

# Imputation for Censored Observations in Survival Studies Allowing for a Positive Cure Rate

Alan Cantor, University of Alabama at Birmingham  
Birmingham, AL

## ABSTRACT

This paper presents a method of imputation for right censored survival data in which the possibility that an individual with a censored survival time is cured is explicitly accommodated. For each right censored observation, first cure or non-cure is imputed. For the non-cures a survival time is imputed. All imputations are based on a Gompertz model and are conditioned on the censored time in the original data.

In a recent report<sup>1</sup>, Royston, Parmar, and Altman (RPA) suggest an approach to the visualization of survival data with right censoring as a complement to Kaplan-Meier plots. They suggest fitting a log-normal distribution to the data. Then a value can be imputed for a censored observation by generating a value for the random variable defined by this log-normal distribution conditioned on survival to the censored time. They then exhibit the enhanced data in the form of a dot plot. In this plot, a value imputed from a censored observation is distinguished from one based on a complete observation by the use of a different plotting symbol. RPA point out that such a plot complements the Kaplan-Meier plot by displaying actual survival times (known and imputed) as opposed to the estimates of cumulative mortality over time. In addition, they argue that, when used for different treatment or prognostic groups, such plots visually present the overlap or separation of these groups in a manner that is more informative than the distance between Kaplan-Meier curves.

RPA acknowledge, as a limitation of their approach, that it may result in some imputed survival times that are unrealistic – in fact beyond the human lifespan. This is especially true if there is a substantial amount of censoring. In such cases, their method can not be used reliably. Instead, it might be reasonable to consider the possibility that the survival data are sampled from a population in which a non-zero proportion is cured. It might, therefore, be appropriate to replace the log-normal distribution with a survival distribution that allows for cure. We would then use that distribution to impute cure or non-cure as well as a survival time for those who are imputed as not cured.

keywords: Survival Analysis, Imputation, Cure Models

## INTRODUCTION

Although much has been written about the use of imputation to replace missing data with imputed values, until recently that literature did not include data that were missing due to censoring. This is, of course, a type of “missingness” that is common in clinical trials with survival as an endpoint, since it is generally not practical to follow each patient until death. Non-parametric imputation has been discussed by Taylor, Murray, and Hsu<sup>2</sup> and Royston, Parmar, and Altman<sup>1</sup> presented the parametric approach described above.

When dealing with disease-specific survival data for which the Kaplan-Meier curve seems to have a horizontal asymptote above the time axis, a non-zero cure rate is plausible. In such cases, any attempt at imputation for right censored observations needs to consider the possibility that a patient with a censored survival time may, in fact, be cured and will not die of the disease being studied at all.

## CURE MODELS

There is a vast and growing literature discussing what are known as “cure models”; i.e. models for survival based on survival functions  $S(t)$  which approach a positive number (instead of zero) as  $t$  approaches infinity. Such models were first discussed by Boag<sup>3</sup> and Berkson and Gage<sup>4</sup>. Both used models of the form  $S(t) = p + (1-p)S^*(t)$  where  $S^*(t)$  is a standard survival function that goes to zero as  $t$  approaches infinity and  $0 < p < 1$ . More generally, if  $h(t)$  is a hazard function, then the associated survival function is

$$S(t) = \exp\left[-\int_0^t h(u)du\right] \text{ and } S(\infty) = \lim_{t \rightarrow \infty} S(t) = \exp\left[-\int_0^\infty h(u)du\right] \text{ which is positive if and only if } \int_0^\infty h(u)du < \infty.$$

## PARAMETRIC IMPUTATION USING A CURE MODEL

We begin by using the original data, including the censored observations, to compute maximum likelihood estimates,  $\hat{\theta}$ , of the parameters,  $\theta = (\theta_1, \theta_2, \dots, \theta_k)'$  of a survival distribution  $S(t, \theta)$  that allows for cure. For a specified survival distribution, this is accomplished with PROC NLP. From the estimated survival function,  $S(t, \hat{\theta})$ , we can then estimate of the probability of cure conditioned on a right censored time of  $t_0$ ,  $S(\infty, \hat{\theta} \mid T > t_0)$ , and the survival function of the non-cures conditioned on a right censored time of  $t_0$ ,  $S(t, \hat{\theta} \mid t_0 < T < \infty)$ . A right censored observation at  $t_0$  can be imputed to be a cure or non-cure and a survival time imputed for the imputed non-cures in three ways.

### FIXED IMPUTATION

If  $S(\infty, \hat{\theta} \mid T > t_0) > .5$ , then impute as a cure. Otherwise impute as a non-cure. If the censored observation is imputed as a non-cure, impute the survival time to be the median of the conditional survival distribution  $S(t, \hat{\theta} \mid t_0 < T < \infty)$ .

### RANDOM IMPUTATION

Using a uniform random number generator, impute the censored observation as a cure with probability  $S(\infty, \hat{\theta} \mid T > t_0)$ . If the censored observation gets imputed as a non-cure, impute a survival time as a value of a random variable from the conditional survival distribution  $S(t, \hat{\theta} \mid t_0 < T < \infty)$ .

### AVERAGING IMPUTATION

Assign weights of one to all patients with uncensored survival times. For each censored survival time,  $t_0$ , create an imputed cured patient with weight equal to  $S(\infty, \hat{\theta} \mid T > t_0)$  and create an imputed non-cured patient with weight equal to  $1 - S(\infty, \hat{\theta} \mid T > t_0)$  and imputed survival time equal to the median of the conditional survival distribution  $S(t, \hat{\theta} \mid t_0 < T < \infty)$ . The sum of the weights of the imputed cures will give us the expected number of cures based on the model and the data. If we divide the time interval into sub-intervals, the sum of the weights for the non-cures in each sub-interval would give us the expected number of survival times in each of these sub-intervals. For plotting purposes we could round each of these sums to the nearest integer.

## THE GOMPERTZ DISTRIBUTION

In 1825, Benjamin Gompertz reported on his studies of a life table for the mortality of an aging male population<sup>5</sup>. In modern terminology, he found that these data fit very well to a model in which the hazard increases exponentially; i.e.  $h(t) = \alpha \exp(\beta t)$  with  $\alpha > 0$  and  $\beta > 0$  and shows that this implies the survival function  $S(t) = \exp[(\alpha/\beta)\{1 - \exp(\beta t)\}]$ . Over 100 years later, Haybittle<sup>6</sup>, studying survivorship of women with breast cancer, noted that the hazard seemed to decrease exponentially. Of course, this leads to the same survival function with  $\beta < 0$ . It is easy to see that the cure rate is  $\exp(\alpha/\beta)$ . It is convenient to write  $\pi = \exp(\alpha/\beta)$ . Then  $S(t) = \pi^{1 - \exp(\beta t)}$ . The survival distribution of the non-cures is  $S(t \mid t < \infty)$  which equals  $[(S(t) - \pi)]/(1 - \pi)$ . For an observation censored at  $t_0$ , the probability of cure conditioned on survival to  $t_0$  is  $\pi^{\exp(\beta t_0)}$  and the survival distribution of the non-cures conditioned on survival to  $t_0$  is

$$S(t \mid t_0 < t < \infty) = [S(t) - \pi]/[S(t_0) - \pi]. \text{ The median of this distribution is } \frac{\log \left\{ \frac{-\log[(\pi^{-\exp(\beta t_0)} + 1) / 2]}{\log(\pi)} \right\}}{\beta}.$$

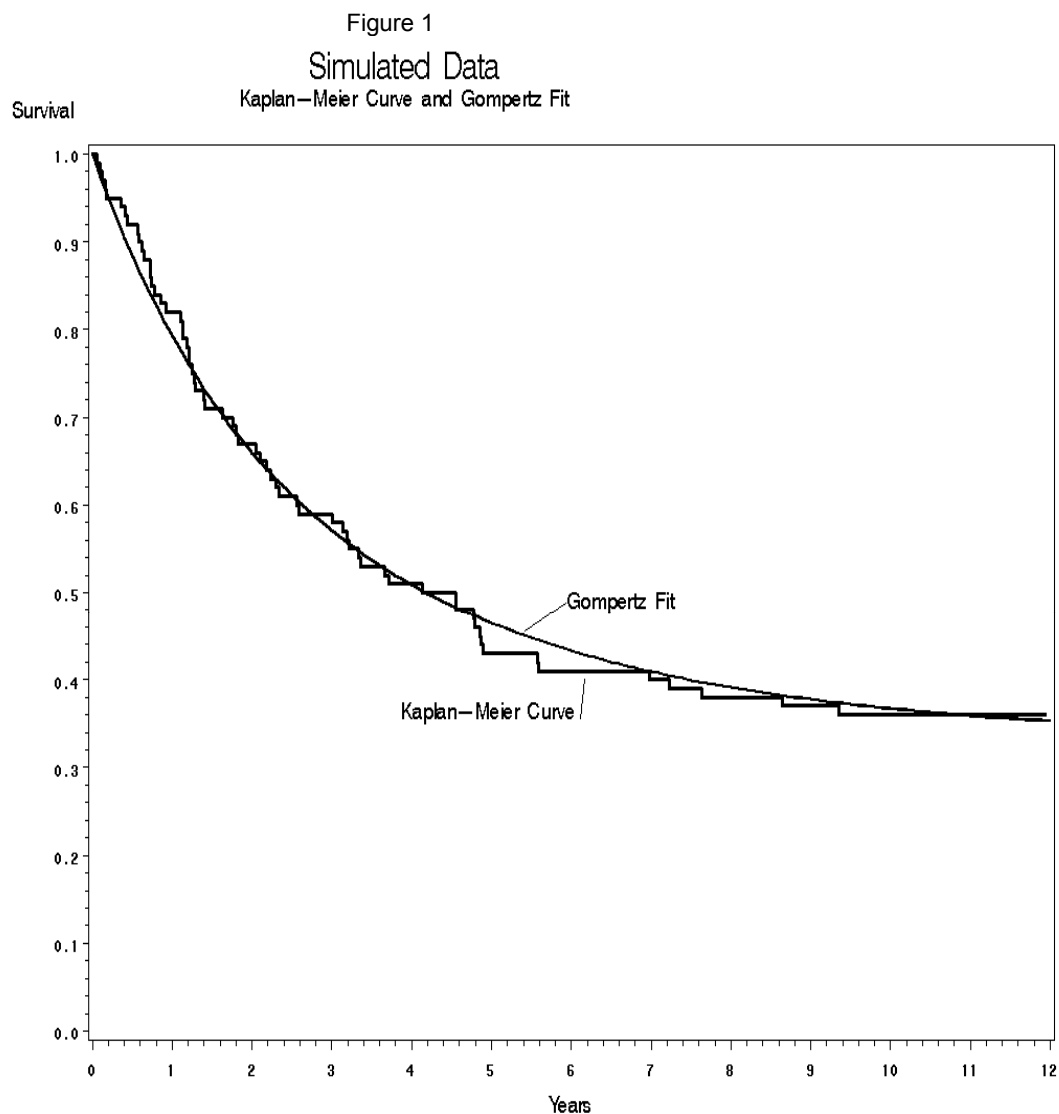
### AN EXAMPLE

As an example of how this works, we consider a set of simulated data. These data represent a simulated clinical trial in which 100 patients are accrued uniformly over three years and followed for an additional nine years after the end of the accrual period. Their survival data were simulated from a Gompertz survival distribution with a cure rate of 35% and a median survival time for the non-cures of two years. Setting  $S(\infty) = \exp(\alpha/\beta) = 0.35$  and  $[S(2) - \pi]/(1 - \pi) = .5$ , we find  $\alpha = .2462$  and  $\beta = -.2345$ . For this dataset, 36 survival times were censored. Thirty-five of them were cured.

The other was censored because of a survival time (over 13 years) that was past the follow-up period. Of course, this survival time could not be observed. Instead a censored time of 10.02 years was observed.

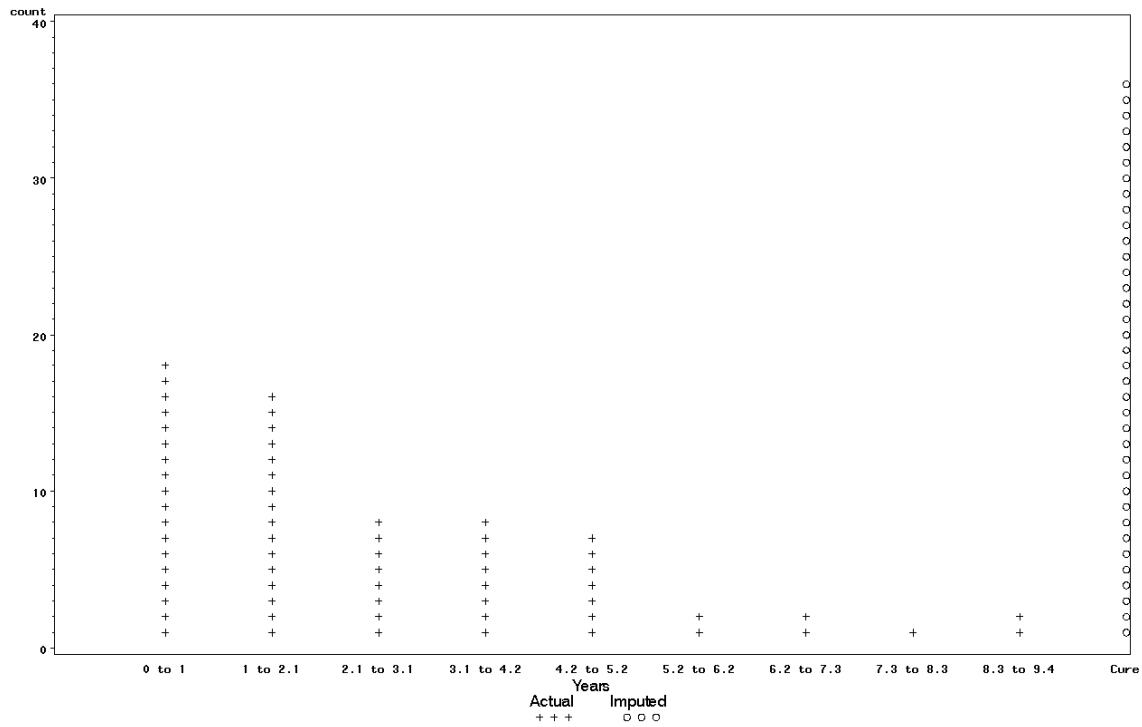
Using PROC NLP to maximize the likelihood, we find maximum likelihood estimators,  $a = 0.2609$  and  $b = -0.2355$ , for  $\alpha$  and  $\beta$  respectively. For an observation censored at time  $t_0$ , the estimates of the conditional survival described above can be written by replacing  $\alpha$ ,  $\beta$ , and  $\pi$  by  $a$ ,  $b$ , and  $p = \exp(a/b) = 0.3304$  respectively.

The Kaplan-Meier curve for this simulated data set is shown in figure 1 along with the Gompertz curve using the maximum likelihood estimates of the parameters.



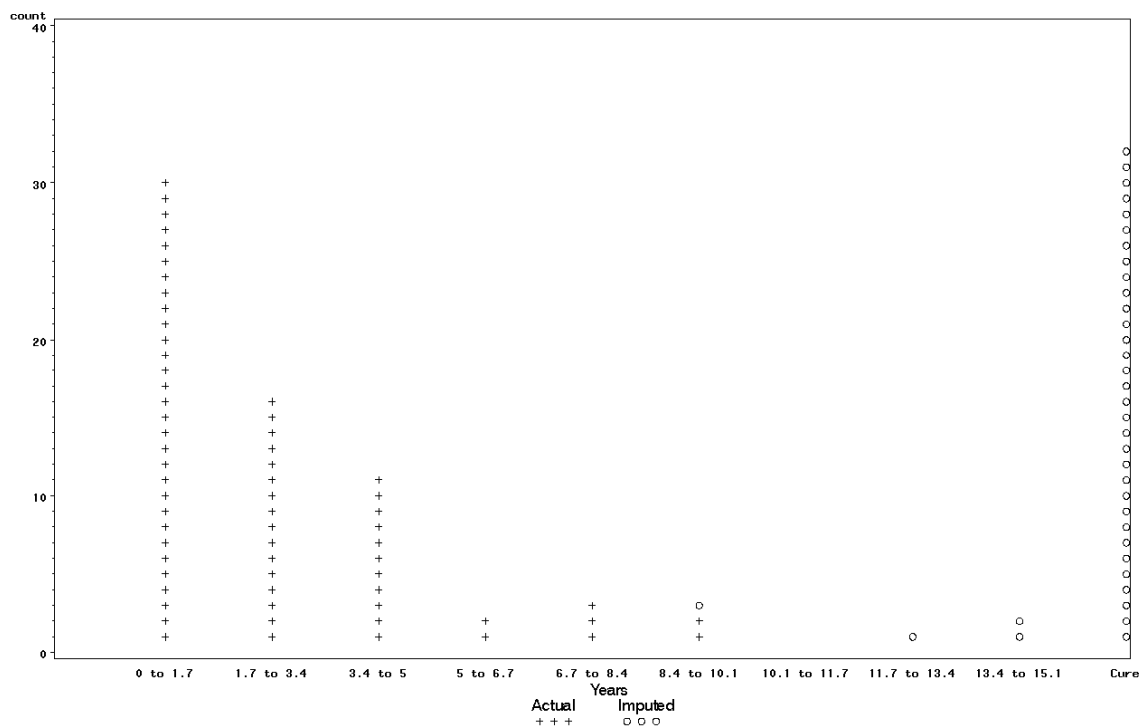
For the method of fixed imputation, the plot of the imputed cures and survival times is shown as figure 2. Note that all 36 of the censored survival times are imputed to be cures. In fact, for a censored time of  $t_0$ , the conditional probability of cure given that time will exceed 0.5 if  $t_0 > \log[\log(0.5)/\log(p)]/b$  which in this case is 1.9 years. Thus, for the data generated as described above, all censored survival times will have conditional probabilities of cure greater than 0.5 and will be imputed as cures

Figure 2  
Fixed Imputation



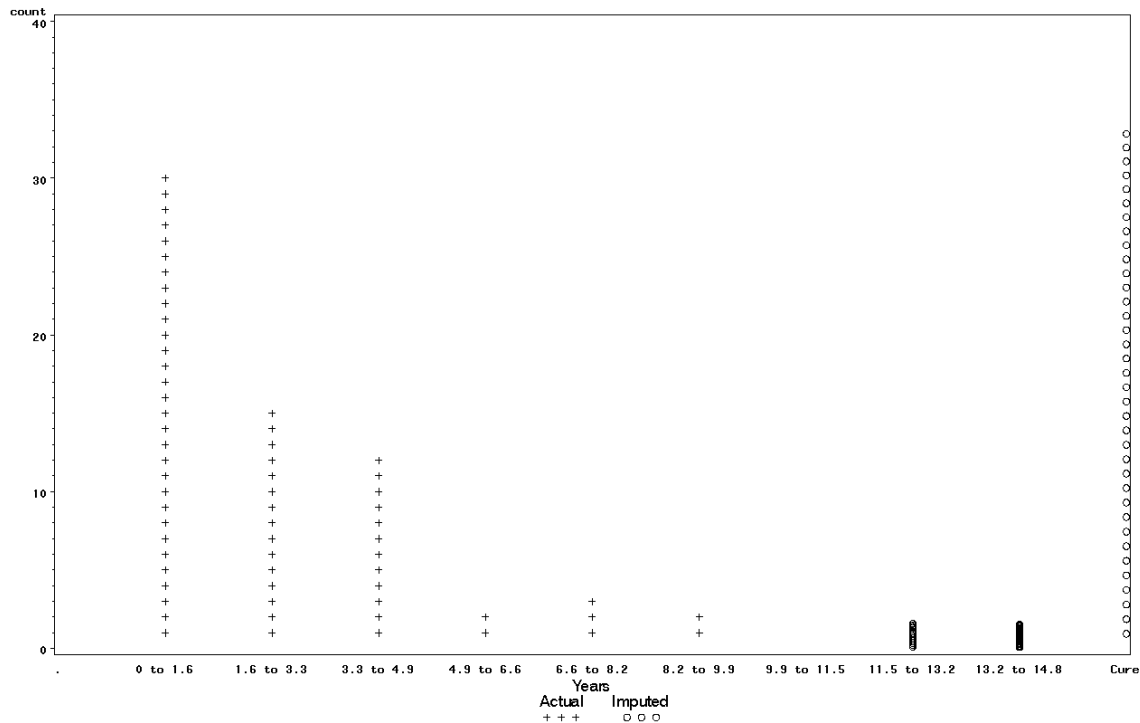
The plot of the imputed cures and survival times for the random imputation method is given in figure 3. In this case, four of the thirty-six censored observations are imputed to be non-cures although three of those are actually cures.

Figure 3  
Random Imputation



For the method of averaging, each of the censored times produces two weighted observations. One is a cure with weight equal to the conditional probability of cure given that censored time. The other is a non-cure with complementary weight and survival time equal to the median of the conditional survival distribution for the non-cures given that censored time. In the plot given as figure 4, the circles representing the imputed observations now are for fractional observations and the heights of each of the last three columns are the expected numbers in these columns. Thus there are 15 and 21 fractional observations in the last two non-cure columns; but each of these columns represents a total of about 1.5 expected survival times in that interval. The column of cures has 36 fractional observations representing an expected total of about 33.

Figure 4  
Averaged Imputation



Although only the plots of the number of imputed cures and the number of survival times in each time interval are presented here, the computer program that produces them also produces datasets containing detailed information about the imputations for each censored observation. A plot of the Kaplan-Meier curve for the original data as well as the Gompertz curve that is fit to the data by using the maximum likelihood estimates of the parameters is also produced. This additional information is available to the user; however imputed values, while useful in visualizing results of survival studies, should not be used in analyses involving statistical inference. Any such analyses, to be valid, would have to take into consideration the additional variance due to the imputation process. Appropriate methods of adjustment for this have yet to be studied.

A computer program, written in SAS®, can be obtained from the author.

SAS and all other SAS Institute Inc. product or services names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

#### References:

1. Royston P, Parmar MKB, Altman DG, Visualizing Length of Survival in Time-to-Event Studies: A Complement to Kaplan-Meier Plots, *Journal of the National Cancer Institute*, 100:2, 2008, 92-97
2. Taylor, JMG, Murray S, Hsu CH, Survival Estimation and Testing Via Multiple Imputation, *Probability & Statistics Letters*, 58, 2002, 221-232
3. Boag JW, Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy, *Journal of the Royal Statistical Society, Series B*, 11, 1949, 15-44.
4. Berkson J and Gage P, Survival Curves for Cancer Patients Following Treatment, *Journal of the American Statistical Association*, 47, 1951, 501-515

5. Gompertz B, On the Nature of the Function Expressive of the Law of Human Mortality and on the New Mode of Determining the Value of Life Contingencies, Philosophical Transactions of the Royal Society, Series A, 115, 1825, 513-580
6. Haybittle JL, The Estimation of the Proportion of Patients Cured After Treatment for Cancer of the Breast, British Journal of Radiology, 32, 1959, 725-733