Lack of identifiability of plateaus

Hein Putter Per Kragh Andersen

Introduction

The presence of a latent sub-population that is cured of the disease of interest must be based on the tail of the time-to-event distribution. The most important practical limitation for establishing the presence of cure is that the tail of the time-to-event distribution is precisely the part of the distribution that is most difficult to estimate. Towards the end of follow-up, the numbers at risk are small, and therefore the estimate of the survival probabilities at the end of follow-up is very imprecise. The follow-up distribution is finite per definition. Looking at the tail of the Kaplan-Meier estimates and trying to say something about the behaviour of the survival function at infinity seems wishful thinking. The statistician can easily be misled, despite the fact that there are mathematically rigorous identifiability results about mixture cure models. The problem with these mathematical results is that they are valid under conditions that are impossible to test based on right-censored survival data.

We will illustrate the difficulties with a simulated data set. Full code in R for all data generation and analyses performed is available in the Quarto markdown document that generated this document. The true underlying data generation mechanism will be revealed towards the end of this document. Based on the simulated data we follow theory from Chapter 4 of the classical book of Maller and Zhou (1996) to first test for the presence of a cure fraction, and subsequently test for sufficiency of follow-up to establish and reliably estimate this cure fraction.

Figure 1 shows Kaplan-Meier survival curves for two values of a binary covariate x, based on 2500 observations from a true underlying survival curve $S(t \,|\, x = 0)$ (black) and 2500 observations from a true underlying survival curve $S(t \,|\, x = 1)$ (red). Both estimated survival curves seem to suggest a plateau, either with identical or different values around 55 to 60%. Figure 2 shows a plot of the estimated censoring distribution in the form of a reverse Kaplan-Meier estimate, assumed (correctly) to be common to $x = 0$ and $x = 1$. The plot strongly suggests a uniform censoring distribution between 2 and 3, which is in fact the true underlying censoring distribution.

The first question to address is whether there is evidence of a plateau. The first thing then is to estimate the proportion of immunes in each group, the natural estimator being the Kaplan-Meier evaluated at the last observed time point, irrespective of whether it is an event or censoring time point, which we will denote by $\hat{p}_n$. These Kaplan-Meier estimates are 0.560

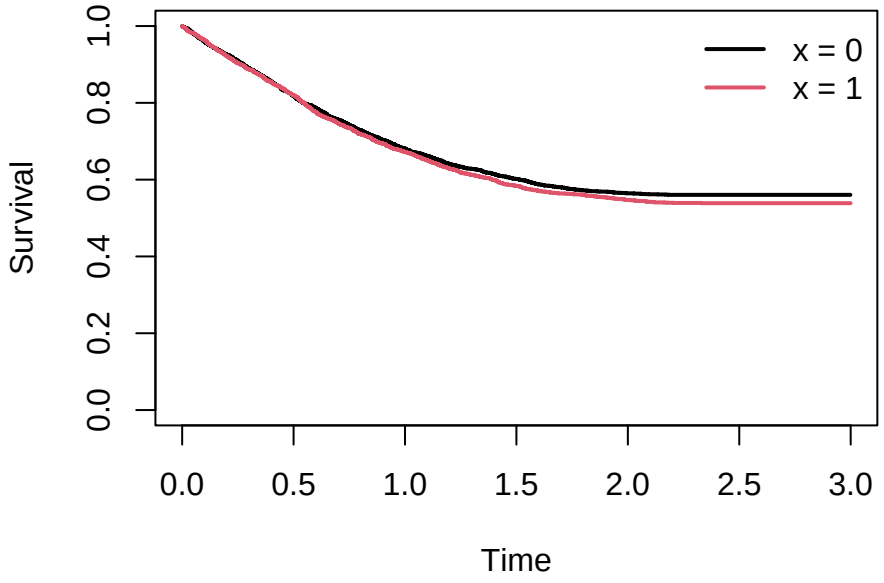

Figure 1: Kaplan-Meier survival curves for two values of a binary covariate suggesting a plateau

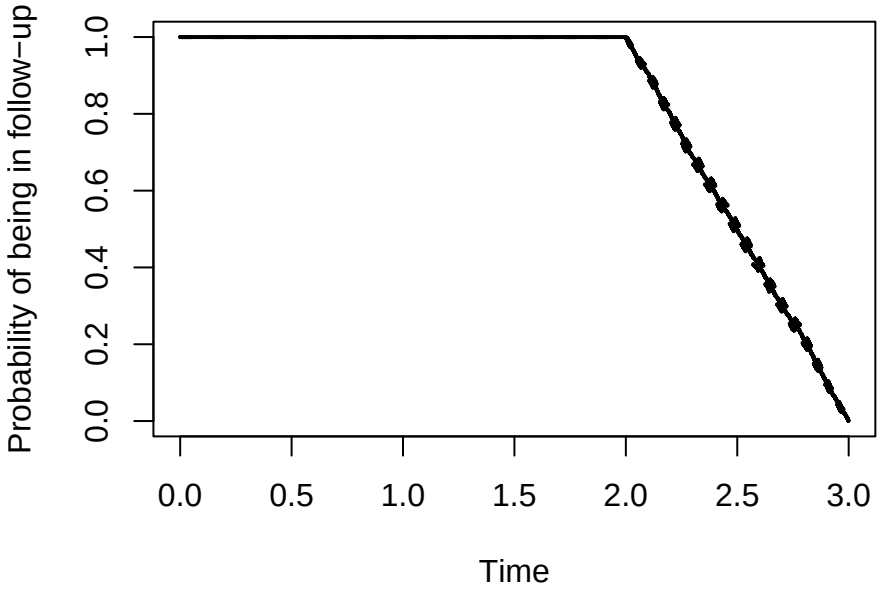

Figure 2: Reverse Kaplan-Meier estimate of the censoring distribution (assumed to be common to $x = 0$ and $x = 1$)

for $x = 0$ and 0.539 for $x = 1$. In what follows in this section, we will follow the theory of Maller and Zhou (1996) in testing for the presence of a plateau and sufficiency of follow-up. The theory is non-parametric, so we will apply it separately for $x = 0$ and $x = 1$. Maller and Zhou (1996) start out by testing $H_{01} : F(\tau_G) = 1$, in Section 4.2, where $F(\cdot)$ is the cumulative distribution function of the time-to-event, and τ_G is the end of follow-up. An alternative way of expressing this null hypothesis is $H_{01} : S(\tau_G) = 0$. If H_{01} is rejected, this is seen as evidence for a plateau, and we would move on to test sufficiency of follow-up, to be addressed later.

Testing for the presence of a plateau

The proposed test for H_{01} is to reject H_{01} if $\hat{p}_n > c_{0.05}$, where $\hat{p}_n$ as before is the Kaplan-Meier evaluated at the last observed time point (event or censored), and $c_{0.05}$ is the 5th percentile of the distribution of $\hat{p}_n$ calculated under H_{01}. To calculate the distribution of $\hat{p}_n$ under H_{01} is difficult in general, it will depend on the unknown distributions F and G of the survival and censoring times. Maller and Zhou (1996) approximate this distribution under a range of distributions of F and G, the percentiles of which are reported in a number of tables. We will not rely on their published tables, since their assumed censoring distributions are different from ours, but follow the reasoning of their procedure, and calculate the percentiles by simulation, separately under $x = 0$ and $x = 1$ (since we are going to test H_{01} in each of the subgroups). We take the censoring distributions (correctly) to be uniform on (2, 3), and, following Maller and Zhou (1996) the event distribution exponential with rates to be estimated from the observed data (separately for $x = 0$ and $x = 1$). The assumption of underlying exponential distributions may well be incorrect, but recall that we have to work from H_{01} anyway. Larger classes of survival models are of course possible, but preliminary analyses suggested that using Weibull distributions for instance do not change the results of our analyses in a major way. We used 1000 simulated data sets of size $n = 2500$ (same as the actual data) with $x = 0$ and with $x = 1$, generating event data from exponential distributions with rates estimated from the data (0.256 for $x = 0$ and 0.274 for $x = 1$) and applying independent uniform censoring on (2, 3). Within each of the simulated data sets we calculated and recorded $\hat{p}_n$. Figure 3 shows histograms of these simulated $\hat{p}_n$ under H_{01}, separately for $x = 0$ (left) and $x = 1$ (right), along with the estimated values from the actual data indicated with vertical lines.

The proportions of simulated $\hat{p}_n$ values under H_{01} that are larger than the estimated $\hat{p}_n$ in the data are 0 for $x = 0$ and 0 for $x = 1$. It is clear that we would reject H_{01}, both for $x = 0$ and $x = 1$, with $p < 0.001$.

Testing sufficiency of follow-up

That leaves the second question: is there sufficient follow-up to establish the presence of these plateaus? This is addressed in Section 4.3 of Maller and Zhou (1996), by defining the second hypothesis $H_{02} : \tau_F \leq \tau_G$, and its complement $H_{02}^c : \tau_F > \tau_G$, where τ_F and τ_G are

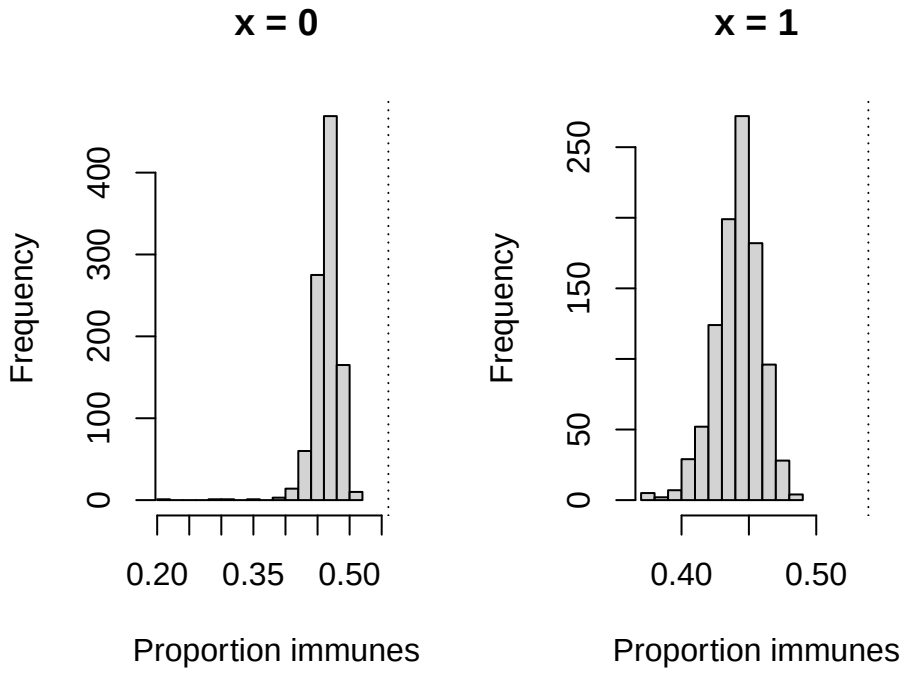

Figure 3: Histogram of the plateau estimates under H_{01}, for $x = 0$ (left) and $x = 1$ (right)

the right-hand side of the supports of F, the event-time distribution and of G, the censoring distribution. Rejecting H_{02}^c, and thereby accepting H_{02} would imply that all of the *potential* event times are contained in the *potential* follow-up. Importantly, this is different from all of the *actual* event times are contained in the *actual* follow-up, which is the only thing that we can really say anything about. But it is what Maller and Zhou (1996), also on much of their subsequent work propagate. In our opinion this is the real flaw in this mathematical approach to the problem of testing sufficiency of follow-up; it suggests mathematical rigor in a setting where many of the underlying components (τ_F, τ_G) are inherently unobservable. Maller and Zhou (1996) argue that we need to "use the information in the sample regarding the magnitude of $\tau_G - \tau_F$, which is effectively a measure of how far the large censored lifetimes (the potential immunes) lie from the main mass of the susceptibles' lifetime". Evidence for sufficient follow-up is to be found in the difference between the largest failure time $t_{\max}$ and the largest uncensored failure time $t^*_{\max}$. This leads to a test statistic for the null hypothesis H_{02}^c of the form $\delta_n = t_{\max} - t^*_{\max}$. Maller and Zhou (1992) also consider $q_n = N_n/n$, with N_n the number of uncensored t_i in $(2t^*_{\max} - t_{\max}, t_{\max}])$. The test will then be to reject H_{02}^c in favour of H_{02} if δ_n or q_n exceed certain critical values, and to accept H_{02}^c otherwise. These critical value then depend on the distribution of the test statistics under H_{02}^c. Again, these distributions can be approximated by simulation.

For $x = 0$ we see in our data that $t_{\max}$, $t^*_{\max}$, δ_n and q_n are given by 2.999, 2.184, 0.815, and 0.057, respectively. For $x = 1$, the observed $t_{\max}$, $t^*_{\max}$ and q_n are given by 3.000, 2.329, 0.671, and 0.026, respectively.

We again use the same setting to approximate the null distributions (under H_{02}^c) for δ_n and q_n, so in essence the same generated data sets were used as before, but this time the test statistics δ_n and q_n were calculated and stored for each generated data sets (separately for $x = 0$ and $x = 1$).

The resulting histograms are shown in Figure 4 and Figure 5. It is clear that both for $x = 0$ and $x = 1$, the estimated values obtained from the data are completely outside the histogram of the values generated under H_{02}^c, again leading us to the clear conclusion that follow-up is sufficient to establish and reliably estimate the plateaus.

Establishing cure? The true underlying data generating mechanism

Fitting a mixture cure model using {smcure} leads to a cure model with estimated plateaus of 0.560 for $x = 0$ and 0.539 for $x = 1$.

Results are in complete accordance with the plateaus seen by the Kaplan-Meiers, so together with the overwhelming evidence that both for $x = 0$ and $x = 1$ there is a plateau, and that we have sufficient follow-up to establish this and estimate its proportion, we can be quite confident about these results.

Or can we? Time to reveal the true nature of the data that we just generated. The true time-to-event distributions for $x = 0$ and $x = 1$ are the same. Figure 6 shows the true underlying hazard (left) and the true underlying survival function (right). The hazard is

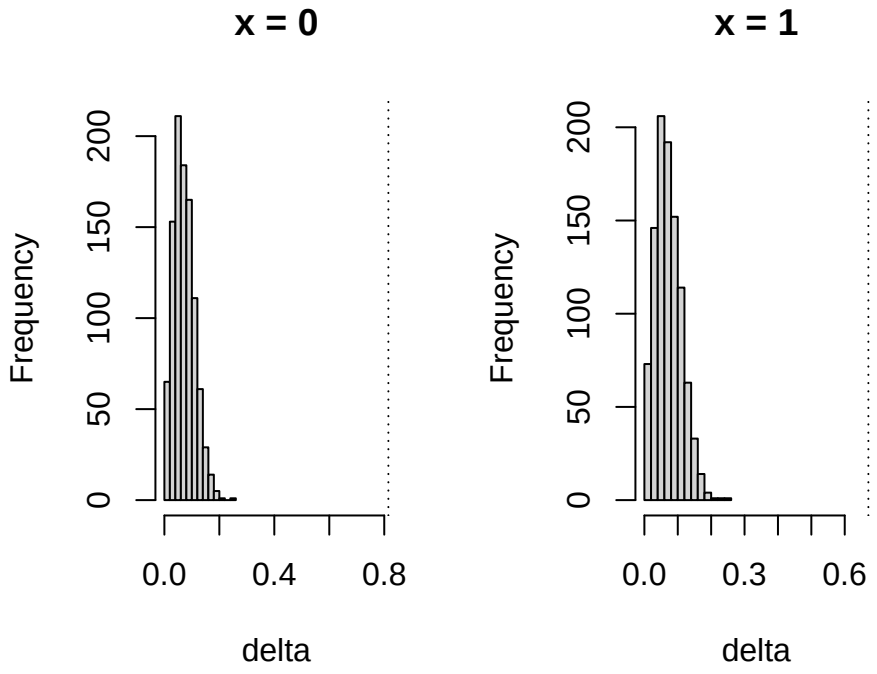

Figure 4: Histogram of d_n under H_{02}^c, for $x = 0$ (left) and $x = 1$ (right)

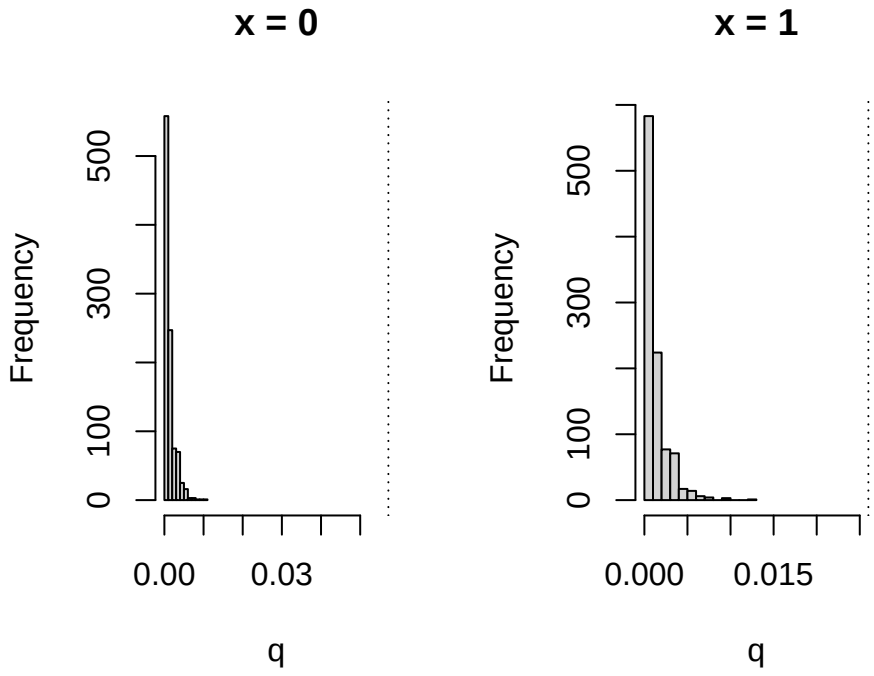

Figure 5: Histogram of q_n under H_{02}^c, for $x = 0$ (left) and $x = 1$ (right)

a sine wave function, periodically close to 0, around $t = 2.5, 6.5, 10.5$ etcetera. The true underlying survival function actually goes to zero as time goes to infinity, so in reality there is no proportion cured at all.

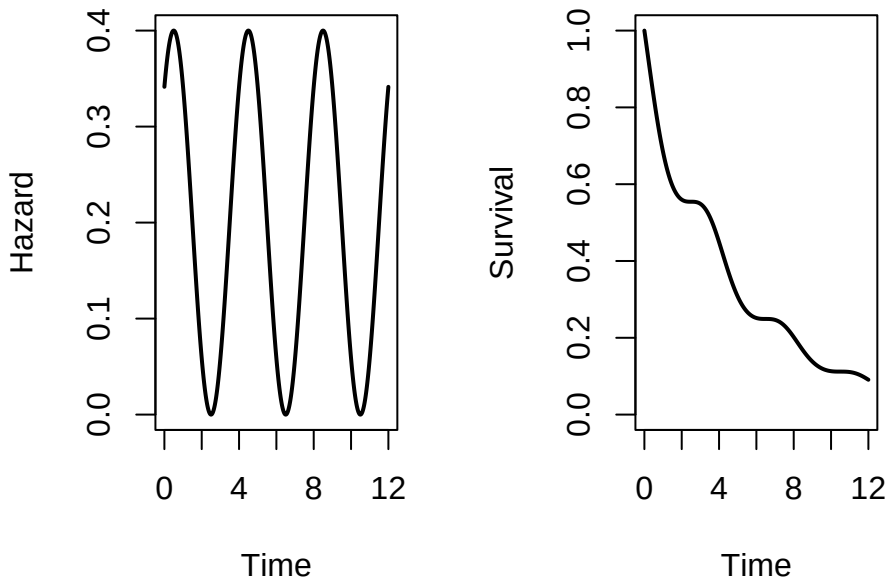

Figure 6: The true underlying hazard (left) and the true underlying survival function (right)

If we use the same uncensored data leading to Figure 1 with uniform censoring on $(2, 3)$, and instead change the censoring interval to $(6, 7)$, this leads to the Kaplan-Meier estimates for $x = 0$ and $x = 1$ as shown in Figure 7. This time a plateau around 25% is visible. Checking the existence of a non-zero plateau and sufficiency of follow-up using the methods of Maller and Zhou (1996) again leads us to be confident about these results, except for the sufficiency of follow-up for $x = 1$. Fitting a cure model to this data gives estimated plateaus of 0.251 for $x = 0$ and 0.246 for $x = 1$. Again we are misled about the existence of a plateau.

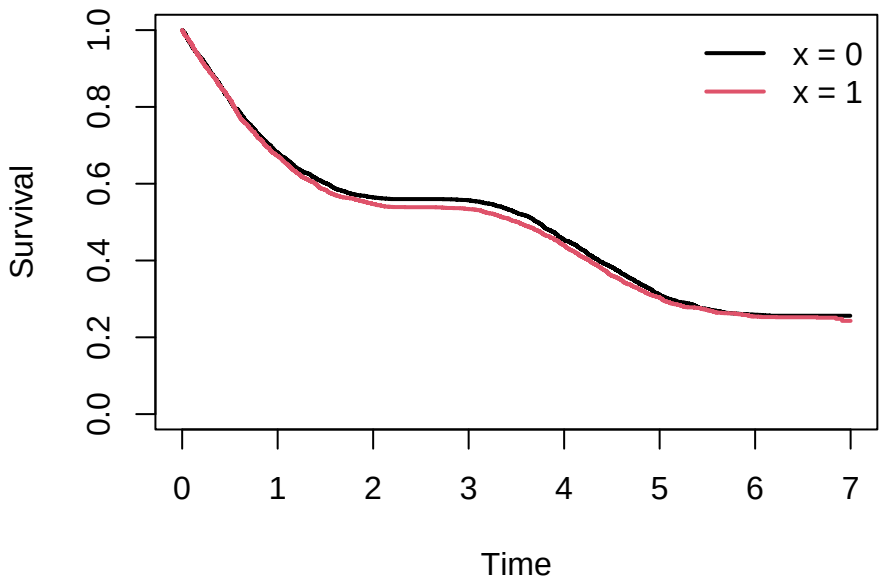

Figure 7: Kaplan-Meier survival curves for x=0 and x=1, uniform censoring on (6, 7)