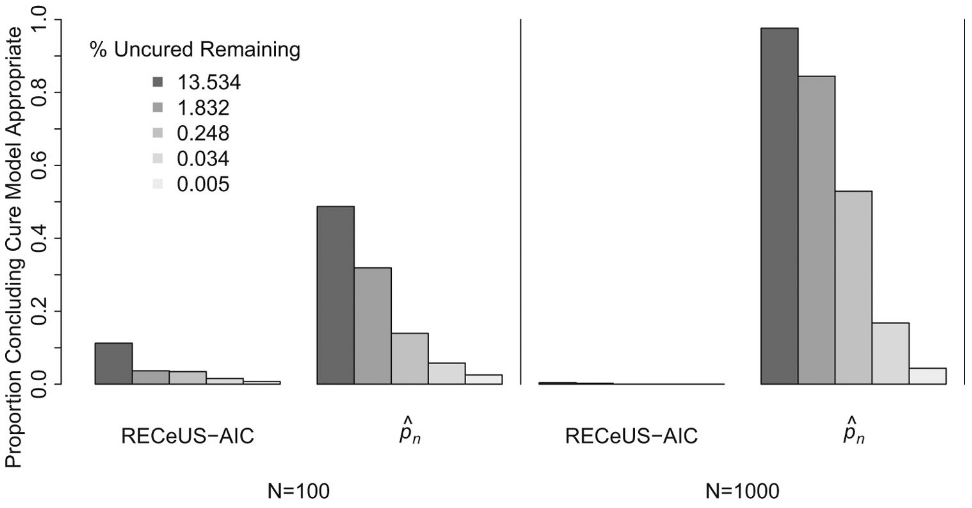


FIGURE D1.

The proportion of 10 000 simulations that (incorrectly) conclude a cure model is appropriate under settings with no cure fraction ($\pi = 0$) and varying percent uncured remaining, $\gamma_0(u)$*, when using an Exponential(1) mixture cure distribution

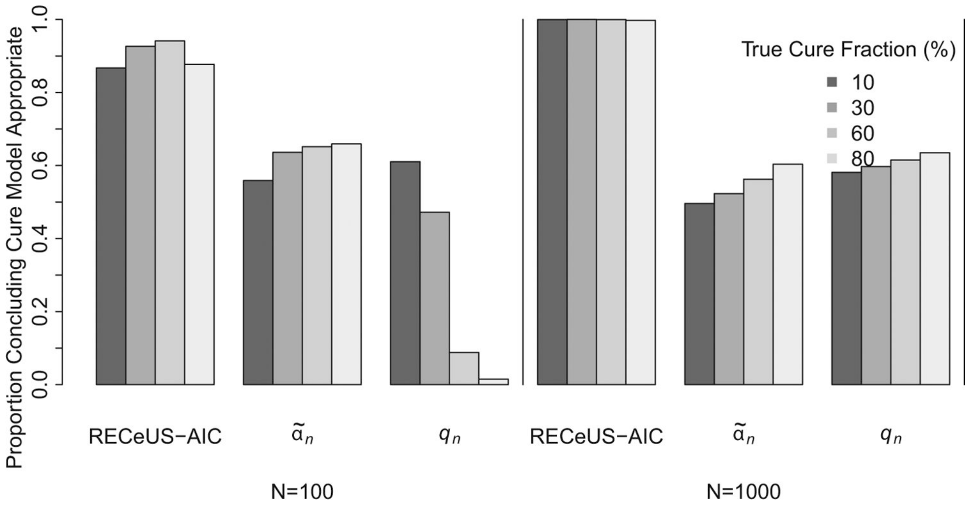


FIGURE D2.

The proportion of 10 000 simulations that (correctly) conclude a cure model is appropriate under settings with a nonzero cure fraction and 0.005% uncured remaining, $\gamma_\pi(0.001)$*, when using an Exponential(1) mixture cure distribution

Broadly, increasing the threshold of $\hat{r}_n$ or decreasing the threshold of $\hat{\pi}_n$ both serve to increase the rates of concluding cure model appropriateness, both when this is desirable and when it is not. We see different specific patterns when we vary the $\hat{\pi}_n$ threshold compared to the $\hat{r}_n$ threshold. We present these specific patterns in Figures D3–D6. We fix either the $\hat{\pi}_n$ or $\hat{r}_n$ threshold and vary the other. For presentation purposes, we also fix the sample size at $n = 100$, but patterns are similar across sample sizes.

First, we fix the $\hat{r}_n$ threshold at 0.05 (patterns are similar within other threshold levels) and we vary the $\hat{\pi}_n$ threshold in {0, 0.025, 0.1}. When the true cure fraction $\boldsymbol{\pi} = 0$, we see minimal change in changing the $\hat{\pi}_n$ threshold with short follow-up, but, for longer follow-up, we find using a larger threshold for $\hat{\pi}_n$ leads to the desirable property of a small frequency of concluding a cure model is appropriate.

When the cure fraction is nonzero, changing the $\hat{\pi}_n$ threshold only impacts the properties when the threshold approaches the true cure fraction. In Figure D4, we see that if $\boldsymbol{\pi}$ = 10%, and we use a threshold of 0.10, then rates of concluding cure model appropriateness undesirably decrease.

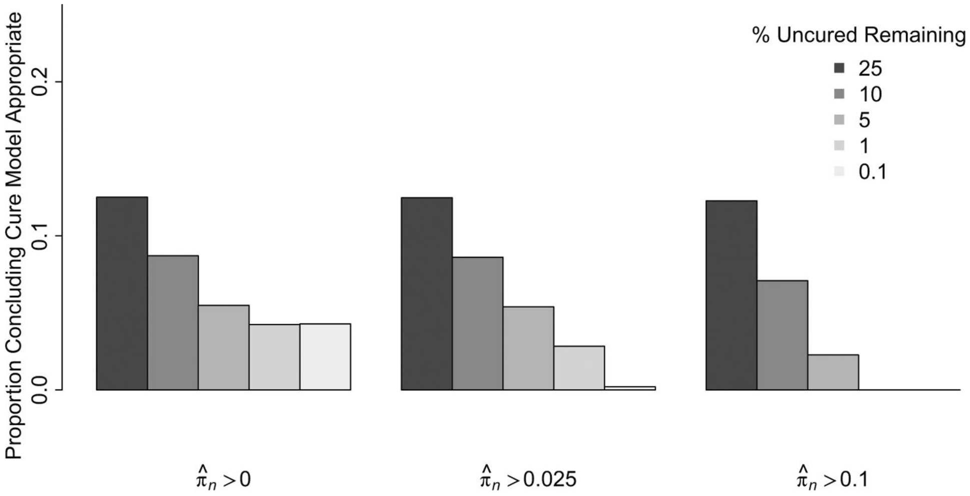


FIGURE D3.

The proportion of 10 000 simulations of RECeUS-AIC that (incorrectly) conclude a cure model is appropriate under settings with no cure fraction ($\boldsymbol{\pi} = 0$) and varying percent uncured remaining, $\gamma_0(u)^*$, across different choices of $\hat{\pi}_n$ thresholds with $n = 100$ and a threshold for $\hat{r}_n$ of 0.05

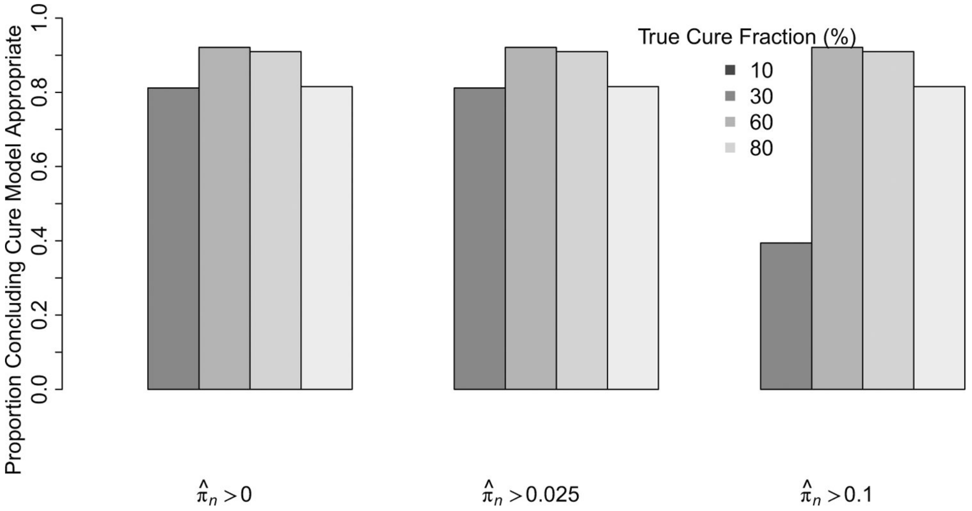


FIGURE D4.

The proportion of 10 000 simulations of RECeUS-AIC that (correctly) conclude a cure model is appropriate under settings with a nonzero cure fraction and 0.1% uncured remaining, $\gamma_\pi(0.001)$*, across different choices of $\hat{\pi}_n$ thresholds with $n = 100$ and a threshold for $\hat{r}_n$ of 0.05

Next, we fix the $\hat{\pi}_n$ threshold at 0.025 and vary the $\hat{r}_n$ threshold in {0.01, 0.05, 0.10}. We only see modest changes by varying this threshold in the rate of concluding cure model appropriateness for $\pi = 0$.

With long follow-up and true cure fraction $\pi > 0$, we observe a large benefit for choosing a threshold of 0.05 over a threshold of 0.01. We observe only a small relative benefit by further increasing from 0.05 to 0.10.

We conclude that the choice of $\hat{\pi}_n$ threshold most impacts the method's properties when the threshold approaches the true cure fraction. We suggest that authors choose a suitably large threshold but one that is smaller than an expected cure fraction. The use of a threshold for $\hat{\pi}_n$ of 0 (ie, using $\hat{r}_n$ alone) does not negatively impact the properties with shorter follow-up, but there are costs (relative to using a nonzero threshold) when a study has longer follow-up with no appreciable gains.

Using a threshold of $\hat{r}_n$ of 0.01 maintains a low rate of concluding cure model appropriateness when $\pi = 0$, but at the cost of low rates when $\pi > 0$. The choice between 0.05 and 0.10 is less impactful.

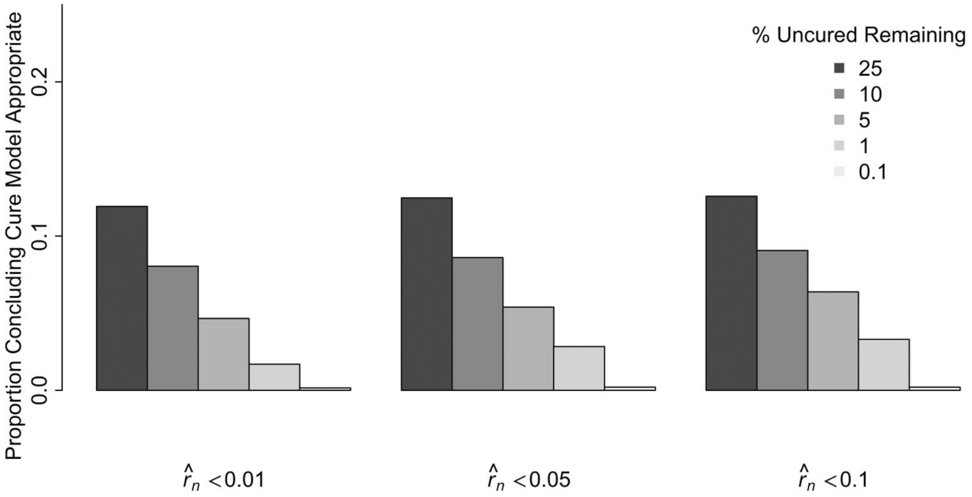


FIGURE D5.

The proportion of 10 000 simulations of RECeUS-AIC that (incorrectly) conclude a cure model is appropriate under settings with no cure fraction ($\pi = 0$) and varying percent uncured remaining, $\gamma_0(u)$, across different choices of $\hat{r}_n$ thresholds with $n = 100$ and a threshold for $\hat{\pi}_n$ of 0.025

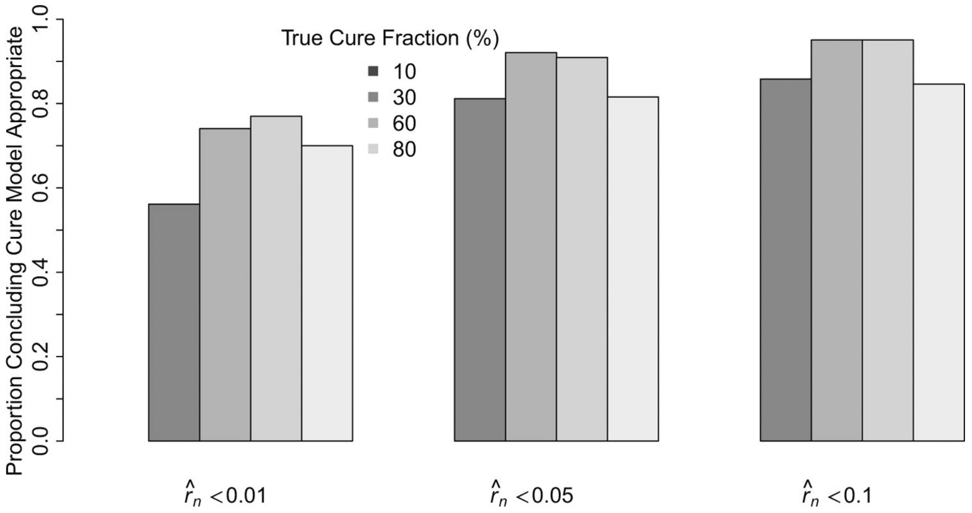


FIGURE D6.

The proportion of 10 000 simulations of RECeUS-AIC that (correctly) conclude a cure model is appropriate under settings with a nonzero cure fraction and 0.1% uncured remaining, $\gamma_\pi(0.001)$*, across different choices of $\hat{r}_n$ thresholds with $n = 100$ and a threshold for $\hat{\pi}_n$ of 0.025

In summary, our proposed method performs well within the wide range of settings under study using the combination of 0.025 for the $\hat{\pi}_n$ threshold and 0.05 as the $\hat{r}_n$ threshold. Researchers interested in a more specific application within this range or settings outside the scope of this article may benefit from updating this combination of thresholds.

D.3 Comparing RECeUS-AIC with correctly-specified RECeUS-Weibull

We can also compare RECeUS-AIC to RECeUS fit with the correctly-specified model. We include this comparison for the Weibull(2, 1) mixture-cure distribution in Table D1. As the results between RECeUS-AIC and RECeUS-Weibull differ, we can immediately conclude that AIC does not always correctly select the Weibull model under these settings. Correct model specification does have improved behavior, but the RECeUS-AIC procedure has acceptable properties and model selection reduces sensitivity to model misspecification (over choosing a single model a priori).

TABLE D1

Rates for concluding sufficient follow-up ($\gamma_{\pi}(u)$) in RECeUS-Weibull and RECeUS-AIC by percentage of uncured remaining, sample size and cure fraction in 10 000 simulations with a Weibull(2, 1) mixture cure distribution

		RECeUS-Weibull[a]					**RECeUS-AIC**[a]				
		% Uncured remaining					**% Uncured remaining**[b]				
Cure fraction	**Sample size**	**25%**	**10%**	**5%**	**1%**	**0.10%**	**25%**	**10%**	**5%**	**1%**	**0.10%**
0%	100	0.024	0.018	0.019	0.035	0.003	0.125	0.086	0.054	0.028	0.002
	250	0.001	0.000	0.000	0.003	0.000	0.067	0.041	0.027	0.009	0.000
	500	0.000	0.000	0.000	0.000	0.000	0.036	0.016	0.006	0.002	0.000
	1000	0.000	0.000	0.000	0.000	0.000	0.011	0.001	0.000	0.000	0.000
10%	100	0.042	0.043	0.081	0.462	0.962	0.130	0.102	0.104	0.327	0.811
	250	0.003	0.003	0.013	0.428	0.998	0.070	0.057	0.068	0.372	0.926
	500	0.000	0.000	0.001	0.383	1.000	0.049	0.032	0.036	0.386	0.969
	1000	0.000	0.000	0.000	0.331	1.000	0.020	0.007	0.007	0.355	0.998
30%	100	0.104	0.130	0.242	0.793	0.996	0.159	0.157	0.214	0.588	0.921
	250	0.013	0.029	0.115	0.889	1.000	0.089	0.085	0.147	0.698	0.985
	500	0.001	0.003	0.041	0.954	1.000	0.058	0.066	0.102	0.809	0.999
	1000	0.000	0.000	0.005	0.990	1.000	0.047	0.034	0.062	0.921	1.000
60%	100	0.243	0.304	0.453	0.882	0.997	0.197	0.243	0.328	0.655	0.909
	250	0.100	0.158	0.377	0.960	1.000	0.140	0.176	0.292	0.775	0.968
	500	0.026	0.074	0.302	0.991	1.000	0.089	0.119	0.255	0.857	0.992
	1000	0.002	0.015	0.215	0.999	1.000	0.072	0.083	0.224	0.924	0.999
80%	100	0.400	0.460	0.574	0.868	0.987	0.201	0.256	0.338	0.577	0.816
	250	0.238	0.328	0.525	0.944	1.000	0.199	0.263	0.382	0.743	0.955
	500	0.120	0.227	0.488	0.982	1.000	0.152	0.210	0.351	0.823	0.982
	1000	0.039	0.116	0.448	0.999	1.000	0.100	0.149	0.339	0.901	0.995

[a] The RECeUS-Weibull procedure employs the correctly-specified Weibull model for estimation.

[b] In this setting, the exact follow-up times generating the data were 1.25, 1.5, 1.75, 2.25, and 2.75. These are the follow-up times corresponding to the percentiles rounded to the nearest quarter in order to more closely represent clinical trial data.

REFERENCES

1. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J R Stat Soc B Methodol. 1949;11(1):15–44.
2. Berkson J, Gage RP. Survival curve for cancer patients following treatment. J Am Stat Assoc. 1952;47(259):501–515.
3. Amico M, Ingrid VK. Cure models in survival analysis. Ann Rev Stat Appl. 2018;5(1):311–342. doi:10.1146/annurev-statistics-031017-100101
4. Sekeres MA, Othus M, List AF, et al. Randomized phase II study of azacitidine alone or in combination with lenalidomide or with vorino stat in higher-risk myelodysplastic syndromes and chronic myelomonocytic leukemia: north american intergroup study SWOG S1117. J Clin Oncol. 2017;35(24):2745–2753. doi:10.1200/jco.2015.66.2510 [PubMed: 28486043]
5. Anderson JE. Bone marrow transplantation for myelodysplasia. Blood Rev. 2000;14(2):63–77. doi:10.1054/blre.2000.0126 [PubMed: 10913969]
6. Maller RA, Zhou X. Survival analysis with long-term survivors. New York, NY: Wiley; 1996.
7. Maller RA, Zhou S. Testing for the presence of immune or cured individuals in censored survival data. Biometrics. 1995;51(4):1197. doi:10.2307/2533253 [PubMed: 8589219]
8. Maller RA, Zhou S. Testing for sufficient follow-up and outliers in survival data. J Am Stat Assoc. 1994;89(428):1499–1506. doi:10.1080/01621459.1994.10476889
9. Shen PS. Testing for sufficient follow-up in survival data. Stat Prob Lett. 2000;49(4):313–322. doi:10.1016/s0167-7152(00)00063-8
10. Yu Y, Tiwari RC, Cronin KA, Feuer EJ. Cure fraction estimation from the mixture cure models for grouped survival data. Stat Med. 2004;23(11):1733–1747. doi:10.1002/sim.1774 [PubMed: 15160405]
11. van der Vaart AW. Asymptotic statistics. Cambridge: Cambridge University Press; 2007.
12. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19(6):716–723. doi:10.1109/TAC.1974.1100705
13. Petersdorf SH, Kopecky KJ, Marilyn S, et al. A phase 3 study of gemtuzumab ozogamicin during induction and postconsolidation therapy in younger patients with acute myeloid leukemia. Blood. 2013;121(24):4854–4860. doi:10.1182/blood-2013-01-466706 [PubMed: 23591789]

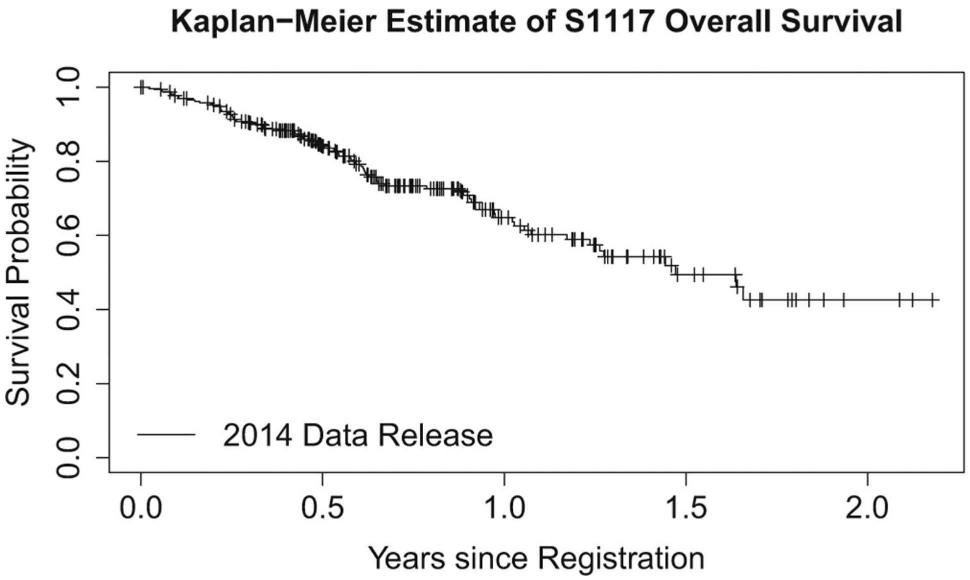


FIGURE 1.

Kaplan-Meier estimate for the survival function based on data of trial S1117 ending in 2014

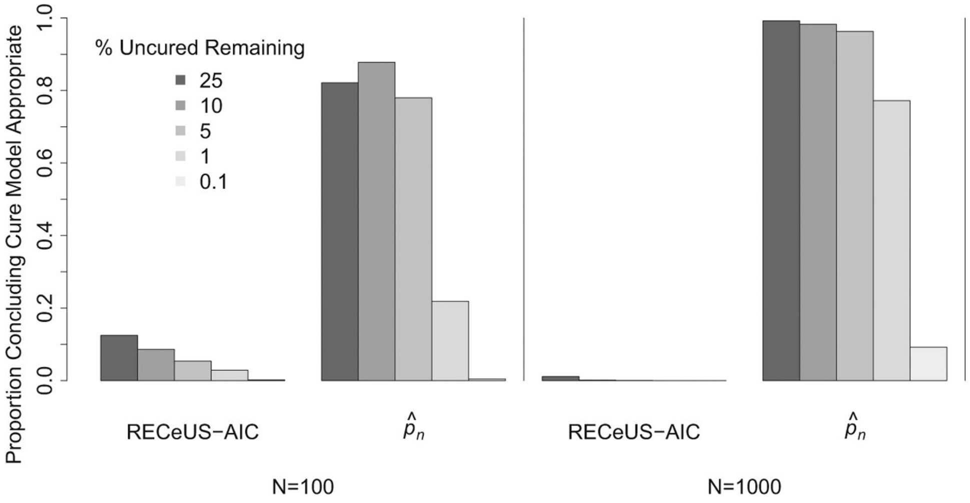


FIGURE 2.

The proportion of 10 000 simulations that (incorrectly) conclude a cure model is appropriate under settings with no cure fraction ($\pi = 0$) and varying percent uncured remaining, $\gamma_0(u)$*

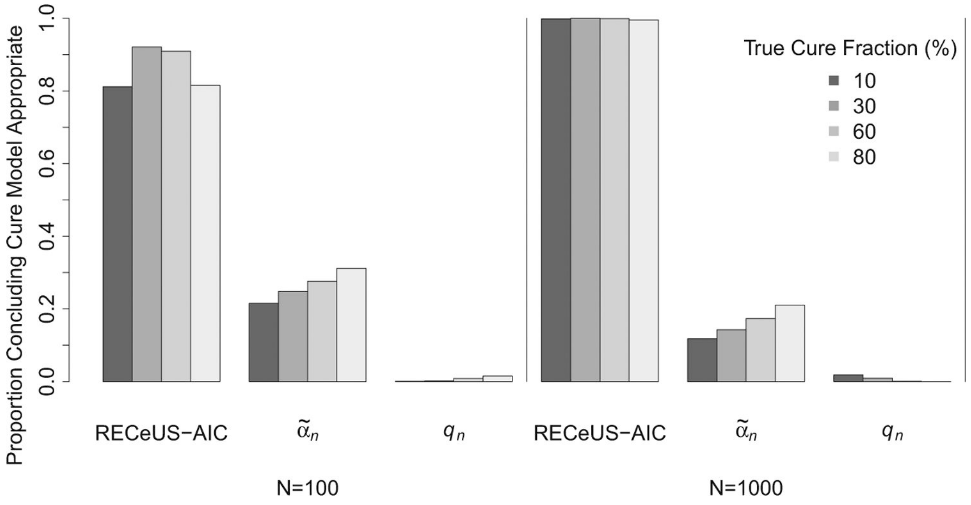


FIGURE 3.

The proportion of 10 000 simulations that (correctly) conclude a cure model is appropriate under settings with a nonzero cure fraction and 0.1% uncured remaining, $\gamma_{\pi}(0.001)$*

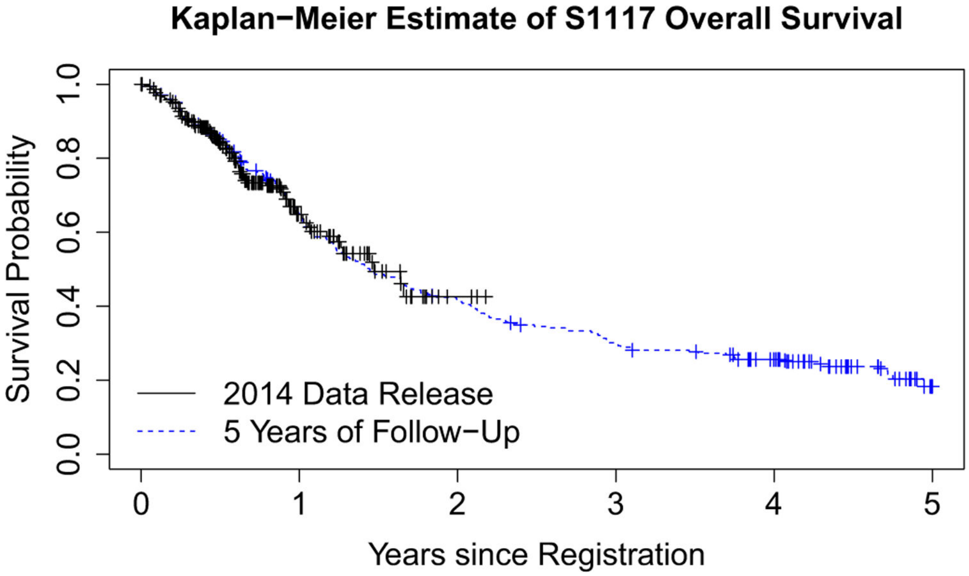


FIGURE 4.

Kaplan-Meier estimates for the survival function based on data through trial S1117 end in 2014 (black) and extended follow-up with five years of total follow-up (blue)

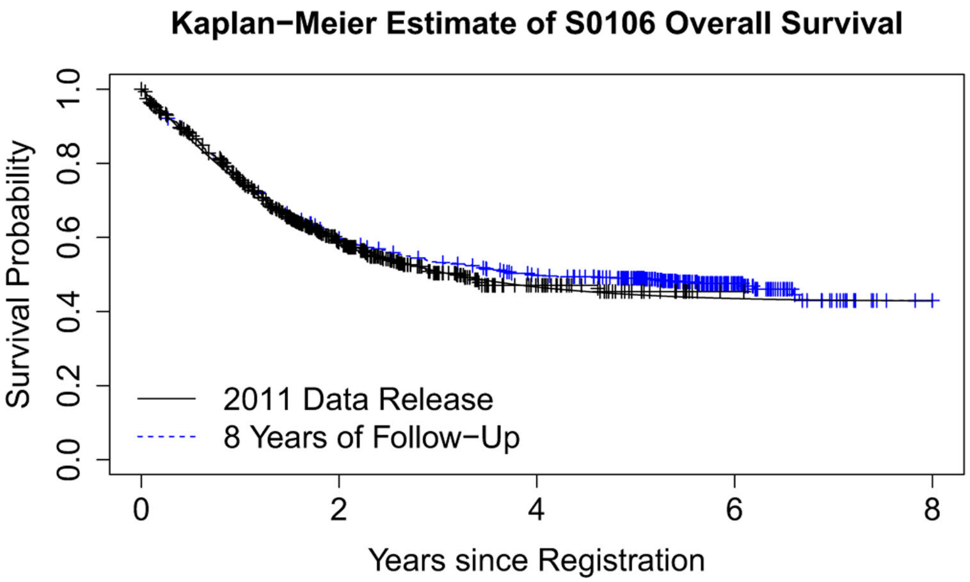


FIGURE 5.

Kaplan-Meier estimates for the survival function based on data through trial S0106 end in 2014 (black) and extended follow-up with data from 2018 (blue). The best-fitting mixture-cure model (mixture-cure Weibull model) based on 2011 data is overlaid in black

TABLE 1

Summary of method properties

	Testing for the presence of immunes only		Testing for sufficient follow-up only		
Desirable properties	d_n [7]	$\hat{p}_n$ [6]	$\hat{\alpha}_n$ [9]	q_n [6]	**Both RECeUS-AIC**
Nonparametric	N	Y	Y	Y	N
Formal hypothesis test	Y	Y	Y	Y	N
No subjective user input[a]	N	N	Y	N	N
Small $\gamma_0(u)$	N/A[b]	Y for small u, else N	N/A	N/A	Y
Large $\gamma_\pi(0.001)$, $\pi > 0$	N/A[b]	N/A	N	N	Y

[a] The d_n test requires evaluating goodness of fit of fitted parametric models. The $\hat{p}_n$ test requires rounding sample size and censoring rates to find a critical value in the published table. The q_n test requires rounding sample size, estimated cure fraction and input of practically sufficient follow-up time to find a critical value in the published table. RECeUS-AIC requires specifying thresholds for $\hat{\pi}_n$ and $\hat{r}_n$.

[b] The d_n test is not assessed in this article.

TABLE 2

Rates for concluding sufficient follow-up ($\gamma_{\pi}(u)^{*}$) for RECeUS-AIC by percentage of uncured remaining, sample size and cure fraction in 10 000 simulations with a Weibull(2, 1) mixture cure distribution

		% Uncured remaining[a]				
Cure fraction	**Sample size**	**25**	**10**	**5**	**1**	**0.1**
0%	100	0.125	0.086	0.054	0.028	0.002
	250	0.067	0.041	0.027	0.009	0.000
	500	0.036	0.016	0.006	0.002	0.000
	1000	0.011	0.001	0.000	0.000	0.000
	True ratio	1.0000	1.0000	1.0000	1.0000	1.0000
10%	100	0.130	0.102	0.104	0.327	0.811
	250	0.070	0.057	0.068	0.372	0.926
	500	0.049	0.032	0.036	0.386	0.969
	1000	0.020	0.007	0.007	0.355	0.998
	True ratio	0.7262	0.5409	0.3292	0.0599	0.0052
30%	100	0.159	0.157	0.214	0.588	0.921
	250	0.089	0.085	0.147	0.698	0.985
	500	0.058	0.066	0.102	0.809	0.999
	1000	0.047	0.034	0.062	0.921	1.000
	True ratio	0.4692	0.2820	0.1406	0.0208	0.0017
60%	100	0.197	0.243	0.328	0.655	0.909
	250	0.140	0.176	0.292	0.775	0.968
	500	0.089	0.119	0.255	0.857	0.992
	1000	0.072	0.083	0.224	0.924	0.999
	True ratio	0.3065	0.1641	0.0756	0.0105	0.0009
80%	100	0.201	0.256	0.338	0.577	0.816
	250	0.199	0.263	0.382	0.743	0.955
	500	0.152	0.210	0.351	0.823	0.982
	1000	0.100	0.149	0.339	0.901	0.995
	True ratio	0.2490	0.1284	0.0578	0.0079	0.0006

[a] In this setting, the exact follow-up times generating the data were 1.25, 1.5, 1.75, 2.25, and 2.75. These are the follow-up times corresponding to the percentiles rounded to the nearest quarter in order to more closely represent clinical trial data.