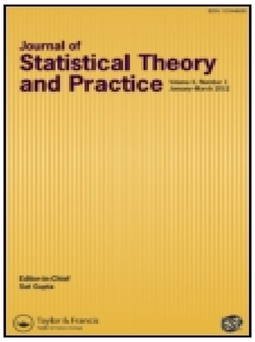

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Theory and Practice

ISSN: 1559-8608 (Print) 1559-8616 (Online) Journal homepage: http://www.tandfonline.com/loi/ujsp20

Survival analysis modeling with hidden censoring

Mahdi Saber Raza & Mark Broom

To cite this article: Mahdi Saber Raza & Mark Broom (2016) Survival analysis modeling with hidden censoring, Journal of Statistical Theory and Practice, 10:2, 375-388, DOI: 10.1080/15598608.2016.1152205

To link to this article: http://dx.doi.org/10.1080/15598608.2016.1152205

Accepted author version posted online: 22 Feb 2016.
Published online: 22 Feb 2016.

JOURNAL OF STATISTICAL THEORY AND PRACTICE
2016, VOL. 10, NO. 2, 375–388
http://dx.doi.org/10.1080/15598608.2016.1152205

Survival analysis modeling with hidden censoring

Mahdi Saber Raza and Mark Broom

Department of Mathematics, City University London, London, United Kingdom

ABSTRACT
There are well-established survival analysis methodologies for data sets that are complete, with accurate information on censoring. But what if they are not complete? In this article we consider how to analyze cases where "hidden censoring" occurs, where individuals have effectively left the study but the hospital is unaware of this. We develop a new Markov chain-based methodology for generating survival curves and hazard functions, and demonstrate this using a breast cancer data set from the Kurdistan region of Iraq.

ARTICLE HISTORY
Received 20 July 2015
Accepted 5 February 2015

KEYWORDS
Survival analysis; censored data; Markov chain; breast cancer; Kurdistan

AMS SUBJECT CLASSIFICATION
62N01

1. Introduction

The modeling of survival has a long history, and there are well-established methodologies for estimating survival probabilities for individuals with a given medical condition (Cox and Oakes 1984; Crowder 2012; Crowder et al. 1991; Lawless 2003). Essentially identical methods are also used for modeling the failure of items, such as components in machines (see, e.g., Barlow and Proschan 1975; Bedford and Cook 2009). The most fundamental functions used in survival analysis are the *survivor function*, which is the probability that an individual survives beyond time t, and the *hazard function*, which is the risk of death (per unit time). Thus, if our individual has lifetime distribution T, following standard terminology (see, e.g., Chap. 2 of Cox and Oakes 1984), the survivor function is

$$S(t) = P[T > t] \tag{1}$$

and the hazard function is

$$h(t) = -\frac{\frac{d}{dt}S(t)}{S(t)}. \tag{2}$$

Expressing the relationship in Eq. (2) the other way round, we obtain

$$S(t) = e^{-\int_0^t h(u)du}. \tag{3}$$

These fundamental properties can be estimated directly from data in a number of ways, but perhaps the simplest and most robust is the Kaplan–Meier estimator, which estimates the hazard function, using the discrete hazard function

CONTACT Mark Broom ✉ Mark.Broom@city.ac.uk Department of Mathematics, City University London, Northampton Square, London, EC1V 0HB, United Kingdom.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujsp.

$$h_j = \frac{d_j}{n_j}, \tag{4}$$

where d_j is the number of observed deaths within a particular (unit) interval, and n_j is the number of individuals at risk at the start of that period. The survivor function is then estimated by

$$\hat{S}(t) = \prod_{j=1}^{t} (1 - h_j). \tag{5}$$

This method automatically takes into account any data censoring, where an individual is known to leave the study at a particular time, by reducing the number at risk n_j by the number of censored individuals c_j. Thus, we update the number at risk as follows:

$$n_{j+1} = n_j - d_j - c_j. \tag{6}$$

A great advantage of this method is that reliable estimates can be obtained without making assumptions about the underlying distribution T. The only information that we need to apply this methodology is, for all times where we take measurements, knowledge of the values of the total number of individuals at risk at the start of the time period and the total number of deaths within the time period, that is, all vallues of n_j and d_j. This then enables robust comparisons between survival curves from different studies, perhaps between different types of treatment, different times, or different countries, and helps clinicians to assess the effectiveness of different approaches. Significant censoring can be factored in as already described without problems, provided that records are sufficiently good to know when contact with patients has been lost. The focus of this article is how to tackle problems when you do not have this knowledge, and significant "hidden" censoring occurs unknown to the researchers, using a real example as a case study. In section 3 we present two models, one without and one with censoring, that address this problem. In fact, our models, in particular the second model, do make parametric assumptions, based as they are on an underlying Markov-chain model. Nevertheless, the models, in particular the simpler first model, are robust to (at least certain types of) deviation from them.

2. A Kurdish breast cancer data set

Breast cancer is the most common cancer in the West, affecting a large number of women (and men) at some point in their lives (World Health Organization [WHO] 2008). There are various risk factors, such as obesity, age, and hormone replacement therapy during menopause (Lan et al. 2013; Robb et al. 2007; Rudat et al. 2013). Breast cancer has been well studied, and treatments are becoming more sophisticated and successful (De Santis et al. 2014). The American Cancer Society reports that around 250,000 breast cancer cases are diagnosed in the United States per year, and of these, almost 10% affect women under the age of 45 years. While this percentage may sound relatively insignificant in comparison to the total number of women diagnosed annually, it is a noteworthy ratio, particularly when compared to other cancers. In women under 40, breast cancer is the leading cause of cancer deaths (Ries et al. 2007). The United Kingdom currently has the 11th highest breast cancer rate, with 89.1 of every 100,000 women every year expected to develop breast cancer (NHS Choices 2011). Breast cancer is also becoming a more common disease in the developing world (Ozmen 2006). In particular, we are interested in the incidence of breast cancer in the Kurdistan Region of Iraq. This has not received a lot of detailed attention; examples include Majid et al. (2009), Othman et al. (2011), Majid et al. (2012), and Shabila et al. (2012) (for other work on breast cancer in the Kurdistan region see also, e.g., Alwan et al. 2000; Hughson 2012; Hussaion 2009; for work on the incidence of breast cancer elswhere in the wider region see Al Tamimi et al. 2010; Dey et al. 2010; Rennert 2006; Sughayer et al. 2006). These papers addressed various important questions, especially related to the incidence of breast cancer, but so far no detailed survival analysis has been carried out for the Kurdish region of Iraq.

We consider a data set of breast cancer patients from Nanakaly Hospital. Nanakaly Hospital is a public-sector hospital in Erbil, the capital of the Kurdistan region of Iraq, was established in 2004, and is funded by the Kurdistan government. It is a center concerned with all types of cancer. The hospital registry department collects data regarding the type of cancer and the age of the patients, the time of diagnosis, and (if appropriate) the time of death, as well as personal details, and these are registered on the statistical database. We have access to the most recent data on breast cancer, minus the personal information.

Detailed times of death were provided, with censoring only at the end of the study period on June 1, 2014; see Figure 1 for an illustration. Analyzing this data set using SPSS provided the Kaplan–Meier survival curve in Figure 2. The function flattens out to effectively a horizontal line, indicating a hazard rate tending to zero. This is clearly not a realistic survival curve. For comparison, a survival curve for a set of breast cancer data from Schumacher et al. (1994) is shown in Figure 3. The problem with the survival curve from Figure 2 is that we calculated it on the assumption that all individuals other than those who died (or were censored by reaching the end of the study period) were still active in the study, but in fact individuals often did not return

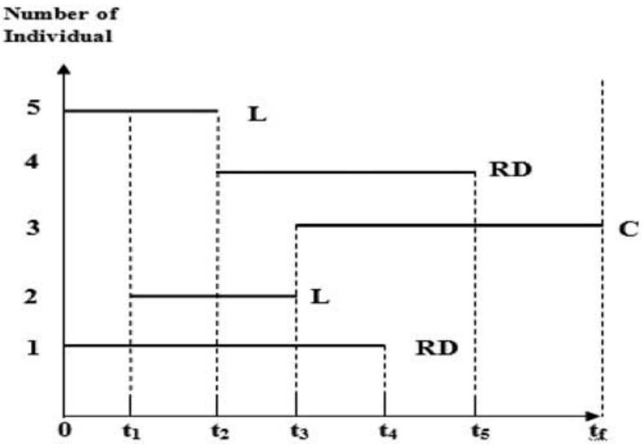

Figure 1. Illustrative plot of survival times including end of period censoring (C), recorded death (D), and hidden censoring, individuals unknowingly lost to the study (L), for the Kurdish data from Nanakaly Hospital.

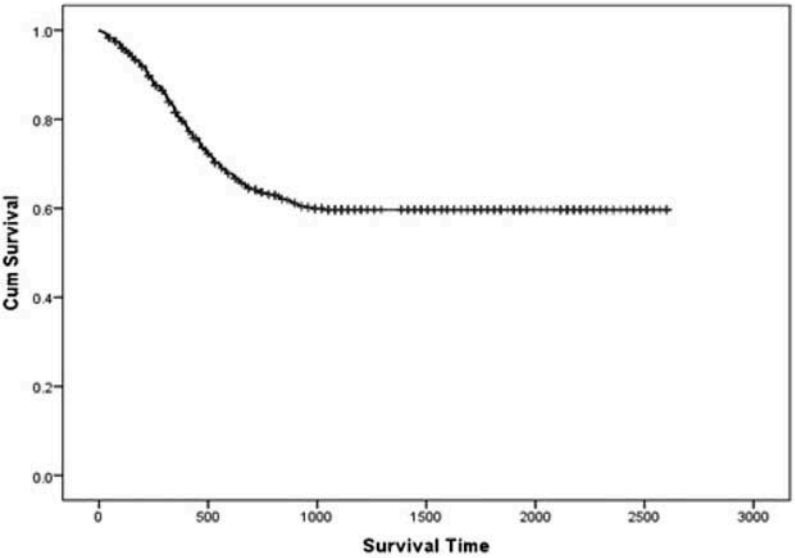

Figure 2. The original survival curve for the Kurdish data from Nanakaly Hospital.

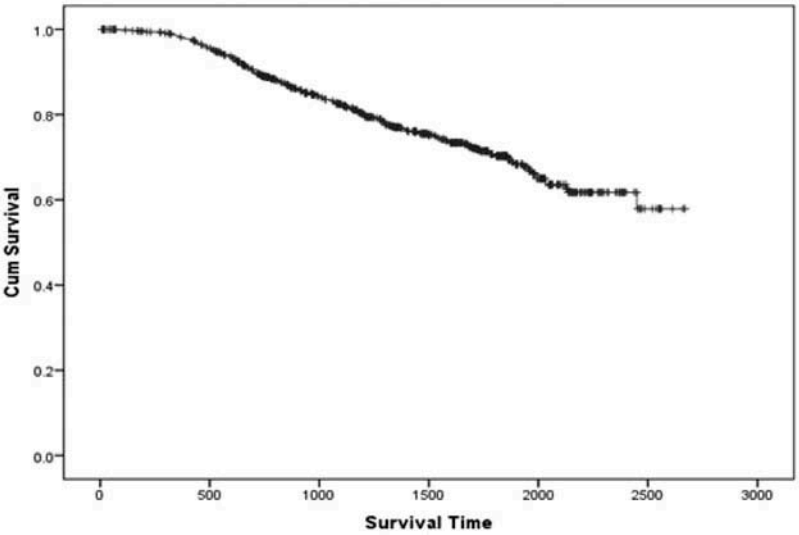

Figure 3. Survival curve for the German data from Schmoor et al. (1996) and Schumacher et al. (1994).

to the hospital after initial treatment, and there are no clear records of when the deaths of these individuals occur, or of which individuals these are. Thus, there is some secret censoring that we do not have knowledge about. We can think of this to mean that while the values of d_j are accurate, the values of n_j are not, and we are (after some time, greatly) overestimating them.

This article presents two related methods for overcoming this problem and obtaining a realistic survival curve for the Nanakaly data. The data will be analyzed more fully in a separate paper.

3. Markov models

3.1. A Markov model without censoring

We first introduce a continuous-time Markov model without overt censoring (Cox and Miller 1965; Grimmett and Stirzaker 2001). In our data the only observed censoring was caused by the end of the study period, although as patients were being recruited all the time during the period, the censoring time could be small, and such censoring could occur for any time less than 2602 days, the time from the earliest record considered to the end of the study period. Thus, all of the individuals censored and removed from the number at risk in the standard way following Eq. (6) (see Eq. (15)) was censoring of this type.

Consider a population of individuals in three categories: either at risk (I), died (D), or who have left the study (without our knowledge), which we call "lost" (L). Individuals simply move from state I to the other two states at constant rates l to L and p to D. We thus have a population as described by Figure 4. We denote the proportion of individuals in states I, L, and D at time t by $P_I(t)$, $P_L(t)$, and $P_D(t)$, respectively. State I cannot be entered, and is left at constant rate per individual $l+p$; thus, we obtain the differential equation (see, e.g., Chap. 8 of Haigh 2002):

$$\frac{d}{dt}P_I(t) = -(l+p)P_I(t). \tag{7}$$

Since at time 0 every individual is in the "at risk" category, so that $P_I(0) = 1,$ we obtain

$$P_I(t) = e^{-(l+p)t}. \tag{8}$$

Since state D is entered at rate $pP_I(t),$ we also have

$$\frac{d}{dt}P_D(t) = pP_I(t), \tag{9}$$

which using Eq. (8) and the fact that $P_D(0) = 0$ yields

$$P_D(t) = \frac{p}{l+p}\left(1 - e^{-(l+p)t}\right). \tag{10}$$

Since $P_L = 1 - P_I - P_D$, we obtain

$$P_L(t) = \frac{l}{l+p}\left(1 - e^{-(l+p)t}\right). \tag{11}$$

Suppose that, as in the original survival plot, we consider the data without realising that the category L exists. We can see from Eqs. (10) and (11), that

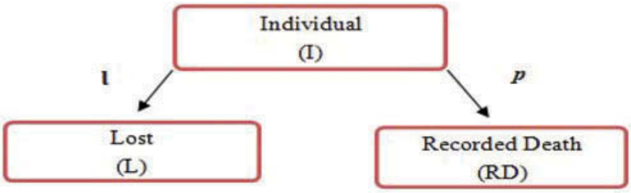

Figure 4. The Markov survival model without censoring.

$$\frac{P_L(t)}{P_D(t)} = \frac{l}{p}. \tag{12}$$

Let us denote the ratio l/p by α, which is the number of "lost" individuals per death. As $t \to \infty$ there are no more observed deaths, as all individuals who have not already died have in fact been "lost." We can obtain an estimate of α by the ratio of the number of individuals apparently still in the study n_∞ and the number who have been observed to die D_∞, yielding

$$\hat{\alpha} = \frac{n_\infty}{D_\infty}. \tag{13}$$

Thus, an easy way to construct a survival curve from the data is to adjust the number at risk, instead of using the formula from Eq. (6), and instead use

$$\hat{n}_{j+1} = \hat{n}_j - d_j - \hat{\alpha} d_j - c_j, \tag{14}$$

which implies that

$$\hat{n}_{j+1} = \hat{n}_j - \frac{D_\infty + n_\infty}{D_\infty} d_j - c_j. \tag{15}$$

We can see that using this updating method, $\hat{\alpha}$ from Eq. (13) is the appropriate estimate to use, by observing that after all observed deaths have been accounted for, using Eq. (15) an extra n_∞ individuals will have been removed from the at-risk category. This is precisely the number of individuals that we noted had been "lost."

The survivor function is then just calculated using Eq. (5), with h_j calculated using

$$h_j = \frac{d_j}{\hat{n}_j}. \tag{16}$$

This method will work well even if the underlying death rate and the rate of loss of individuals vary in time, as long as they vary in proportion with each other. If this is not the case, there would be a bias in the estimates of the hazard function h_j that we obtain. To tackle this problem we would need to have some more specific information about the way in which the loss of individuals into the L category differed from the rate of deaths, and this is likely to be problem specific, for example, depending upon hospital procedures, so we do not discuss any specific methodological ideas in this paper.

As mentioned in section 2, there is overt censoring in this population caused by the end of the study period. This creates a potentially significant problem, because even the "lost" individuals are censored in this way, so without adjustment the number of individuals at risk can be underestimated due to double counting (effectively, the same individual being lost and then censored can be removed twice). This in turn leads to a lower estimate of α than would otherwise be the case. In the alternative model that follows, we show a different, higher, estimate of α, and there is some discrepancy between the estimated survival curves of the two methods as a result, since the preceding errors cause an overestimate of the survival curve for the first model. We discuss this issue later (see Figures 6 and 7).

3.2. A Markov model with censoring

More generally, we would like to allow for observed censoring as well as hidden censoring within our model. Observed censoring occurs either when an individual is still alive at the end of the study period, or where the person leaves the study before the end but the hospital is aware of it. Hidden censoring occurs when the person leaves the study but the hospital is not aware of it. Thus, we now add an extra "censored" category C to our model, where individuals move from I to C at rate q. Importantly, individuals also move from the lost category L to C at the same rate q. This is clearly appropriate for our data set, since the only overt censoring is due to the end of the study, and thus any individual will reach this at the same time, whether in category I or L. We thus now have a population as described by Figure 5. We note that for individuals censored because we know that they have dropped out of the study prior to the end time, it would seem reasonable to assume that these and the "lost" individuals would be entirely separate, so that the transition rate q from state L to state C would be absent.

Following the transitions in Figure 5, there is a constant rate of departure per individual from state I, so that we have

$$\frac{d}{dt}P_I(t) = -(l+p+q)P_I(t), \tag{17}$$

and similarly to before, we obtain

$$P_I(t) = e^{-(l+p+q)t}. \tag{18}$$

We also still have

$$\frac{d}{dt}P_D(t) = pP_I(t), \tag{19}$$

which using Eq. (18) and the fact that $P_D(0) = 0$ yields

$$P_D(t) = \frac{p}{l+p+q}(1-e^{-(l+p+q)t}). \tag{20}$$

For the lost category L, we have entry to the state at rate $lP_I(t)$ and departure from the state at rate $qP_L(t)$, giving

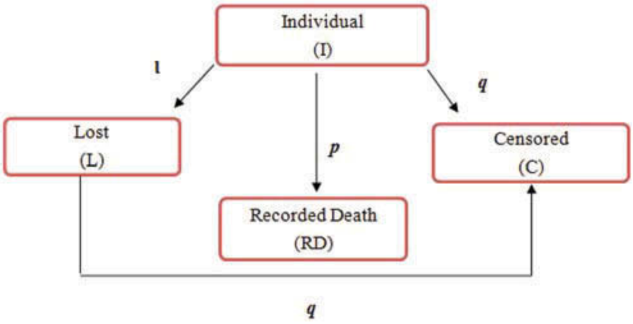

Figure 5. The Markov survival model with censoring.

$$\frac{d}{dt}P_L(t) = lP_I(t) - qP_L(t). \tag{21}$$

Using standard methods (see, e.g., Chap. 8 of Haigh 2002), together with the fact that $P_L(t) = 0$, yields

$$P_L(t) = e^{-qt}\frac{l}{l+p}\left(1 - e^{-(l+p)t}\right). \tag{22}$$

Finally, since $P_C = 1 - P_I - P_D - P_L$, we obtain

$$P_C(t) = \frac{l+q}{l+p+q}\left(1 - e^{-(l+p+q)t}\right) - e^{-qt}\frac{l}{l+p}\left(1 - e^{-(l+p)t}\right). \tag{23}$$

It is clear that the death rate for individuals in the at risk category *I*, that is, the correct hazard function, is simply

$$h_c(t) = p. \tag{24}$$

Using Eq. (3), the true survivor function for our model is thus simply

$$S_c(t) = e^{-pt}. \tag{25}$$

Since we perceive individuals from class L as being in category I too, we observe an apparent hazard function of

$$h_a(t) = \frac{P_I}{P_I + P_L}p. \tag{26}$$

Substituting the appropriate terms from Eqs. (18) and (22) and rearranging gives

$$h_a(t) = \frac{p(l+p)}{p + le^{(l+p)t}}. \tag{27}$$

Following Eq. (3), the apparent survivor function is thus

$$S_a(t) = e^{-\int_0^t h_a(u)du}, \tag{28}$$

which rearranges to

$$S_a(t) = \frac{l + pe^{-(l+p)t}}{l+p} = \frac{\alpha + (e^{-pt})^{1+\alpha}}{1+\alpha}. \tag{29}$$

Thus, we can express $S_c(t)$ in terms of $S_a(t)$ as follows:

$$S_c(t) = ((1+\alpha)S_a(t) - \alpha)^{1/(1+\alpha)}. \tag{30}$$

The apparent survival curve $S_a(t)$ from Eq. (29) flattens out to a limiting value $\alpha/(1+\alpha)$. We can thus estimate α by equating this theoretical limit with the observed limiting value of the survival curve from the data which we shall denote by s_∞, giving

$$\tilde{\alpha} = \frac{s_\infty}{1 - s_\infty}. \tag{31}$$

This yields the conversion formula from the apparent to the corrected survival curve as

$$S_c(t) = \left(\frac{1}{1-s_\infty} S_a(t) - \frac{s_\infty}{1-s_\infty}\right)^{1-s_\infty}. \tag{32}$$

We note that our final solutions for the survivor function and the hazard function do not contain q at all. In fact, this means that these solutions are unaffected if q is replaced by a time-dependent function $q(t)$. This is important, as the censoring time is in reality directly related to the rate of recruitment into the study, which may be influenced by nonrandom factors. It also means that we can apply this method to cases without censoring as in the previous method of section 3.1. We also note the discrepency between our two estimates of α. In general when overt censoring occurs, $\hat{\alpha}$ will be smaller than $\tilde{\alpha}$, because the first model neglects the influence of censoring in the estimation procedure.

4. Applying our models to the Kurdish data

From the Kurdish data we obtained the following values: $n_\infty = 232$, $D_\infty = 240$, and $s_\infty = 0.5969$, which gives the two alternative estimates of $\hat{\alpha} = 0.9667$ and $\tilde{\alpha} = 1.4807$ lost individuals per death. Applying our method from section 3.1 gives the adjusted survival curve from Figure 6. Using the alternative method from section 3.2 gives the adjusted survival curve from Figure 7.

We can see that the two alternative survival curves generated by our methods now resemble the survival curve from the German data from Figure 3. The curves in our case are clearly lower than that of the German data, indicating poorer survival rates among the Kurdish patients. There are a number of reasons for this, including later diagnosis, less efficient treatment regimes, and different patient demographics, which we will consider in a later paper. Comparing the two curves from Figures 6 and 7, we see that initially the two curves are roughly the same, but for later times, the curve in Figure 6 is clearly above that in Figure 7. We should also note that our methods are likely not to be very accurate near

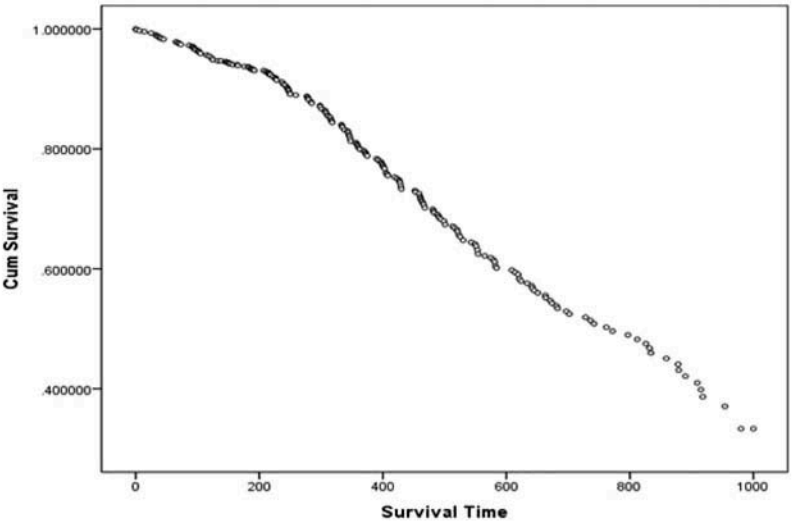

Figure 6. An adjusted survival curve for the Nanakaly data using the method without censoring from section 3.1.

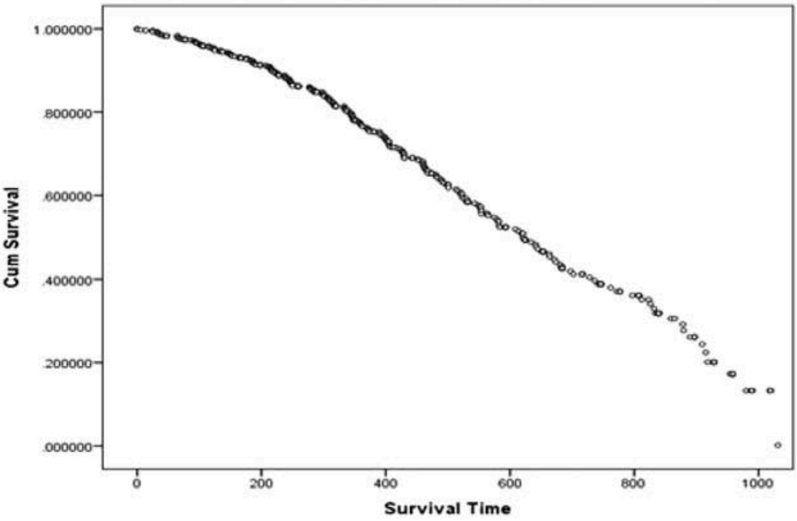

Figure 7. An adjusted survival curve for the Nanakaly data using the method with censoring from section 3.2.

the end of the curves, that is, when the last of the recorded deaths occur. Thus, in the case of the Nanakaly data, the curves beyond about 700 days are likely to be inaccurate.

We should also note that the preceding methodology can be applied to any survival curve, so that if we have a survival curve from a subset of the data, or for patients with particular properties, then the method of adjusting the original survival curve is completely unchanged.

5. Simulations

In this section we consider simulations to investigate the validity of our modelling procedure. We consider the example German data from Figure 3, as we have an accurate survival function for this. For each simulation, we chose a distribution and simulated each individual from the German data being “lost” following this distribution. Thus, if death happens before the individual is lost, we observe the death, but if the individual is lost first we assume that they are still in the study, and we do not observe their death, if it occurs. This thus replicates what happens in the Nanakaly data, and the situation that we are modeling.

The models that we have considered are Markov with constant rate, which would yield an exponentially distributed time of loss. We considered various values of this distribution. One set of simulations considered a mean loss time of 2000 days. Given the length of the German study, this accounted for quite a significant loss of data. One example run of this is shown in Figure 8, where the apparent survival probability after 2000 days has only fallen to approximately 0.8 instead of the true value of just over 0.6 as a result. The survival curves generated for our two models are shown in Figures 9 and 10, respectively. We can see that in both cases, the models significantly correct the survival function from the apparent survival function shown in Figure 8. The first model gives a somewhat conservative correction, which is higher than the true survival function in Figure 3. As

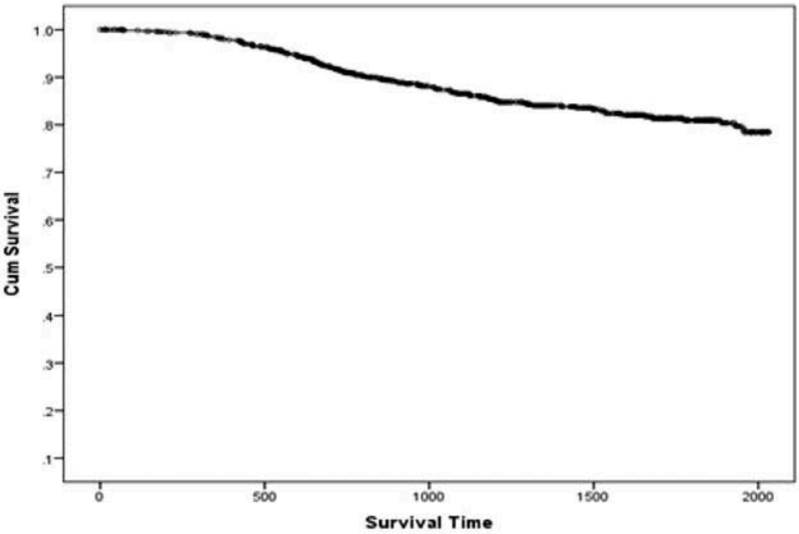

Figure 8. The survival curve for a sample simulation of loss from the German data, where loss of individuals occurs following an exponential time with mean 2000 days.

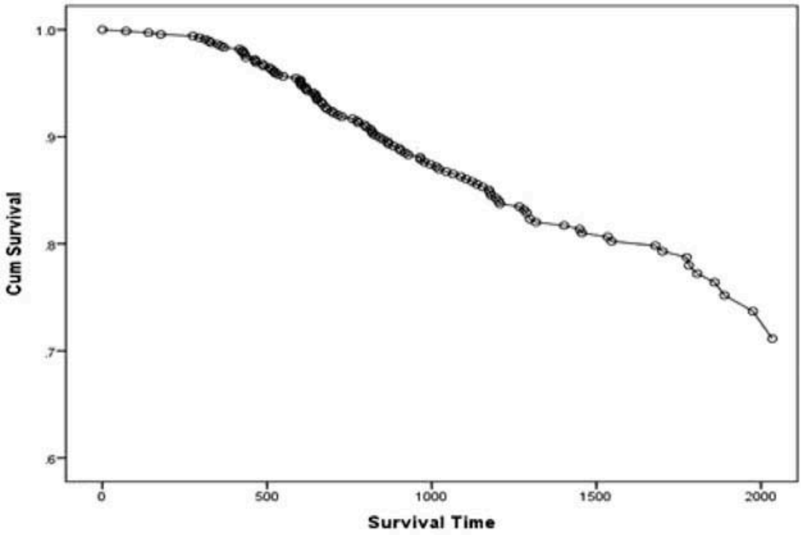

Figure 9. An adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 2000 days, using the method without censoring from section 3.1.

explained in the final paragraph of section 3.1, this is because of the double counting of lost and censored individuals. The estimate of the second model is clearly better, with a closely comparable survival curve. These curves are typical of different simulations with the same mean loss time.

Exponential distributions with higher means yield even better results, as the level of loss is diminished, so the amount of adjustment that needs to be carried out through our procedure is reduced. For exponential distributions with lower means (in particular below

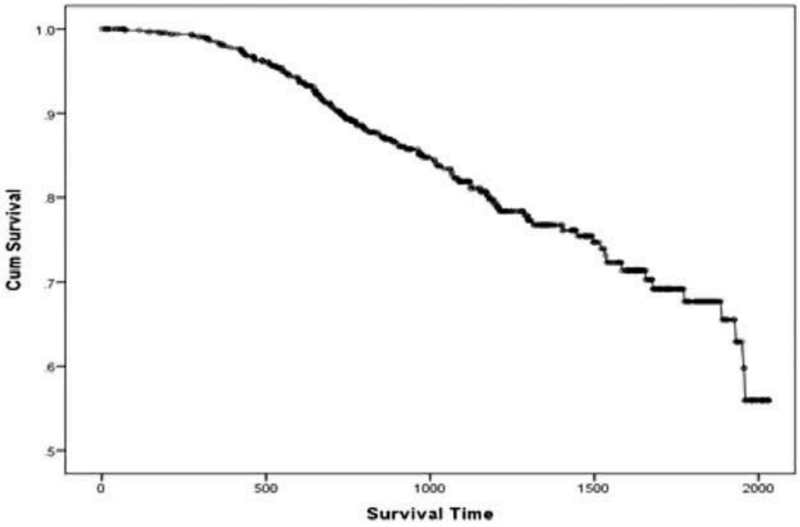

Figure 10. An adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 2000 using the method with censoring from section 3.2.

500 days), and thus very large data loss, figures became increasingly less accurate, as a larger proportion of deaths was missed, and this led to overestimates of survival rates.

We also considered nonexponential distributions, for example, Gamma$(2, \theta)$, for patient loss. When this led to a large number of lost individuals (for sufficiently high means this did not, and thus as earlier the corrections were not large and were accurate), we would expect our model to perform worse in such circumstances, as this would indicate that the underlying Markov assumption was not correct. This was indeed the case, although the models still corrected the false apparent curves to a significant effect, and as for exponential distributions with small means, the effect was generally to produce slightly conservative survival functions, which overestimated the true survival curve.

Thus, we see that our models perform well in many circumstances, and even when less accurate they are always an improvement on considering the apparent survival curves from the unadjusted data.

6. Discussion

In this article we have developed a new method for performing a survival analysis on a set of data where there are important unknown factors, namely, secret censoring of the data, so that the number of individuals apparently at risk is greater than those actually at risk. In particular, we have shown how to adjust a Kaplan–Meier analysis to find a survival curve in such circumstances, and also shown how to estimate a true hazard (survivor) function to the biased one obtained directly from the data.

A limitation of our methodology is that it is based upon a Markov chain, so transition rates are assumed constant over time. In fact, some relaxation of these assumptions, such as making the censoring rate q time dependent, does not affect the model accuracy. Similarly, allowing the parameters l and p to be time dependent does not affect the model, provided that they vary with $\alpha = l/p$ constant. This, however, is not always reasonable, and it is possible to envisage some situations where this is far from being the case. In such circumstances our predictions would not be reliable. Similarly, if different groups of individuals have different rates with different l/p ratios, this might also affect the results. We claim, however, that in circumstances where the problems outlined occur, our model is a good first step, and a considerable improvement on making no adjustment. This is demonstrated in section 5, where we simulated the loss of individuals from the German data set and compared the resulting survival curves from our models with those from the original data.

This leads to the question, how prevalent will the problems that we have described be? With sufficiently accurate records and follow-up of individuals they will not occur, and of course a better solution than applying our methods is to have these processes in place. Nevertheless, in reality they often will not be in place. This is particularly the case in regions with a history of upheaval and developing medical services. It can be argued that these are precisely the regions that most need accurate survival models, so the application of our methods can be of significant value.

Funding

Mahdi Saber Raza was supported by a scholarship under the Human Capacity Development Program (HCDP) sponsored by the Kurdistan Regional Government (KRG). This program is implemented and executed by the Ministry of Higher Education and Scientific Research HCDP-R1.No.1063.

References

Al Tamimi, D., A. Mohamed, A. Ayesha, K. Ammar, and A. Amal. 2010. Portion expression profile and prevalence pattern of the molecular classes of breast cancer—A Saudi population based study. *BioMed Central Cancer* 10 (223):1–13.

Alwan, N. A., W. Al-Kubaisy, and K. Al-Rawaq. 2000. Assessment of response to tamoxifen among Iraqi patients with advanced breast cancer. *East Mediterranean Health Journal* 6:475–82.

Barlow, R., and F. Proschan. 1975. *Statistical theory of reliability and life testing probability models*. Austin, TX: Holt, Rinehart and Winston.

Bedford, T., and R. Cook. 2009. *Probabilistic risk analysis foundation and methods*. New York, NY: Cambridge University Press.

Cox, R., and H. Miller. 1965. *The theory of stochastic processes*. London, UK: Methuen & Co.

Cox, R., and D. Oakes. 1984. *Analysis of survival data*. London, UK: Chapman and Hall.

Crowder, M. 2012. *Multivariate survival analysis and computing risks*. New York, NY: CRC Press.

Crowder, M., A. Kimber, R. Smith, and T. Sweeting. 1991. *Statistical analysis of reliability data*. London, UK: Chapman and Hall.

De Santis, C., J. Ma, L. Bryan, and A. Jemal. 2014. Breast cancer statistics, 2013. *CAA Cancer Journal for Clinicians* 64:52–62.

Dey, S., A. S. Soliman, A. Hablas, I. A. Seifeldin, K. Ismail, M. Ramadan, H. El-Hamzawy, M. L. Wilson, M. Banerjee, P. Boffetta, J. Harford, and S. D. Merajver. 2010. Urban–rural differences in breast cancer incidence by hormone receptor status across 6 years in Egypt. *Breast Cancer Research and Treatment* 120:149–160.

Grimmett, G., and D. Stirzaker. 2001. *Probability and random processes*, 3rd ed. New York, NY: Oxford University Press.

Haigh, J. 2002. *Probability models*. London, UK: Springer.

Hughson, M. D. 2012. A population-based study of Kurdish breast cancer in northern Iraq: Hormone receptor and HER2 status. A comparison with Arabic women and United States SEER data. *BMC Women's Health* 12 (16):1–10.

Hussaion, A. H., and P. M. Aziz. 2009. The incidence rate of breast cancer in Suleimani Governorate in 2006: Preliminary study. *Journal of Zankoy Suleimani* 12(1, Part A):59–65.

Lan, N. H., W. Laohasiriwong, and J. Stewart. 2013. Survival probability and prognostic factors for breast cancer patients in Vietnam. *Global Health Action*, 6:18860.

Lawless, J. F. 2003. *Statistical models and methods for lifetime data*, 2nd ed. Hoboken, NJ: John Wiley & Sons.

Majid, R. A., H. A. Mohammed, H. M. Saeed, B. M. Safar, R. M. Rashid, and M. D. Hughson. 2009. Breast cancer in Kurdish women of northern Iraq: Incidence, clinical stage, and case control analysis of parity and family risk. *BMC Women's Health* 9 (33).

Majid, R. A., H. A. Mohammed, H. A. Hassan, W. A. Abdulmahdi, R. M. Rashid, and M. D. Hughson. 2012. A population-based study of Kurdish breast cancer in northern Iraq: Hormone receptor and HER2 status. A comparison with Arabic women and United States SEER data. *BMC Women's Health* 12 (16):1–10.

NHS Choices. 2011. Unhealthy lifestyles linked to UK cancer rates. http://www.nhs.uk/news/2011/01January/Pages/unhealthy-lifestyles-linked-to-UK-cancer-rates.aspx

Othman, R. T., R. Abdulljabar, A. Saeed, S. S. Kittani, H. M. Sulaiman, S. A. Mohammed, R. M. Rashid, and Hussein, N. R. 2011. Cancer incidence rates in the Kurdistan region/Iraq from 2007–2009. *Asian Pacific Journal of Cancer Prevention* 12 (5):1261–64.

Ozmen, V. 2006. Screening and registering programs for breast cancer in Turkey and in the world. *Journal of Breast Health* 2 (2):55–58.

Rennert, G. 2006. Breast cancer. In *Cancer incidence in the four member countries (Cyprus, Egypt, Israel, and Jordan) of the Middle-East Cancer Consortium (MECC) compared with US SEER*, Ed. L. S. Friedman, B. K. Edwards, L. A. G. Reiss, J. L. Young, 73–81. Bethesda, MD: National Cancer Institute, NIH Pub No. 06-5873.

Ries, L. A. G., D. Melbert, M. Krapcho, A. Mariotto, B. A. Miller, E. J. Feuer, L. Clegg, M. J. Horner, N. Howlader, M. P. Eisner, M. Reichman, and B. K. Edwards. eds. 2007. *SEER cancer statistics review, 1975–2004*. National Cancer Institute.

Robb, C., W. E. Haley, L. Balducci, M. Extermann, E. A. Perkins, B. J. Small, and J. Mortimer. 2007. Impact of breast cancer survivorship on quality of life in older women. *Critical Reviews in Oncology/Hematology* 62 (1):84–91.

Rudat, V., B. Nuha, T. Saleh, and A. Mousa. 2013. Body mass index and breast cancer risk: A retrospective multi-institutional analysis in Saudi Arabia. *Advances in Breast Cancer Research* 2:7–10.

Schmoor, C., M. Olschewski, and S. Martin. 1996. Randomized and non-randomized patients in clinical trials: Experiences with comprehensive cohort studies. *Statistics in Medicine* 15:263–71.

Schumacher, M., G. Bastert, H. Bojar, K. Hubner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. A. Neumann, and H. F. Rauschecker. 1994. Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology* 12 (10):2086–93.

Shabila, N., G. Namir, N. Al-Tawil, S. Tariq, T. Al-Hadithi, S. Egbert, and V. Kelsey. 2012. Iraqi primary care system in Kurdistan region: Providers perspectives on problems and opportunities for improvement. *BioMed Central Womens Health* 12 (21):1–9.

Sughayer, M. A., M. A. Maha, M. Suleiman, and A. Mahmoud. 2006. Prevalence of hormone receptors and HER2/neu in breast cancer cases in Jordan. *Pathology Oncology Research* 12 (2):83–86.

World Health Organization. 2008. The global burden of disease: 2004 Update. www.who.int/evidence/bod.