Asian Journal of Probability and Statistics

Volume 25, Issue 3, Page 45-69, 2023; Article no.AJPAS.108897
ISSN: 2582-0230

The Use of Survival Analysis Modelling with Incomplete Data with Application to Breast Cancer

Mahdi Saber Raza [a] and Mark Broom [b*]

[a] Department of Software and Informatics Engineering, College of Engineering, University of Salahaddin, Erbil, Iraq.
[b] Department of Mathematics, City University London, Northampton Square, London EC1V 0HB, UK.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2023/v25i3563

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/108897

Original Research Article

Received: 13/09/2023
Accepted: 17/11/2023
Published: 24/11/2023

Abstract

There are strong survival analysis methodologies for data sets which are complete, with accurate information on censoring. But what if they are not complete? In an earlier paper we built a methodology for estimating survival probabilities and hazard functions in a health setting, using breast cancer data from the Kurdistan region of Iraq, for censored and uncensored data when a substantial portion of individuals are lost to the study. In this paper we build on these models to consider further issues based upon the accuracy of the records of patient death, where deaths often occur beyond the hospital in family settings and patients ceasing treatment and contact with the hospital may or may not represent their death; thus the record of their time of death may not be accurate. We develop a new Markov chain-based methodology for generating survival curves and hazard functions, and demonstrate this using a different breast cancer dataset from the Kurdistan region of Iraq.

**Corresponding author: Email: Mark.Broom@city.ac.uk;*

Keywords: Survival analysis; Markov model; breast cancer; censoring data.

1 Introduction

Survival analysis is primarily concerned with modelling and analysing time-to-event data, with events generally referred to as "failures." Some examples are time until an electrical component fails, time to first recurrence of a tumour (i.e., length of remission) after initial treatment [1].

It is possible that a "failure" time will not be observed due to deliberate design or random censoring. In this study this would occur if a patient is still alive at the end of a clinical trial period or has moved away [2,3]. The primary reason for developing specialized models and procedures for failure time data is the necessity of obtaining methods of analysis that accommodate censoring. Survival analysis can then be thought of as a collection of statistical procedures that accommodate time-to-event censored data. Previously, incomplete data were treated as missing data and omitted. This loss of information introduced bias in estimated quantities. The procedures discussed here avoid bias and are more powerful as they utilize the partial information available on a subject or item [1]. Survival analysis is the study of the occurrence and timing of events. Covariates are studied to determine their effect on survival duration. Censoring and time-dependent covariates (time-varying explanatory variables) are central to survival analysis [4,5].

The survival function is the most well-known function in survival analysis; it describes the probability that an individual survives up to time *t*. Related to this is the hazard function, which is the risk of death (per unit time) [6,7]. If our individual has lifetime distribution T, following standard terminology (see Raza and Broom, 2016) the survival function is

$$S(t) = 1 - F(t) = P\{T > t\}. \tag{1}$$

The hazard rate or hazard function is

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{2}$$

i.e.

$$h(t)dt = P\{t < T < t + dt \mid T > t\} \tag{3}$$

$=P$ {death in the interval (t, $t+dt$) given survival past time t}.

Integrating $h(t)$,

$$\int_0^t h(u)du = \int_0^t \frac{f(u)}{1 - F(u)} du = -\log(1 - F(u))\Big|_0^t$$
$$= -\log(1 - F(t)) = -\log S(t),$$

which leads to the important expression

$$\log S(t) = -\int_0^t h(u)du. \tag{4}$$

Notice that F(+∞) = 1 (i.e., S(+∞) =0) iff $\int_0^\infty h(u)du = \infty$.

Continuity will be assumed but concepts and formulae can be modified to include jumps in the density function when it is important [8]. 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 [9,10].

$$h_j = \frac{d_j}{n_j}. \tag{5}$$

Here 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

$$\hat{S}(t) = \prod_{j=1}^{t}(1-h_j). \tag{6}$$

This method automatically takes into account any data censoring, where an individual is known to leave the study at a particular time; in Raza and Broom [8] we reduced the number at risk n by the number of censored individual's c_j. Thus, the updated number at risk was as follows:

$$n_{j+1} = n_j - d_j - c_j. \tag{7}$$

A great advantage of this method is that reliable estimates can be obtained without making assumptions about the underlying distribution of estimated times [11]. 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, i.e. all values 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 described above without problems, provided that records are sufficiently good to know when contact with patients has been lost. In Raza and Broom [8] we considered situations where we do not have this knowledge, and significant "hidden" censoring occurs unknown to the researchers, using a real example as a case study (Nanakaly Hospital data). There we presented two models, one without and one with censoring, which addressed 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. For details of these models see Figs. 5 and 6. Nevertheless the models, in particular the simpler first model, were robust to (at least certain types of) deviation from such assumptions [12,13].

There are numerous studies on survival analysis in different countries, on various types of diseases. However, there are hardly any survival analyses studies performed on breast cancer by Iraqi researchers in general and researchers in the Kurdistan Region in particular. Due to the importance of this for both society and the individuals [14,15], in this paper we look at the models described above (with and without censoring) based on an updated structure applied to data from Hewa Hospital. This accounts for the missing reports on times of death of a number of patients who did not return for their follow up appointments, a phenomenon which is indicative of a larger problem with the patients' reaction toward this particular diagnosis which may be due to poor health education and general lack of awareness of the importance of keeping detailed and complete hospital records. The reason may include a general fear of disease or death and the hope of receiving better treatment elsewhere. Hospital record consistencies and general compliance appears to be correlated to economic status and doctor-patient interaction, and also the apparent termination of a patient's follow up treatment may simply be due to bad bookkeeping [16]. In particular it is clear that actual death records are almost entirely absent in the data that was obtained. That is the basic problem leading to not having the real time of death.

Detailed times of death were provided, with censoring only at the end of the study period on 1st June 2014, see Fig. 1 for an illustration. Analysing the above data using SPSS provided the Kaplan-Meier survival curve in Fig. 2. The function flattens out to effectively a horizontal line, indicating a hazard rate tending to zero. In particular, Fig. 3 shows that the probability of death appears very small after 700 days as the curve flatters out around this

time [17]. Since we see a small number of patients remaining for several thousand days, the data after this period was removed after time 659 as in Fig. 3.

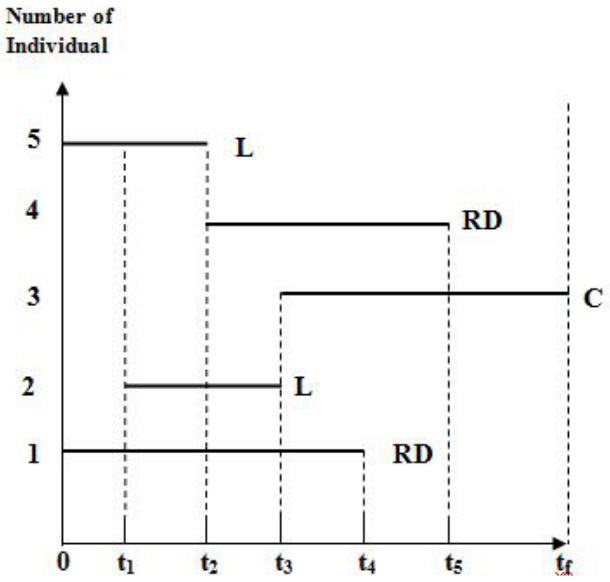

Fig. 1. Illustrative plot of survival times including end of period censoring (C), recorded death (RD) and hidden censoring, individuals unknowingly lost to the study (L), for the Kurdish data from Hewa Hospital

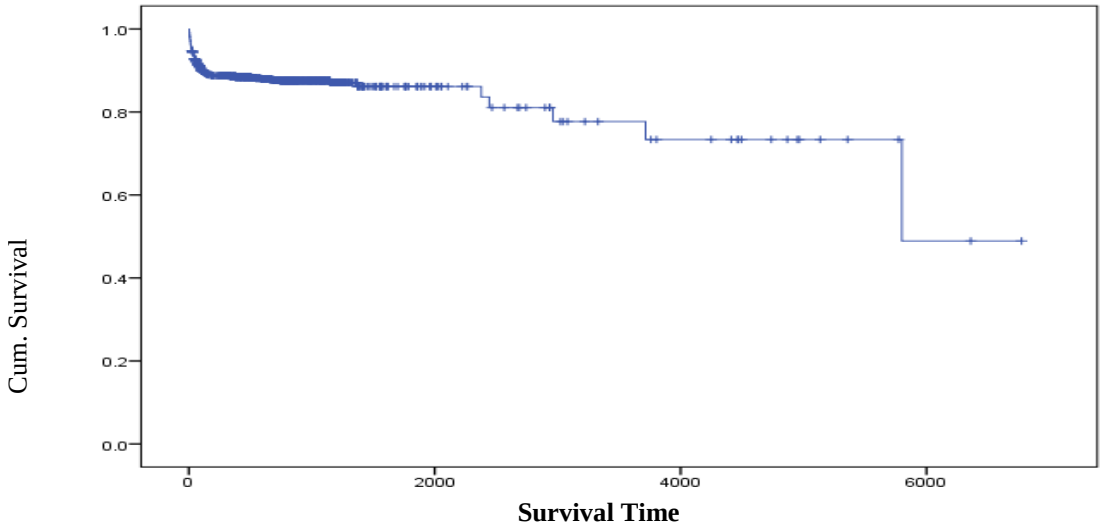

Fig. 2. Survival curve including censoring for the Kurdish data from Hewa hospital

The survival curve in Fig. 3 is clearly not realistic. For comparison, a survival curve for a set of breast cancer data from Schumacher et al. [18] is shown in Fig. 4. The problem with the survival curve from Fig. 3 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 to the hospital after initial treatment, and there are no clear records of when the deaths of these individuals occurred, or of which individuals these were. Thus, there is some secret censoring that we do not have knowledge about. We can think of this to mean that whilst the values of d_j are accurate, the values of n_j are not, and we are (after some time, greatly) overestimating them. As in the Nanakaly data in Raza and Broom [8], the flatness of the curve is likely because only a small number of patients remain in the study after this time. In summary, to find a good

model for the Hewa Hospital data we began by plotting the survival curve using the Kaplan Meier method and it shows that, the curve is not reliable when compared to the Kaplan Meier curve for the German data model.

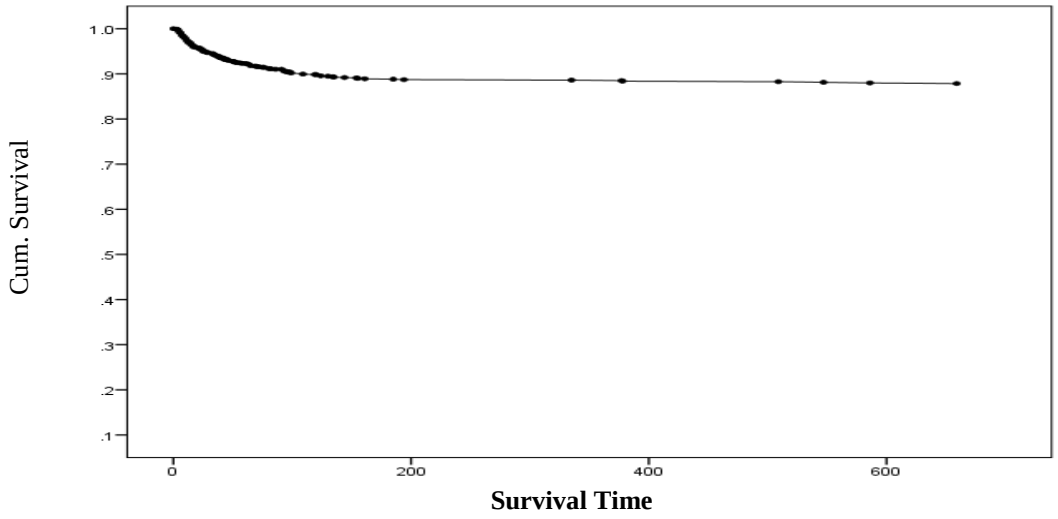

Fig. 3. Survival curve including censoring for the Kurdish data from Hewa hospital data for 700 days

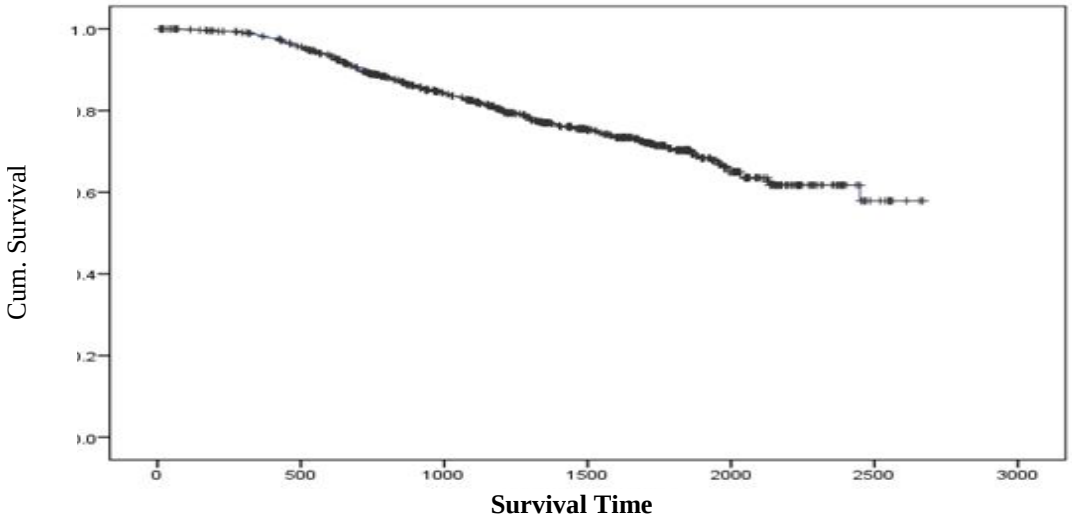

Fig. 4. Survival curve for the German data from Schmoor et al. (1996) and Schumacher et al. (1994).

2 Markov Model

2.1 Markov model without censoring for Hewa data

In general, the process of collecting data in the health sector or any other sector in a developing country such as Iraq is not easy, because there is no accurate database system. The most dependable data are available in the official records but not obtainable readily. One inevitably has to refer to numerous government agencies to obtain relevant information from official sources such as the Ministry of Health, especially for information regarding the time of death.

The Hewa data includes general information about the breast cancer patients, including their age, religion, tumour size, tumour grade, lymph nodes, exercise, educational level, family history, breast feeding, smoking, drinking alcohol, occupation, progesterone, estrogen, menopause, marital status and income. In addition there

are three times involved in this data, time of admission, time of diagnosis and time of death. The time of admission refers to the first time when the patient visits the hospital and the staff of the hospital administration register general information about her. The diagnosis time means the time when the doctor diagnoses the patient and refers them to the laboratory to make necessary required tests based on the symptoms that they are suffering from. Finally, there is the actual time of death of the patient.

In initial analysis later in this study we use *z* as an intermediate measurement for the rate of real death (which is not given) from the time of diagnosis. This serves to account for the missing reports on times of death of a number of patients who did not return for the follow up appointments, leading to the hospital not having the real time of death, which is reflected in the results of the survival curve when we applied the program at the first step, because we used the time of admission instead of the time of death. Clearly there are differences between the time of admission and the time of death, and this leads to significant problems with the analysis. That is why we extend the analysis of the survival curve using a Markov process. This is the most natural and simple extension to the model to try to deal with the absence of the actual time of death. Here when we applied the Markov process we considered two modifications, one without censoring and one with censoring [19].

Nevertheless, there are some serious limitations to this model. For example, if individuals do not get censored at constant rate, the censoring time will follow another distribution rather than an exponential distribution, which arises from the Markov model, and we will obtain a different picture to that obtained from a Markov process. Potentially more significant problems result from the lack of knowledge of the times of death. Deaths are assumed to follow a Markov process from the time of diagnosis category and there are two main source of error. Firstly, the rate (z) of this process is unknown and had to be estimated and we have thus considered a range of values. Secondly, again this may not be a Markov process, which would also affect the shape of the survival function.

We addressed this problem through estimating these numbers by constructing two new models, as described above, each using Markov chains and estimating the number at risk $\tilde{n}$ and deaths $\tilde{d}$. Fig. 5 represents the original structure Markov survival model with no censoring.

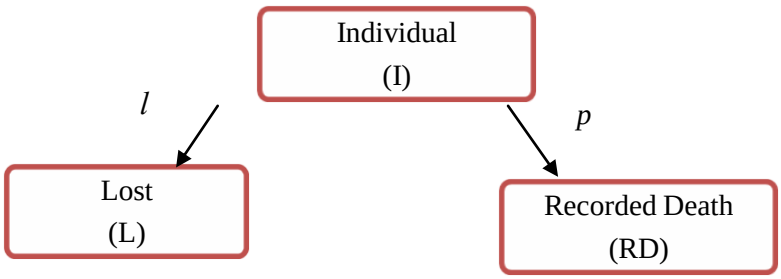

Fig. 5. The Markov survival model without censoring for Nanakaly data

Consider a population of individuals in three categories; either at risk (I), died (RD) or who have left the study (without our knowledge), which we shall call "lost" (L). Individuals simply move from state (I) to the other two states at constant rates (l) to (L) and (p) to (RD). We thus have a population as described by Fig. 5. Denoting the proportion of individuals in states (I), (L) and (RD) at time t by $P_I(t)$, $P_L(t)$ and $P_{RD}(t)$ respectively.

More generally we need to allow for observed censoring as well as hidden censoring within our model. 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 dataset, 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 then have a population as described by Fig. 6. 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, and so that the transition rate *q* from state (L) to state (C) would be absent.

Fig. 7 shows the first model for the Hewa data; here we have four stages, three of them are the same as for the first model of the Nanakaly data plus on extra stage from recorded death (where we use the admission time as a proxy for recorded death) to death (z). Here the same conditions and protocol will be required for the continuous-time Markov chain as applies in the Nanakaly data Model I, but for the different sample space;

$$\Omega = [I, L, RD, D].$$

The difference between the time of admission (RD) and the time of death (D) occurs because when the patient leaves the study, this may not indicate their death, but simply that they are possibly lost in the study or they have not been followed up, as is shown in the transition stages in Fig. 6 with censoring. However, the term (D) represents the patients that are actually dead in the study, as shown in Fig. 7 without censoring.

.

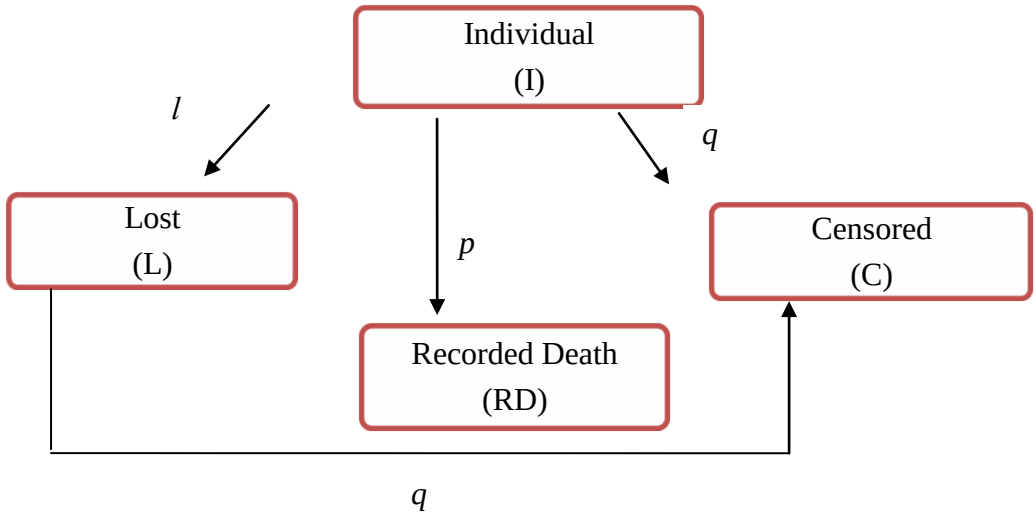

Fig. 6. The Markov survival model with censoring for Nanakaly data

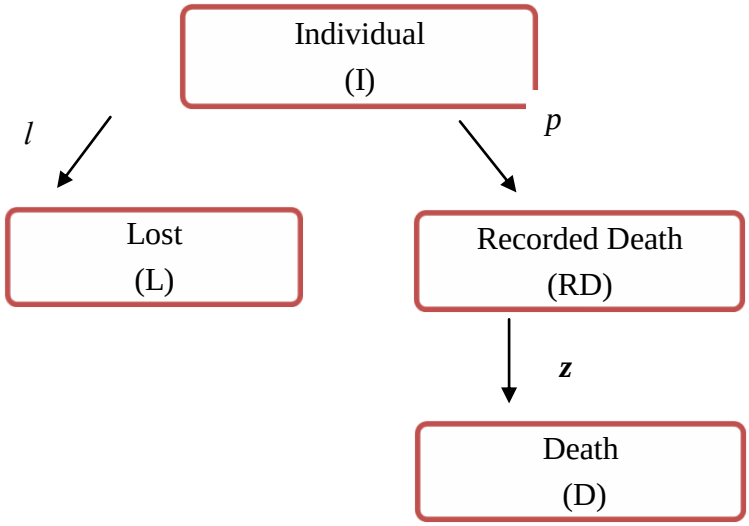

Fig. 7. The Markov survival model without censoring for Hewa data

The state probabilities at time t, are represented by the following:

$$P = \begin{pmatrix} P_I(t) & P_L(t) & P_{RD}(t) & P_D(t) \end{pmatrix}. \tag{8}$$

They depend continuously on time with constant transition rate represented by:

$$P' = \frac{dP}{dt} = PQ \tag{9}$$

for a given transition matrix $Q = (q_{is})$, where q_{is} is the rate of flow from ($i \to s$), which is a |P| * |P| matrix of transition rates if it fulfils the following two conditions:

a) *Q* has no negative off-diagonal entries, i.e. $q_{is} \geq 0$ for all $i \neq s$.

b) *Q* has row sums equal to zero, or $\sum_s q_{is} = 0$ for all i.

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1s} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2s} & \cdots & q_{2n} \\ . & . & \cdots & . & \cdots & . \\ q_{i1} & q_{i2} & \cdots & q_{is} & \cdots & q_{in} \\ . & . & \cdots & . & \cdots & . \\ q_{n1} & q_{n2} & \cdots & q_{ns} & \cdots & q_{nn} \end{pmatrix} \tag{10}$$

The solution of (9) identifies the following equation for *P*:

$$p = p_0 e^{tQ} \tag{11}$$

subject to initial conditions $P_I(0) = P_0$. The transition rate matrix Q for Figure 7 is given by:

$$Q = \begin{pmatrix} -(l+p) & l & p & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -z & z \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{12}$$

We then have

$$\begin{pmatrix} \frac{d}{dt}P_I(t) & \frac{d}{dt}P_L(t) & \frac{d}{dt}P_{RD}(t) & \frac{d}{dt}P_D(t) \end{pmatrix} = \begin{pmatrix} P_I(t) & P_L(t) & P_{RD}(t) & P_D(t) \end{pmatrix} \begin{pmatrix} -(l+p) & l & p & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -z & z \\ 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow$$

$$\begin{pmatrix} \frac{d}{dt}P_I(t) & \frac{d}{dt}P_L(t) & \frac{d}{dt}P_{RD}(t) & \frac{d}{dt}P_D(t) \end{pmatrix} =$$

$$\begin{pmatrix} -(l+p)P_I(t) & l\,P_I(t) & p\,P_I(t) - z\,P_{RD}(t) & z\,P_{RD}(t) \end{pmatrix}. \tag{13}$$

The transition rate out of state I is the same as in Model I from Raza and Broom [8], i.e. as in equation (14),

$$\frac{d}{dt}P_I(t) = 0 - (l+p)P_I(t) \Rightarrow P_I(t) = k_1 e^{-(l+p)t} = e^{-(l+p)t}. \quad (14)$$

Similarly the equation, initial condition and solution for (L) are identical to before, i.e. $P_L(t) = \frac{l}{l+p}(1-e^{-(l+p)t})$.

For state (RD) the flow in equals $p\,P_I(t)$ and flow out equals $z\,P_{RD}(t)$.

$$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - z\,P_{RD}(t) \Rightarrow$$
$$\frac{d}{dt}P_{RD}(t) + z\,P_{RD}(t) = p\,P_I(t)$$

.

Multiplying both sides by the integrating factor e^{zt} we get

$$(\frac{d}{dt}P_{RD}(t) + z\,P_{RD}(t))e^{zt} = p\,P_I(t)e^{zt} \Rightarrow$$
$$\frac{d}{dt}(e^{zt}\,P_{RD}(t)) = p\,P_I(t)e^{zt}$$
$$= p\,e^{zt}\,e^{-(l+p)t} \Rightarrow e^{zt}\,P_{RD}(t) = \int P\,e^{(z-l-p)t}dt \Rightarrow$$
$$e^{zt}\,P_{RD}(t) = \frac{p}{z-l-p}\,e^{(z-l-p)t} + k_5. \quad (15)$$

Dividing both sides of equation (15) by e^{zt} we get the following equation 16.

$$P_{RD}(t) = k_5\,e^{-zt} + \frac{p}{z-l-p}\,e^{-(l+p)t}. \quad (16)$$

To find the value of constant k_5 we use the fact that at time zero $P_{RD}(0) = 0$ which leads to

$$P_{RD}(t) = \frac{p}{l+p-z}[e^{-zt} - e^{-(l+p)t}]. \quad (17)$$

As already mentioned, the transition rate from state (RD) to state (D) is (z), and so the flow into state (D) is equal to $z\,P_{RD}(t)$ and the flow out is equal to zero. Thus

$$\frac{d}{dt}P_D(t) = z\,P_{RD}(t) - 0. \quad (18)$$

By substituting equation (17) into the above formula we get

$\frac{d}{dt}P_D(t) = z\frac{p}{l+p-z}[e^{-zt} - e^{-(l+p)t}]$. If we let $w = \frac{p}{l+p-z}$ and then integrate we obtain equation (19) below:

$$P_D(t) = -we^{-zt} + \frac{zw}{l+p}e^{-(l+p)t} + k_6 . \tag{19}$$

When $t = 0 \Rightarrow P_D(0) = 0$ then we obtain

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

Table 1 shows a summary of the above solutions:

Table 1. Summary of all state transitions for model I Hewa data

State	In flow	Out flow	Equations for state probabilities	Probability values
I	0	$(l+p)P_I(t)$	$\frac{d}{dt}P_I(t) = 0 - (l+p)P_I(t)$	$P_I(t) = e^{-(l+p)t}$
L	$l\,P_I(t)$	0	$\frac{d}{dt}P_L(t) = l\,P_I(t) - 0$	$P_L(t) = \frac{l}{l+p}(1-e^{-(l+p)t})$
RD	$p\,P_I(t)$	$z\,P_{RD}(t)$	$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - z\frac{p}{l+p-z}(e^{-zt} - e^{-(l+p)t})$	$P_{RD}(t) = \frac{p}{l+p-z}[e^{-zt} - e^{-(l+p)t}]$
D	$z\,P_{RD}(t)$	0	$\frac{d}{dt}P_D(t) = z\frac{p}{l+p-z}(e^{-zt} - e^{-(l+p)t}) - 0$	$P_D(t) = \frac{p}{l+p-z}[(1-e^{-zt}) - \frac{z}{(l+p)}(1-e^{-(l+p)}$

The results in Table 1 are helpful in setting up the second Hewa model (now including censoring), but these should be applied using the same mathematical formalism as for the first model (which does not account for censoring).

We can derive, from the Markov process (see Fig. 7), estimates of the number of individuals moving from state (I) to (RD) and then (D). In the calculation below we shall use the following terms:

$\tilde{d}_t$ is the estimated number of real deaths, while d_t is the number of recorded deaths, and x_t is the probability of death of an individual in the (RD) category, all within the t th time interval starting at T_t and ending at T_{t+1}.

In time interval T_t to T_{t+1}, d_t individuals move from (I) to (RD) to (D). Here, recorded death (RD) is the transaction from individuals (I) to recorded death (RD) at rate p; recall that this does not mean that all of the individuals are dead, but some of them may be lost to the study. However, death (D) means that the individuals really have died, and this is assumed to happen at the start of the interval. Individuals in the record death category (RD) can then move to death (D) by rate z, at the start of the subsequent time interval, which they do with probability x_t .Thus the number of individuals remaining in state (RD) at the end of the time interval is

$\sum_{i=1}^{t} d_i \prod_{j=i}^{i-1}(1-x_j)$ and, so to estimate the number of deaths for Hewa patients, the following equation is used:

$$\tilde{d}_{t+1} = x_t \sum_{i=1}^{t} d_i \prod_{j=i}^{t-1}(1-x_j) , \tag{21}$$

where

$$\tilde{d}_1 = 0 \;\; , \;\; \tilde{d}_2 = x_1 d_1 \;\; , \;\; \tilde{d}_3 = x_2 d_2 + x_2(1-x_1)d_1 \tag{22}$$

and

$$x_t = 1 - \exp(-z(T_{t+1} - T_t)). \tag{23}$$

We can write this as a recurrence relation in $\tilde{d}_t$ following equation (21) as follows:

$$\tilde{d}_{t+1} = x_t[d_t + \frac{1 - x_{t-1}}{x_{t-1}} \tilde{d}_t]. \tag{24}$$

The number of individuals dying in the *t th* period is the sum of all the probabilities of the death of individuals whose deaths were recorded before this. To estimate the number at risk ($\tilde{n}_{t+1}$) we use a similar method to the first model in Raza and Broom [8], and so need the value of $\hat{o}$ as in the following equation:

$$\hat{o} = \frac{\sum_{t=0}^{n_\tau} \tilde{d}_t}{n_\tau + \sum_{t=0}^{n_\tau} \tilde{d}_t}. \tag{25}$$

The equation of the total number of patients n_τ is

$$n_\tau = N_r + \sum_{t=0}^{n_\tau} d_t - \sum_{t=0}^{n_\tau} \tilde{d}_t \tag{26}$$

where, N_r is the number of individuals remaining in the study.

Thus,

$$\tilde{n}_{t+1} = \tilde{n}_t - \frac{\tilde{d}_t}{\hat{o}} - c_t, \tag{27}$$

The adjusted hazard function $\hat{h}_a(t)$ and the adjusted survival function $\hat{S}_a(t)$ are given by:

$$\hat{h}_a(t) = \frac{\tilde{d}_t}{\tilde{n}_t}, \tag{28}$$

$$\hat{s}_a(t) = 1 - \hat{h}_a(t), \tag{29}$$

$$\hat{S}_a(t) = \prod_{k=0}^{t-1} \hat{s}(k). \tag{30}$$

Recall that due to only a small number of events occurring later in the study, we cut off the data at time 659 just after one of the death events thus letting $\tau = 659$ $n_\tau = 569.639$ and $\sum_{t=0}^{n_\tau} d_t = 137$. Using $z = 0.005$ we obtain $\sum_{t=0}^{n_\tau} \tilde{d}_t = 127.361$ and $\hat{o} = 0.182$.

2.2 Markov model with censoring for Hewa data

The second Markov Chain structure model for the Hewa data is represented in Fig. 8 below:

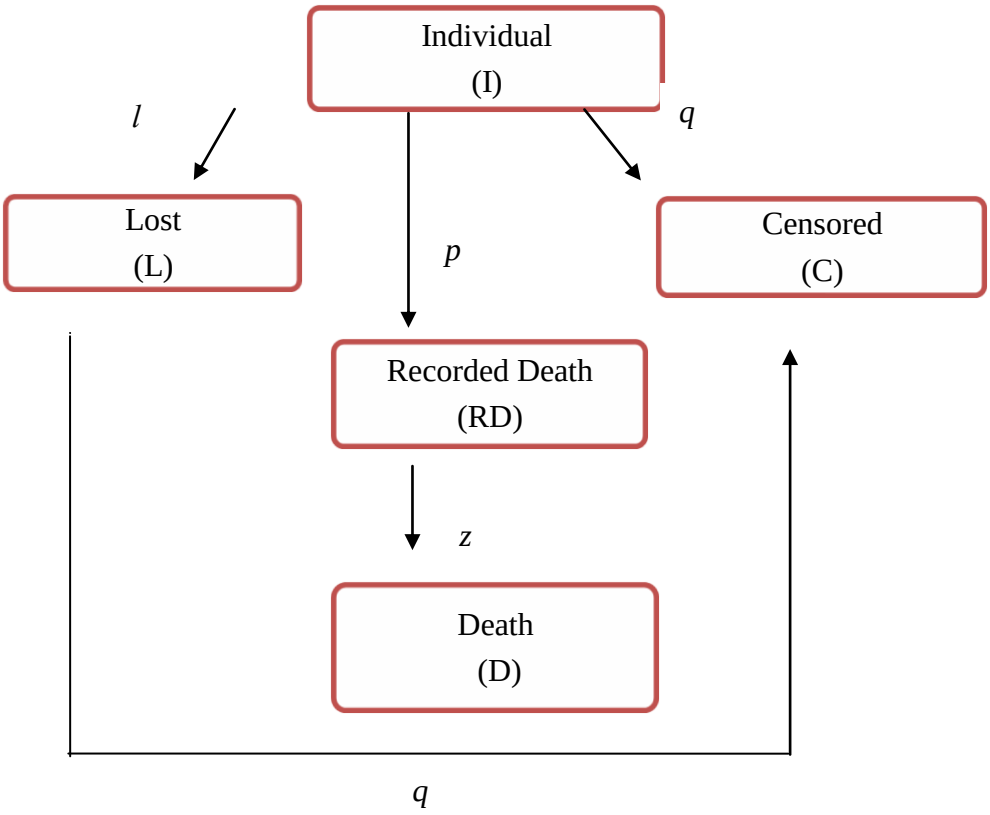

Fig. 8. The Markov survival model with censoring

The probabilities at time t of being in state (I, L,C, RD and D) respectively are represented by the following vector:

$$P = \begin{pmatrix} P_I(t) & P_L(t) & P_C(t) & P_{RD}(t) & P_D(t) \end{pmatrix}. \qquad (31)$$

The transition rate matrix Q is given by

$$Q = \begin{pmatrix} -(l+q+p) & l & q & p & 0 \\ 0 & -q & q & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -z & z \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \qquad (32)$$

The derivatives from equation (9) are represented by:

$$\left(\frac{d}{dt}P_I(t)\;\;\frac{d}{dt}P_L(t)\;\;\frac{d}{dt}P_C(t)\;\;\frac{d}{dt}P_{RD}(t)\;\;\frac{d}{dt}P_D(t)\right)=\left(P_I(t)\;\;P_L(t)\;\;P_C(t)\;\;P_{RD}(t)\;\;P_D(t)\right)\begin{pmatrix}-(l+q+p) & l & q & p & 0\\ 0 & -q & q & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -z & z\\ 0 & 0 & 0 & 0 & 0\end{pmatrix}\Rightarrow$$

$$\left(\frac{d}{dt}P_I(t)\;\;\frac{d}{dt}P_L(t)\;\;\frac{d}{dt}P_C(t)\;\;\frac{d}{dt}P_{RD}(t)\;\;\frac{d}{dt}P_D(t)\right)=$$

$$(-(l+q+p)P_I(t)\;\;l\,P_I(t)-q\,P_L(t)\;\;q(P_I(t)+P_L(t))\;\;p\,P_I(t)-z\,P_{RD}(t)\;\;z\,P_{RD}(t)) \quad (33)$$

The transition rates out of the I state are the same as in the second model for the Nanakaly data, for more detail see Raza and Broom [8], but for the Hewa model we add the rate of transition from state (I) to state (RD) (as opposed to state D) given by (*p*). Thus the equation, initial condition and probability for state (I) are the same as before, i.e.

$$P_I(t)=e^{-(l+p)t}. \quad (34)$$

Similarly, the equation, initial condition and probability for states (L) and (C) are identical to in the previous model so

$$P_L(t)=\frac{l}{l+p}(1-e^{-(l+p)t}). \quad (35)$$

$$P_C(t)=\frac{l+q}{l+p+q}(1-e^{-(l+p+q)t})-e^{-qt}\frac{l}{l+p}(1-e^{-(l+p)t}). \quad (36)$$

For the state (RD) the flow in equals $p\,P_I(t)$ and the flow out is $z\,P_{RD}(t)$. Thus

$$\frac{d}{dt}P_{RD}(t)=p\,P_I(t)-z\,P_{RD}(t). \quad (37)$$

$$\frac{d}{dt}P_{RD}(t)+z\,P_{RD}(t)=p\,P_I(t). \quad (38)$$

Thus as in the previous Hewa model (see equation 19) we get

$$\frac{d}{dt}(e^{zt}\,P_{RD}(t))=e^{zt}\frac{d}{dt}P_{RD}(t)+P_{RD}(t)\,ze^{zt}=p\,P_I(t)e^{zt}=p\,e^{zt}\,e^{-(l+q+p)t}\Rightarrow$$

$$e^{zt}\,P_{RD}(t)=\int p\,e^{(z-l-q-p)t}dt=\frac{p}{z-l-q-p}\,e^{(z-l-q-p)t}+k_5. \quad (39)$$

To find the value of the constant k_5 we consider time zero, so that

$$k_5+\frac{P}{z-l-q-p}=0\Rightarrow k_5=\frac{p}{l+q+p-z} \quad (40)$$

and substituting equation (39) into equation (40) we get equation (41),

$$P_{RD}(t) = \frac{p}{l+q+p-z}[e^{-zt} - e^{-(l+q+p)t}]. \tag{41}$$

Now the last transition from state (RD) to state (D) has rate (z). The value of the flow in is equal to $z\,P_{RD}(t)$ and the flow out is equal to zero. Thus

$$\frac{d}{dt}(P_D) = z\,P_{RD} - 0. \tag{42}$$

By substituting equation (41) into the above formula we get

$\frac{d}{dt}P_D(t) = z\frac{p}{l+q+p-z}[e^{-zt} - e^{-(l+q+p)t}]$. If we let $w = \frac{p}{l+q+p-z}$, and then integrate here we obtain equation (43) below:

$$P_D(t) = -we^{-zt} + \frac{zw}{l+q+p}e^{-(l+q+p)t} + k_6. \tag{43}$$

When $t = 0$ then $P_D(0) = 0$ and thus

$$k_6 = w - \frac{zw}{l+q+p}. \tag{44}$$

Substituting equation (44) into equation (43) we get

$$P_D(t) = \frac{p}{l+q+p-z}[(1-e^{-zt}) - \frac{z}{(l+q+p)}(1-e^{-(l+q+p)t})]. \tag{45}$$

Table 2 is a summary of all the above solutions.

Table 2. Summary of all state transitions for model II Hewa data

State	In flow	Out flow	Equations for state probabilities	Probability values
I	0	$(l+p+q)P_I(t)$	$\frac{d}{dt}P_I(t) = 0 - (l+p+q)P_I(t)$	$P_I(t) = e^{-(l+p)t}$
L	$l\,P_I(t)$	0	$\frac{d}{dt}P_L(t) = l\,P_I(t) - 0$	$P_L(t) = \frac{l}{l+p}(1-e^{-(l+p)t})$
C	$q\,P_I(t)$	0	$\frac{d}{dt}P_C(t) = q\,P_I(t) - 0$	$P_C(t) = \frac{l+q}{l+p+q}(1-e^{-(l+p+q)t}) - e^{-qt}\frac{l}{l+p}(1-e^{-(l+p)t})$
RD	$p\,P_I(t)$	$z\,P_{RD}(t)$	$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - z\frac{p}{l+q+p-z}(e^{-zt} - e^{-(l+q+p)t})$	$P_{RD}(t) = \frac{p}{l+q+p-z}[e^{-zt} - e^{-(l+q+p)t}]$
D	$z\,P_{RD}(t)$	0	$\frac{d}{dt}P_D(t) = z\frac{p}{l+q+p-z}(e^{-zt} - e^{-(l+q+p)t}) - 0$	$P_D(t) = \frac{p}{l+q+p-z}[(1-e^{-zt}) - \frac{z}{(l+q+p)}(1-e^{-(l+q+p)t})]$

Using equations (34) and (41) we can estimate the second model for the Hewa data to determine the survival curve in the Hospital. We consider the estimated hazard function $\hat{h}_c(t)$, represented below:

$$\hat{h}_c(t) = \frac{z\,P_{RD}(t)}{P_{RD}(t) + P_I(t)}\,. \tag{46}$$

This is z times the ratio of the rate of deaths in the population among the at risk individuals (in categories RD and I) and the number of at risk individuals.

Firstly we use the real hazard, function to find the real survival function as shown below:

$$\hat{h}_c(t) = \frac{z\,P_{RD}(t)}{P_{RD}(t) + P_I(t)} = \frac{z\dfrac{p}{l+p+q-z}\left[e^{-zt} - e^{-(l+p+q)t}\right]}{\dfrac{p}{l+p+q-z}\left[e^{-zt} - e^{-(l+p+q)t}\right] + e^{-(l+p+q)t}} \tag{47}$$

$$= \frac{z\,p\left[e^{(l+p+q-z)t} - 1\right]}{l+q-z+p\,e^{(l+p+q-z)t}}. \tag{48}$$

The real survival function is then given by:

$$\hat{S}_c(t) = e^{-\int_0^t h_c(u)\,du}$$

where, using equation (48),

$$\int_0^t h_c(u)du = \int_0^t \frac{z\,p\left[e^{(l+p+q-z)u} - 1\right]}{l+q-z+p\,e^{(l+p+q-z)u}}du\,.$$

Letting $v = e^{(l+p+q-z)u}$ we have

$$\frac{dv}{du} = (l+p+q-z)\,e^{(l+p+q-z)u} \Rightarrow$$

$$\int_0^t h_c(u)du = \int_1^{e^{(l+p+q-z)t}} \frac{z\,p\left[v-1\right]}{(l+q-z+p\,v)}\,\frac{1}{v\,(l+p+q-z)}\,dv. \tag{49}$$

Using partial fractions we obtain

$$\int_1^{e^{(l+p+q-z)t}} \frac{z\,p\left[v-1\right]}{(l+q-z+pv)\,v\,(l+p+q-z)}\,dv = \int_1^{e^{(l+p+q-z)t}} \frac{1}{l+p+q-z}\left[\frac{-z\,p}{(l+q-z)\,v} + \frac{(l+p+q-z)\,z\,p}{(l+q-z)(l+q-z+pv)}\right]dv$$

$$\frac{z\,p}{(l+p+q-z)(l+q-z)}\left[-(l+p+q-z)t + \frac{(l+p+q-z)}{p}\left(\ln\left(l+q-z+p\,e^{(l+p+q-z)t}\right) - \ln\left(l+q-z+p\right)\right)\right].$$

To simplify the above equation we set $h = l+p+q-z$ and $w = e^{(l+p+q-z)t}$. This gives

$$\int_0^t h_c(u)du = \frac{z\,p}{h(h-p)}\left[-ht + \frac{h}{p}\left(\ln(h-p+pw) - \ln h\right)\right] =$$

$$-\frac{z}{h-p}\,pt + \frac{z}{h-p}[\ln(h-p+pw) - \ln h] \Rightarrow$$

$$-\int_0^t h_c(u)du = \frac{z}{h-p}\,pt + \frac{z}{h-p}\left[\ln\frac{h}{h-p+pw}\right]. \tag{50}$$

Using equation (50) the real survival function $\hat{S}_c(t)$ then takes the following form:

$$\hat{S}_c(t) = \exp\left[-\ln\left(\frac{p\,e^{(l+p+q-z)t} + l + q - z}{(l+p+q-z)e^{pt}}\right)^{\frac{z}{l+q-z}}\right] = \left[\frac{(l+p+q-z)e^{pt}}{p\,e^{(l+p+q-z)t} + l + q - z}\right]^{\frac{z}{l+q-z}}. \tag{51}$$

We denote *l/p*, the ratio of probabilities for an individual to be lost to the study or die, respectively, by α, to account for the right censoring in this population caused by the end of the study period. We can use the number of recorded deaths to estimate the number of lost individuals provided we can estimate $\alpha = l/p$.

3 Estimating the p, α, l, q and z Values

At low t, the survival probability satisfies $S(t) \simeq e^{-pt}$, so that $S(0)/S(t) \simeq e^{pt}$, so we estimate the rate of recorded death *p* using the following equation

$$\hat{p} = \frac{1}{t}\ln\left(\frac{S(0)}{S(t)}\right). \tag{52}$$

We tested different values of time *t* for equation (52). In principle the lower the *t* value, the better ($\hat{q}$), assuming that there are enough events for a reliable estimate. For very small *t*, there is little data to use to estimate the value of *p*. There was sufficient data at time *t* =10, where there had been 29 recorded deaths. Here $S(t=10) = 0.9790$ and so we obtain the following estimate of *p*,

$$\hat{p} = \frac{1}{10}\ln\left(\frac{1}{0.9790}\right) \simeq 0.00212.$$

To estimate α we apply the same methods as in Raza and Broom [8] and using equation 53:

$$S(t) = \frac{\alpha + (e^{-pt})^{1+\alpha}}{1+\alpha} \Rightarrow S(\infty) = \frac{\alpha}{1+\alpha}, \tag{53}$$

where $S(\infty)$ is the limiting apparent survival probability for the data. In practice, in contrast to Raza and Broom [8], a small number of individuals remained in the study indefinitely, and so we had to choose a practical cut-off value. We selected the value associated with time ***t***=750**,** which led to an estimated value of $\tilde{\alpha} = 7.07029$. We selected this value because it was close to the equivalent cut off value from Fig. 3 and it gives us an estimated survival curve very close to the survival curve from the real data except for when there are few individuals left in the study. We see this in Figs. 9 and 10.

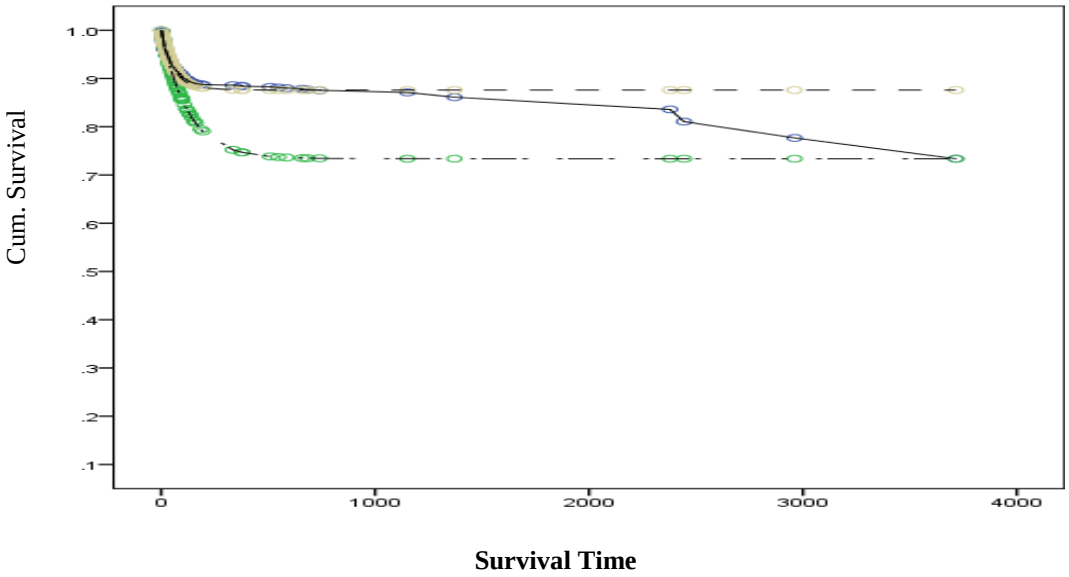

Fig. 9. Kaplan Meier method for apparent and estimated survival functions at time 3715 days (-) is adjusted with cut off up to 750 days, (o) is Real data and (-o) is adjusted without cut off.

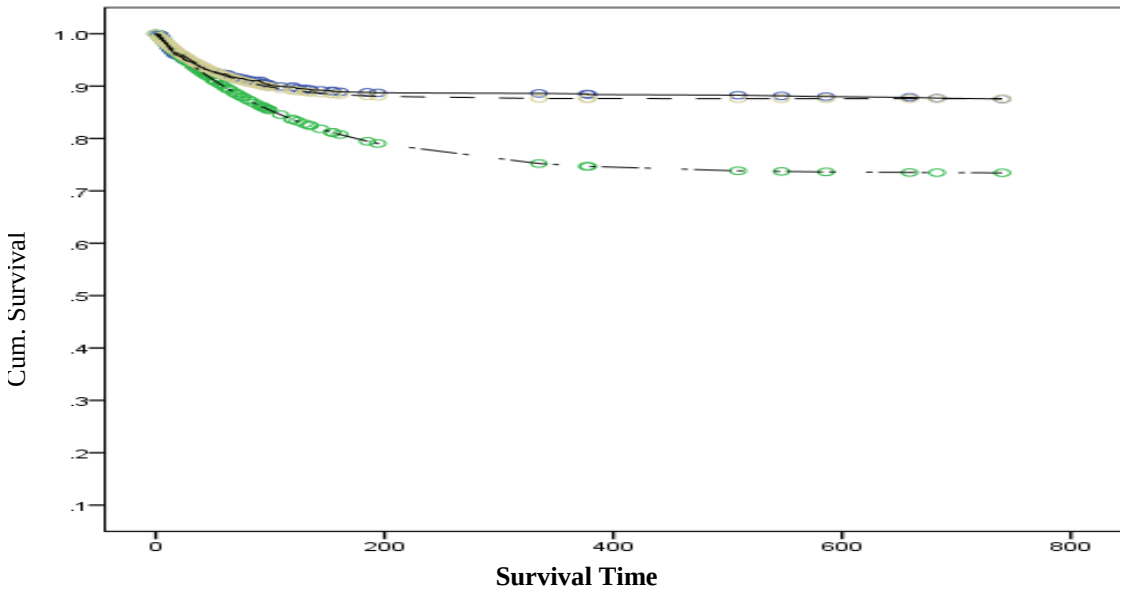

Fig. 10. Kaplan Meier method for apparent and estimated survival function at up tp time 750 days (-) is adjusted with cut off up to 750 days, (o) is Real data and (-o) is adjusted without cut off.

To find the rate of loss of individuals, denoted by l, we use the definition of α ,

$\alpha = \frac{l}{p} \Rightarrow l = \alpha\, p$. Thus we have $\hat{l} = 0.01499$.

The rate of censoring individuals q at time t can be estimated using the equation below.

$$P_c(t) = \frac{l+q}{l+p+q}(1-\exp^{-(l+p+q)t}) - e^{-qt}\frac{l}{l+p}(1-e^{-(l+p)t}),$$

where $P_c(t)$ is the proportion of censored individuals at time t.

After testing different value of *t* we conclude that a small value of *t* has not enough censored individuals for reliable prediction. In practice censoring here is not a homogeneous Markov process (as implicit in the model); note for the models in Raza and Broom [8] we saw that this was not important for that model, but here it is. A large number of censored individuals between *t*=700 to *t*=750 also made these values unreliable. A sensible choice is *t* = 1000 because the variations in $\hat{q}$ are larger below t = 1000. Then the estimated q value is $\hat{q} = 0.00160$. The discussion above is illustrated by Table 3.

Table 3. The estimated rate of censored individuals ($\hat{q}$) at different time (t)

t	200	500	700	750	875	**1000**	2000
$\hat{q}$	0.00093	0.00078	0.00085	0.00211	0.00182	**0.00160**	0.00159

4 Applying Our Models to the Kurdish Data

4.1 Without censoring

From the Hewa data, the value obtained for *S(∞)* (the limiting value of the survival curve) is 0.87607, which gives an estimate of $\tilde{\alpha}$ = 7.07029 individuals lost per death event. Our method applied without censoring then gives the adjusted Kaplan-Meier survival curve in Fig. 11. This figure says that the probability of surviving up to 50 days is 0.990 and up to 100 days is 0.971, during these periods (i.e. up to 100 days) there were 113 death and 51 censored patients. Up to 700 days, there is a probability of survival 0.707, where an additional 20 patients died and 383 patients were censored, supposing that *z* =0.005.

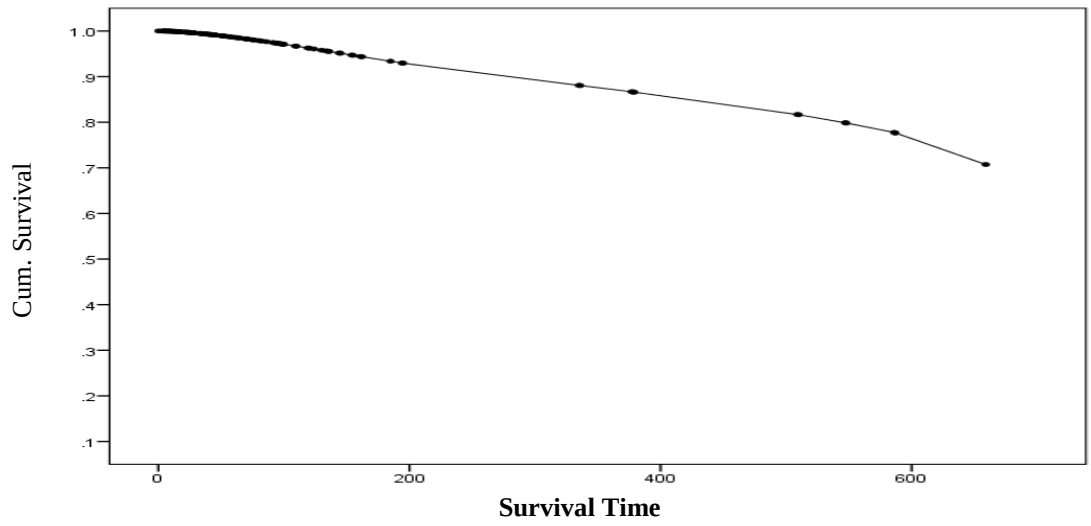

Fig. 11. Adjusted survival curve for the Hewa data using the method without censoring (z=0.005)

Since z depends upon an intuitive estimate using little evidence, we considered z to be 10 times bigger and 10 times smaller than 0.005. As above, we took the times of death to be given by the Markov process starting at the time of diagnosis which we discussed in Section 2.2. Fig. 12 is the equivalent survival curve when *z* =0.05, which shows a different shape of survival curve for the same period compared to Fig. 10. For the first period of 50 days the survival probability is 0.941 and for the second period up to 100 days it is 0.889, while for the period up to 700 days the probability is 0.744.

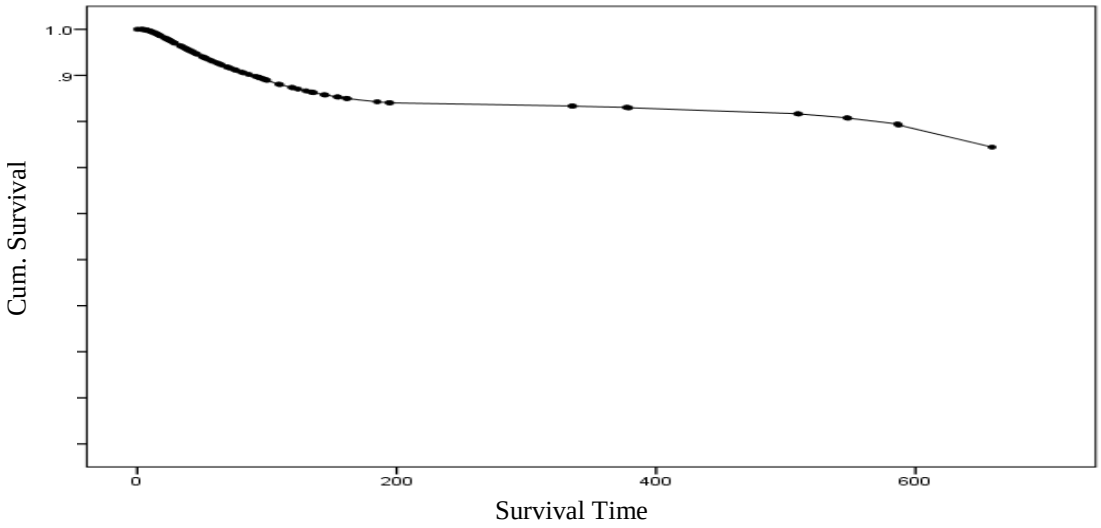

Fig. 12. Adjusted survival curve for the Hewa data using the method without censoring (z=0.05)

Fig. 13, the survival curve when z=0.005, shows different survival probabilities again for the same period respectively. For up to 50 and up to 100 days the survival probabilities are 0.999 and 0.997 respectively, while the survival probability up to 700 days is 0.929.

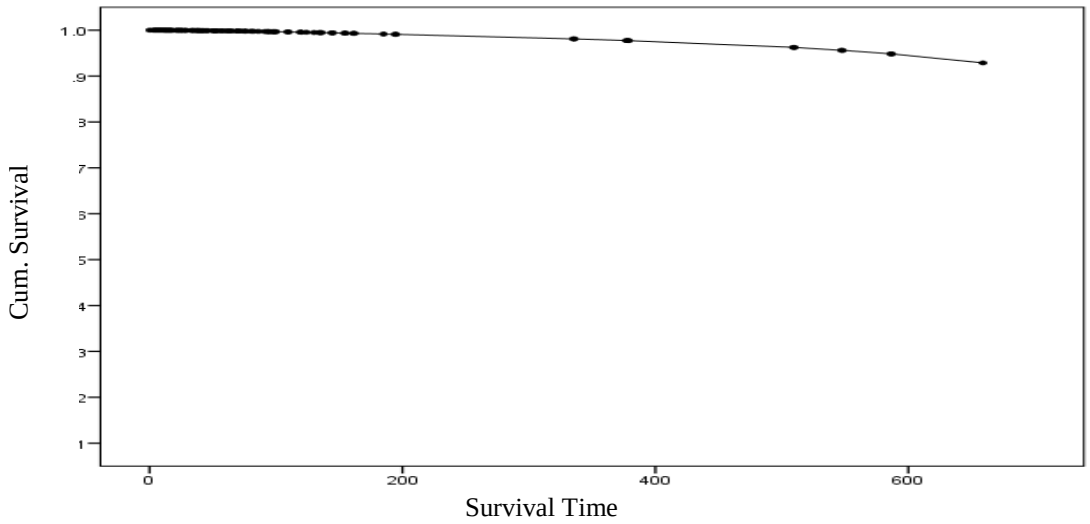

Fig. 13. Adjusted survival curve for the Hewa data using the method without censoring (z=0.0005)

4.2 With censoring

For the survival curve for the with censoring model we again need to estimate is *z*. For this case we use the previous *z* value of the first model without censoring, which was chosen to be equal to 0.005. We see this in Fig. 14.

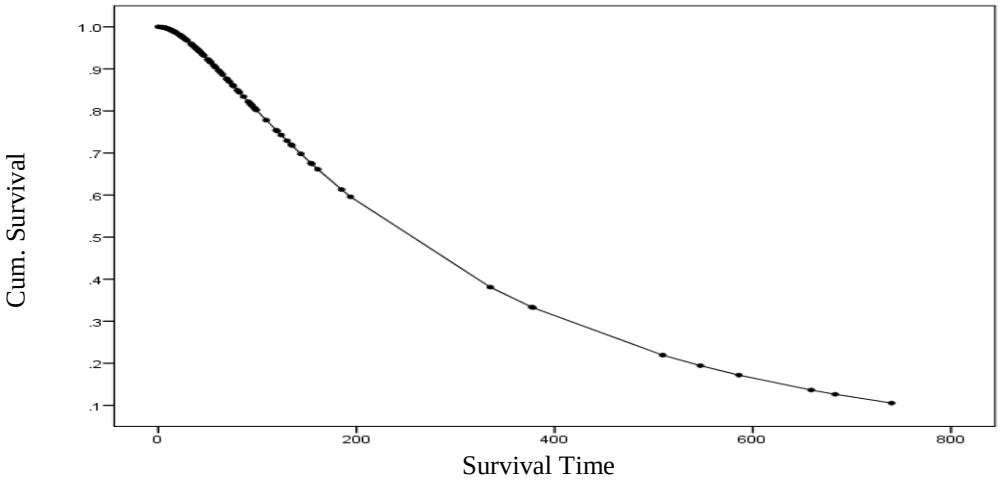

Fig. 14. Adjusted survival curve for the Hewa data using the method with censoring (z=0.005)

Figs. 15 and 16 show the equivalent figures using *z* equal to 0.05 and 0.0005 in the same way as for the previous model.

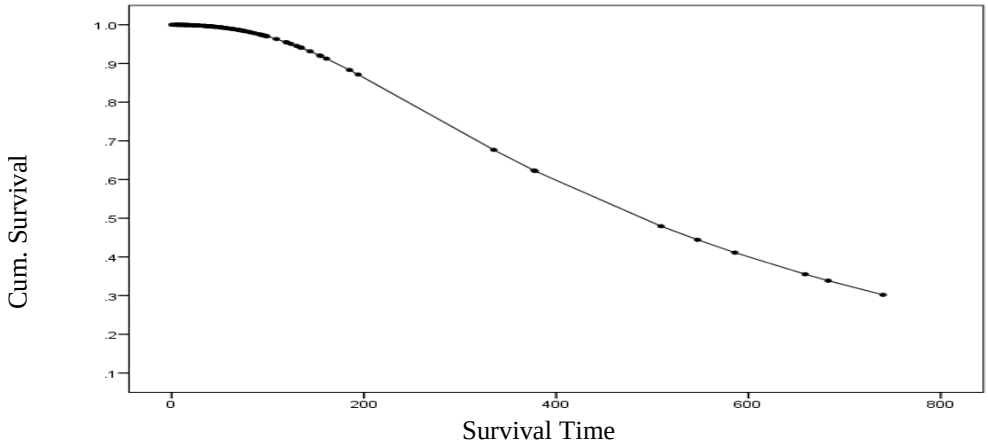

Fig. 15. Adjusted survival curve for the Hewa data using the method with censoring (z =0.05)

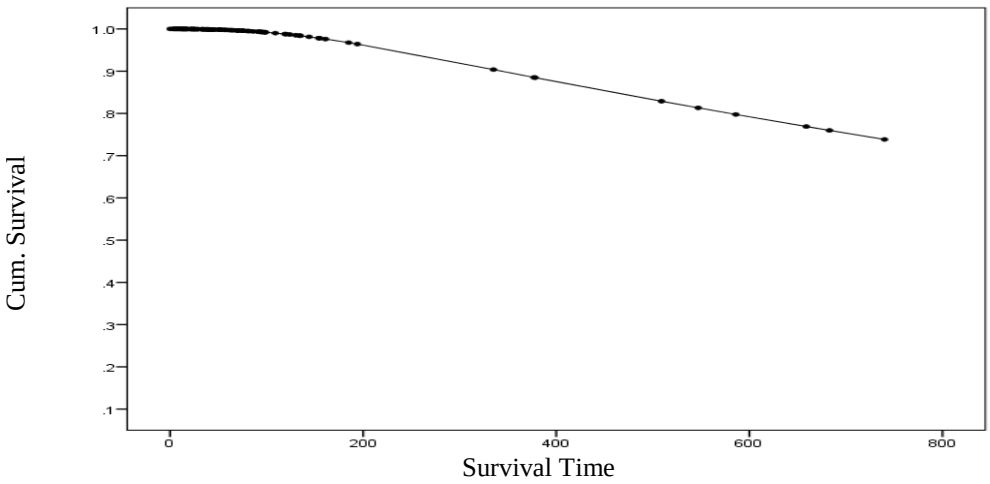

Fig. 16. Adjusted survival curve for the Hewa data using the method with censoring (z=0.0005)

For the second model we used the Markov Chain to estimate the rate of recorded deaths, the rate of censoring, the rate of losing individuals and the rate of death (p, q, l and z)..The survival curves arising from the two models are considerably different (Figs. 11 and 14), partly because of non-homogeneous censoring in the data. Observe that while the survival curve corresponding to z = 0.005 appears to be approaching zero, setting z = 0.05 yields a curve decreasing more slowly. Finally, for z = 0.0005, we obtain an even slower drop of the survival function, which does not appear to be approaching zero in the observed timeframe, casting doubt on the reliability of this estimate.

5 Discussion

In this paper we considered the modelling of survival data when crucial information is missing. Indeed, in the scenarios we consider, towards the end of the study missing data constitutes the vast majority of the dataset, making accurate estimation very challenging. The models in this paper are built on those from our earlier work, Raza and Broom [8], which considered how to deal with a specific type of missing data in a health setting, where individuals leave a medical study without notifying the hospital concerned. For the new model there is an additional problem of missing data, caused by uncertainty about the time of death compared to (what could be taken as) the recorded time of death. The original and new works are both based upon a set of breast cancer data from a hospital in the Kurdistan region of Iraq, Nanakaly and Hewa hospitals, respectively.

Here for the Hewa dataset there were thus two problems with the data, the problem of "lost" individuals, but also the problem of the absence of definitive times of death. We developed two models that tried to overcome these issues. We generated a new Markov chain model for two distinct cases, for data with and without censoring. For the first of our two models, the without censoring case, we estimate the number of observation and the number of deaths in a natural way. Here we adjusted the true number of individuals at risk $\tilde{n}_t$ and the estimated number of deaths $\tilde{d}_t$, which depends upon the rate of transition from the Recorded Death class to the Death class (z). The estimated hazard rate was then adjusted from that using the original data d_t/n_t to $\tilde{d}_t/\tilde{n}_t$. As mentioned in the previous paper for the Nanakaly data there is a consistent scaling which preserves the order of the risk factors, though not in quite as straightforward a manner, as $\tilde{d}_t$ is a weighting of the number of recorded deaths from different time periods.

However, for the second model, with censoring, we need to estimate crucial parameter values, and for appropriately chosen estimates we get realistic survival function curves. However, estimates are made with little information. Thus, while our survival curves are plausible, we cannot rely on them. Here we used the Markov Chain to estimate the rate of recorded death, the rate of censoring, the rate of losing individuals and the rate of death (p, q, l and z) in our model. Further, we saw that the survival curves arising from the two models were considerably different (see Figs. 9 and 11), partly at least as a consequence of the non-homogeneous censoring in the data. Thus, the conclusions from our analysis of the risk factors in the Hewa data are not robust. We believe that it is not possible to obtain such robust conclusions from the current data. 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 leads on 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. 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 which most need accurate survival models and so the application of our methods can be of significant value.

The problems stem from an incomplete database due to various reasons based on what the Region has gone through in the past. The Region separated from the central Government in 1991 causing an internal conflict after that era. Every sector has been affected by this abnormal environment including the health sector. In comparison to various other diseases, the rate of breast cancer among women in the Region has risen

dramatically. This is why we chose this disease to study in Kurdish society. The research involves the application of survival analysis in order to find new tools and ways to illustrate the importance of knowing the survival rate. When we first started collecting the data, we found that some data was missing which may be due to either those recording the data not realising the importance of the details or the patients not returning to the hospitals for a follow-up check. We have therefore made some adjustments to the data by proposing the use of a new model developed using mathematical methods.

The best way to collect data that we can depend on is the use electronic health records (EHR) as recognized by the Academy of Medical Royal Colleges in 2008 (https://www.aomrc.org.uk/publications/) statements/doc_view/217-academy-statement-the-case-and-vision-for-patient-focusedrecords.html). The most important thing we must look at is the list of clinical record headers, each with a description of what should be logged under each header. In addition to the clinical categories, the full set of record headers should include admission, handover, discharge, outpatient, referrals, communications, and space for special remarks. The recorded data in the Kurdish hospital Hewa did not meet these standards (HSCIC, 2013).

The specific subject of our data, breast cancer, is the most common type of cancer in women in both developed and developing countries. The incidence of breast cancer is increasing in developing countries due to increased life expectancy, increased urbanization and wider adoption of western lifestyles. Although some risk reduction might be achieved with prevention, these strategies cannot eliminate the majority of breast cancers that develop especially in low and middle-income countries where breast cancer is diagnosed in very late stages [20]. Early detection is therefore required to improve the outcome of breast cancer and remains the cornerstone of breast cancer control.

The World Health Organization promotes breast cancer control within the context of national cancer control programs integrated with non-communicable disease control and prevention. In order to make efficient use of the medical equipment and trained staff available, the WHO has adopted a programme heavily focusing on early detection and prevention [21]. Given the relative underdevelopment of medical infrastructure across the region and the consequent difficulty of arranging regular check-ups, it is crucial to educate the female population on early signs of breast cancer as well as the associated risks [22].

Mortality rates related to breast cancer around the Eastern Mediterranean are steadily increasing even as the overall incidence is lower than across the developed world. One explanation proffered by Mahdi et al. [23] is that many cases of breast cancer appear to be diagnosed at an advanced stage of the disease, complicating treatment independent of the medical resources available. A useful case study is presented by Iraq. In the decade from 2009 to 2019, the incidence of breast cancer almost doubled from 9.5/100,000 to 18.2/100,000, making it the leading tumour diagnosis as well as the second most lethal cancer in Iraq [24]. Recent data suggests that in addition to rising incidence rates, breast cancer in Iraqi women is diagnosed at later stages and younger ages when measured against the corresponding figures from western countries [25,26]. This combination of trends makes the reported mortality rate one of the highest in the world, according to recent comparative literature [20].

As a result of war, internal displacement, and the chaos and disorder associated with civil strife over the last 40 years, the Iraqi healthcare sector has not been able to develop at rates comparable to other countries of the region. While some government initiatives have been launched in an attempt to reverse the tide and improve certain key indicators [25], health outcomes have broadly deteriorated [27,28].

The breast cancer early detection (BCED) programme was evaluated over nine years in a study of a total of 360 patients at the Medical City Teaching Hospital and the national cancer research center (INCRC) in Baghdad. During the studied period, the proportion of women presenting with stage IV cancer significantly fell from 15.2% in the first year to 9.1% at the end of the study [29]. At the same time, the proportion of advanced stage (III and IV) diagnoses remains alarmingly high.

We note that the awareness and knowledge of breast cancer risks and detection techniques among Iraqi women was the focus of a study carried out in the Kurdistan Region of Northern Iraq. The convenience sampling technique (CST) was used to select 400 female Kurdish patients presenting at either the maternity teaching hospital in Erbil or one of the primary healthcare centres (PHCs). At first glance, the results of the study were encouraging; for instance over half of the surveyed population reported some knowledge of mammography, and almost 90% were aware of the beneficial effects of early detection on prognosis. However, some misconceptions

remain relatively widespread; roughly a third assumed breast cancer to be altogether preventable. Furthermore, the level of accurate knowledge of breast cancer risks and early detection techniques correlated heavily with socioeconomic status, suggesting the necessity of further outreach programs targeting poorer and less literate communities [30].

In summary, the treatment of breast cancer in the region presents particular challenges compounded by the lack of accurate data and is an appropriate area to apply our methodology. However, while of course the supply of such data is the preferable solution, we believe that our methods are of potential value in many cases where there are significant problems with the accuracy and availability of data.

6 Conclusion

All over the world, especially in the developing regions, the rates of breast cancer, the most common malignancy in women, constituting just under one fifth of cancers in females, are on the rise. Kurdistan-Iraq is no exception, with an age-adjusted incidence rate of 68.9 per 100,000 per year; in fact, breast cancer is the most prevalent cancer among the population (affecting about one third of female cancer patients), with particularly alarming rates among the younger demographic, according to the Kurdistan-Iraqi Cancer Registry. In order to tackle this problem, a precise understanding of the survival rate is essential. In order to do so, in this work we adopt the Cox regression and Kaplan-Meier methods.

The main conclusion is that we have developed a new method for performing a survival analysis on a set of data where there are important unknown factors. 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 from the biased one obtained directly from the data. Using data from Hewa hospital we generated a new model in two cases; with and without censoring. For these models we need to estimate the number of deaths and also crucial parameter values. For some parameter estimates we get realistic survival function curves. However, estimates are made with little information. Thus, while survival curves are plausible we cannot rely on them. These models are nonetheless a significant improvement on trying to estimate survival curves without them.

Disclaimer

Much of the work in this paper was originally published in a thesis produced by the first author.

The Thesis is available at this link: https://openaccess.city.ac.uk/id/eprint/15877/1/Raza,%20Mahdi.pdf

[As per journal policy, Thesis article can be published as a journal article, provided it is not published in any other journal]

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Tableman M, Kim Js. Survival Analysis Using S. New York: Chapman & Hall/CRC; 2004.

[2] Majid R, Mohammed H, Hassan H, Abdulmahdi W, Rashid R, Hughson M. 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. 2012;12(16):1-10. Available:http://www.biomedcentral.com/1472-6874/12/16

[3] Compton CC, Byrd DR, Garcia-Aguilar J, Kurtzman S, Olawaiye HA, Washington MK. Cancer Survival Analysis; 2012,21 September 2013. Available:http://www.springer.com/978-1-4614-2079-8

[4] Lawless JF. Statistical Models and Methods for Lifetime Data, 2nd Ed. New Jersey: John Wiley and Sons, INC; 2003.

[5] Bedford T, Cook R. Probabilistic Risk Analysis Foundation and Methods. USA: Cambridge University Press; 2009.

[6] Ziaei J, Zohreh S, Iraj A, Saeed D, Ali P, Jalil V. Survival Analysis of Breast Cancer Patients in Northwest Iran. Asian Pacific Journal of Cancer Prevention. 2013;14(1):39-42.

[7] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. The Stata Journal. 2009;9(2):265–290.

[8] Raza MS, Broom M. Survival analysis modeling with hidden censoring. Journal of Statistical Theory and Practice. 2016;10(2):375-388.

[9] Miller LD, Johanna S, Joshy G, Vinsensius B, Liza V, Alexander P, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Sciences of the United State of America. 2005;102(38):13550-13555.

[10] Hamdan H, Garibaldi JM. Modelling Survival Prediction in Medical Data. Intelligent Modelling and Analysis (IMA) Research Group. University of Nottingham: UK; 2009.

[11] Abuelghar WM, Elsaeed MM, Tamara TF, Elaithy MI, Ali MS. Measurement of serum estradiol / progesterone ratio on the day of embryo transfer to predict clinical pregnancies in injection (ICSI) cycles. Is this of real clinical value? Middle East Fertility Society Journal. 2013;18(1):31-37.

[12] Abadi A, Farzaneh A, Chris B, Parvin Y. Breast cancer survival analysis: Applying the generalized gamma distribution under different conditions of the proportional hazards and accelerated failure time assumptions. International Journal of Preventive Medicine. 2012;3(9):644–651.

[13] Saleem M, Raza A. On Bayesian Analysis of the exponential survival time assuming the exponential Censor Time. Pakistan Journal of Science. 2011;63(1):44-48.

[14] Tabatabai MA, Wayne M Eby, Nadim Nimeh, Karan P Singh. Role of metastasis in Hypertabastic Survival Analysis of Breast Cancer: Interaction with Clinical and Gene Expression Variables. Libertas Academica Ltd. 2012;5:1-17.

[15] Clayforth C, Fritschi L, McEvoy SP, Byrne MJ, Ingram D, Sterrett G, Harvey JM, Joseph D, Jamrozik K. Five-year survival from breast cancer in Western Australia over a decade. The Breast. 2007;16(4):375-381.

[16] Al-Riad al Sharif, Investigations and reportages: Hospital (hiwa) Oncology and Hematology in Sulaimaniyah. The Union, daily political newspaper. 2012;3707(Sep. 16, 2012).
Sited in Jan. 18, 2015.
Available:http://www.alitthad.com/paper.php?name=News&file=article&sid=125630

[17] Breslow NE. Analysis of survival data under a proportional hazards model. International Statistical Review. 1975;43:45-57.

[18] Schumacher M, Bastert G, Bojar H, Hubner K, Olschewski M, Sauerbrei W, et al. Randomized 2x2 Trial Evaluating hormonal Treatment and the Duration of Chemotherapy in Node-Positive Breast Cancer Patients. Journal of Clinical Oncology. 1994;12(10):2086-2093.

[19] Alwan NA. Breast Cancer Incidence Among Iraqi Women Profiled; 2010.
Available:http://www.sciencedaily.com/releases/2010/03/100311074127.htm
Accessed on April, 25, 2013.

[20] IARC. Estimated age-standardized mortality rates (World) in 2020, breast, females, all ages. CANCER TODAY. Lyon: International Agency for Research on Cancer; 2020b.
Retrieved July 28, 2022.
Available:https://gco.iarc.fr/today.

[21] Lyons G, Sankaranarayanan R, Millar A, Slama S. Scaling up cancer care in the WHO Eastern Mediterranean Region. Eastern Mediterranean Health Journal. 2018;24(01):104-110.

[22] Duggan C, Dvaladze A, Rositch AF, Ginsburg O, Yip C, Horton S, et al. The breast health global initiative 2018 global summit on improving breast healthcare through resource- stratified phased implementation: Methods and overview Cancer. 2020;126(S10):2339–2352.

[23] Mahdi H, Mula-Hussain L, Ramzi ZS, Tolba M, Abdel-Rahman O, Abu-Gheida I, et al. Cancer Burden Among Arab-World Females in 2020: Working Toward Improving Outcomes. JCO Glob Oncol. 2022 Mar;8:e2100415.
DOI: 10.1200/GO.21.00415
PMID: 35259001; PMCID: PMC8920429.

[24] Ministry of Health and Environment, Republic of Iraq. Annual Report Iraqi Cancer Registry 2020. Iraqi Cancer Board; 2021.
Retrieved on 25 August, 2022.
Available:https://moh.gov.iq/upload/2991322580.pdf

[25] Alwan N, Kerr D. Cancer control in war-torn Iraq. The Lancet Oncology. 2018;19(3):291–292.

[26] Alwan N. General Oncology care in Iraq. In: Al-Shamsi H, Abu-Gheida I, Iqbal F, Al-Awadhi A. Cancer in the Arab World. Singapore: Springer. 2022;353-362.

[27] World Health Organization (WHO). The Global Health Observatory, Health workforce; 2022.
Retrieved 18 July, 2022.
Available:https://www.who.int/data/gho/data/themes/topics/health-workforce

[28] Physicians for Human Rights. Challenges Faced by the Iraqi Health Sector in Responding to COVID-19; 2021.
Retrieved July 28, 2022.
Available:https://phr.org/our-work/resources/challenges-faced-by-the-iraqi-health-sector-in-responding-to-covid-19/

[29] Alwan NAS, Tawfeeq FN, Maallah MH, Sattar SA, Saleh WA. The Stage of Breast Cancer at the Time of Diagnosis: Correlation with the Clinicopathological Findings among Iraqi Patients. Journal of Neoplasm. 2017a;2(3):22.

[30] Husamaldien L, Dauod A. Knowledge, awareness and practices about breast and cervical cancer in a group of women in Erbil city-Iraq. Tikrit Medical Journal. 2016;21(2):54-66.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/108897