Journal of Statistical Theory and Practice (2026) 20:9
https://doi.org/10.1007/s42519-025-00520-9

ORIGINAL ARTICLE

Applying Cumulative Survival Functions to Age Comparison Data Sets on Breast Cancer

Mahdi Saber Raza[1] · **Mark Broom**[2]

Received: 3 April 2025 / Accepted: 10 November 2025

Abstract

A comparative examination of breast cancer survival between two data sets from the Kurdistan area of Iraq and a corresponding data set from Germany is the aim of this paper. Using breast cancer data from the Kurdistan area of Iraq, both censored and unfiltered, we developed a methodology in a previous publication (2016) for predicting survival probabilities and hazard functions in a health context when a significant fraction of participants are lost to the research. This study follows earlier research (2023) where we had to use unique estimation methods to address the two Iraqi datasets' filtering problems. In particular, the data from Nanakaly hospital in the city of Erbil and Hewa hospitals in the city of Sulamani involved problems with hidden censoring affecting the survival time, leading to significant biases in survival curves generated using standard methods, and we had developed new Markov chain-based methods for generating survival curves providing adjusted Kaplan Meier analyses. Due to the availability of a reliable survival function, we chose to work with a German data set from the W. Sauerbrei Institute for Medical Biometry and Informatics, University of Freiburg—Germany. Our data analysis leads us to the conclusion that younger German women had a higher breast cancer survival rate than patients from the Kurdistan Region of Iraq.

Keywords Survival analysis · Age study · Kaplan-Meier · Cumulative survival function · Breast cancer

✉ Mark Broom
mark.broom.1@city.ac.uk

Mahdi Saber Raza
mahdi.raza@su.edu.krd

[1] Department of Software and Informatics Engineering, Salahaddin University-Erbil/Iraq College of Engineering, Erbil, Iraq

[2] Department of Mathematics, School of Science and Technology, City St. George's, University of London, London, UK

Published online: 08 December 2025

1 Introduction

One of the most prevalent malignancies in women globally, breast cancer ranks first in terms of causes of mortality for women, ahead of lung cancer [1]. Approximately 2.09 million women receive a breast cancer diagnosis each year, and 627000 of them die as a result of the illness [2] and [3]). Breast cancer is the most prevalent cancer in Iraq and the leading cause of death for Iraqi women.

Patients' survival (or time to event) data is frequently gathered and evaluated in clinical trials. Both the overall number of incidents and the rate at which an event of interest occurs can be compared between two or more groups using such data. The normal distribution relevant for many other forms of data does not apply to survival data; data are non-negative and can be skewed. This could be due to events that happen quickly after an action (like a patient dying after surgery) or events that happen primarily over a lengthy period of time (like life expectancy in the UK). More significantly, they are frequently censored (have missing or partial data), which can happen for a number of reasons. A branch of statistics called survival analysis is used to model this kind of data. It is frequently used in medical research to evaluate various treatments within a clinical trial or to analyse the time to disease remission, progression, or death for patient cohorts [4].

Modelling and evaluating time-to-event data often referred to as "failures" is the main focus of survival analysis. Examples include the duration of a tumour's first recurrence (also known as the length of remission) following initial treatment and the period before an electrical component breaks [5]. A "failure" time might not be recorded because of intentional planning or arbitrary censorship. This would happen in this study if a patient has moved away or is still alive at the conclusion of a clinical trial period [6], [7]. The need for techniques of analysis that allow for filtering is the main driver behind the development of specialised models and procedures for failure time data. Then, survival analysis can be viewed as a set of statistical techniques that take into account data that has been suppressed by time to occurrence. The impact of covariates on survival duration is investigated.

Although censoring and time-dependent covariates are frequently used in survival analysis, they are not exclusive to it. These characteristics can also be handled by other methods, such as some regression models and longitudinal data analysis. But because survival analysis is created especially for analysing time-to-event data, it works especially well with censoring and time-dependent covariates, which are common in this type of data [8].

The most well-known function in survival analysis is the survival function, which expresses the likelihood that a person will live up to time t. The hazard function, or the risk of mortality (per unit time), is associated with this [9], [10]. According to accepted nomenclature (see [11], if our individual has lifetime T, the survival function is

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

The hazard rate or hazard function is

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

This can be derived as follows:

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

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

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

which leads to the important expression

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

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

Although continuity will be presumed, ideas and formulae can be changed to incorporate density function jumps when necessary [12]. There are other methods for directly estimating these basic attributes from data, but the Kaplan-Meier estimator, which uses the discrete hazard function to predict the hazard function, is arguably the most straightforward and reliable [13, 14].

It is common practice to compare cancer survival across various groups. For instance, among calendar times, socioeconomic classes, or geographical areas. Numerous research, such as the International Cancer Benchmarking Partnership [15], EUROCARE, and CONCORD studies, compare cancer survival across nations.

Since mortality from other causes varies across the groups under comparison, comparing cancer survival is made more difficult. This will affect the likelihood that the malignancy will really cause death. Therefore, estimating net survival—which is defined as survival in the hypothetical world where death from other causes is impossible—is typically the goal of comparisons of cancer survival among populations. This makes it possible to compare cancer survival in a "fair" way. Since patients in the actual world run the chance of dying from other causes, we never see net survival. As a result, assumptions must be made in order to estimate net survival.

When comparing cancer survival, age is a significant factor. As people age, their net survival for most malignancies declines. Therefore, it is crucial to "adjust" for variations in the age distribution when comparing survival across groups. When a single summary measure is required, traditional age standardization is used, where

an estimate of relative survival is obtained separately in age groups and a weighted average calculated with the weights reflecting an international standard population [16].

Age-related disparities in survival are common. For instance, global comparisons for lung, prostate, colorectal and breast cancer revealed greater disparities in relative survival as age increased [17]. These significant variations may be overlooked if only an overall average is provided. Age-standardized relative survival, on the other hand, will remain a useful summary metric. Relative survival, or the ratio of all-cause survival to projected survival, has historically been used to assess net survival. When the focus is on the overall average of net survival, Pohar Perme et al. [18] demonstrated that the widely used Ederer II method and other approaches may be biased and proposed an alternative (the Pohar Perme method).

Statistical modelling is another method for estimating net survival [19]. A number of assumptions can be included in statistical models, such as proportional excess hazards, only specific interactions, or the idea that the true net survival can never rise [20]. Using breast cancer data from the Kurdistan area of Iraq, both censored and unfiltered, in Raza and Broom [11] we developed a methodology for predicting survival probabilities and hazard functions in a health context where a significant number of participants are lost to the research. We considered how to analyse cases where "hidden censoring" occurs, where individuals have effectively left the study but the hospital is unaware of this. We developed a new Markov chain-based methodology for generating survival curves and hazard functions and demonstrated this using a breast cancer dataset from the Kurdistan region of Iraq. In Raza and Broom [12], we expanded on these models to examine additional concerns regarding the accuracy of patient death records. For example, deaths frequently take place in family settings outside of hospitals, and patients ceasing treatment and contact with the hospital may or may not indicate their death, therefore, the record of their time of death may not be accurate. We also showed how to generate survival curves and hazard functions using a separate breast cancer dataset from the Kurdistan region of Iraq using a new Markov chain-based methodology.

A comparative examination of breast cancer survival between two data sets from the Kurdistan area of Iraq and a corresponding data set from Germany is the aim of this paper. In particular, the data from Nanakaly hospital in the city of Erbil and Hewa hospitals in the city of Sulamani involved problems with hidden censoring affecting the survival time, leading to significant biases in survival curves generated using standard methods, and we had developed new Markov chain-based methods for generating survival curves providing adjusted Kaplan Meier analyses. Due to the availability of a reliable survival function, we chose to work with a German data set from the Institute for Medical Biometry and Statistics, University of Freiburg—Germany. Our data analysis leads us to the conclusion that younger German women had a higher breast cancer survival rate than patients from the Kurdistan Region of Iraq.

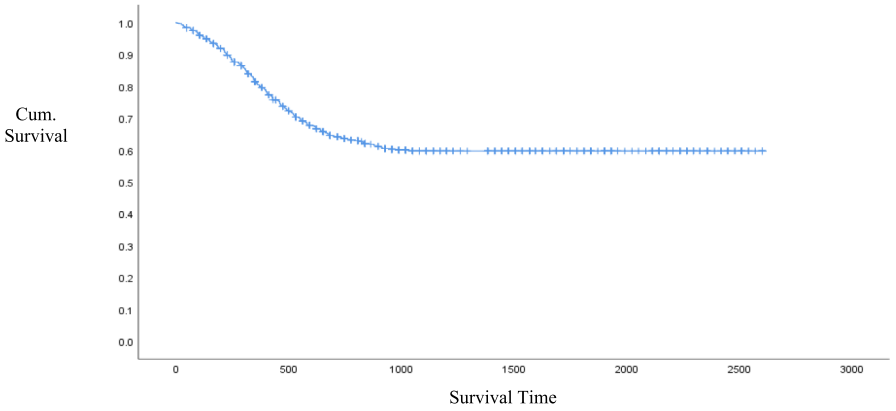

Fig. 1 Cumulative survival function curve for the Nanakaly unadjusted data

2 Survival Analysis

The goal of survival analysis is to examine the time interval between enrolling in a study and a subsequent event. Although survival analysis is helpful in a variety of scenarios beyond mortality, its name derives from its initial focus on the interval between treatment and death.

2.1 Survival Analysis for the Nanakaly Data

The condition of breast cancer at Nanakaly Hospital in the Kurdistan Region is examined in this part. This section's study revisits the work published by Raza and Broom [11]. Using the Nanakaly data, unadjusted for concealed censoring, as displayed in Fig. 1, we first calculate the survival curve for patients with breast cancer (for a description of the problems and the process for generating appropriately adjusted data, see Sect. 2.4). Here the Kaplan-Meier curve for the original data (713 patients) has been used. We see that the survival curve becomes very flat after a while, indicating that many patients have left the study and only a few actually remain from which mortality data can be acquired.

2.2 Analysis for the Hewa Data

In this section we consider breast cancer data from Hewa hospital in the Kurdistan Region, revisiting work published in [12]. As with the previous section for the Nanakaly data, we consider the unadjusted data in Fig. 2 (again see Sect. 2.4 for a discussion of the issues and the methodology for producing suitably adjusted data for this Hewa dataset). Here we show the Kaplan–Meier curve for the original data, involving 1163 patients.

Figure 2 again shows that the probability of death appears very small after 700 days as the curve flatters out around this time. As in the Nanakaly data, this is likely

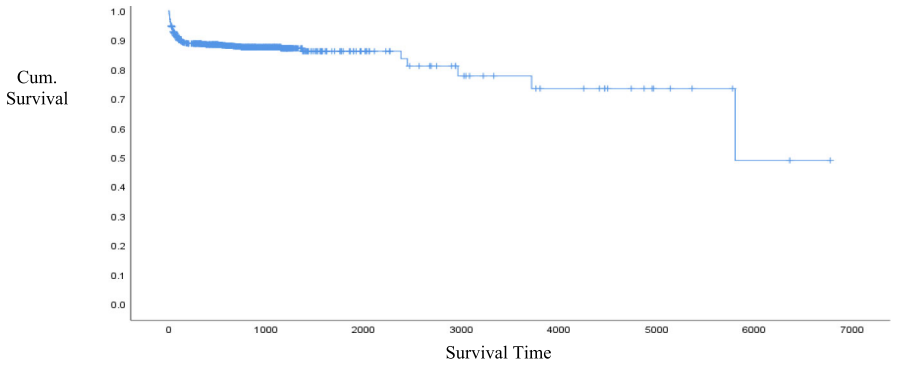

Fig. 2 Cumulative survival function curve for Hewa unadjusted data

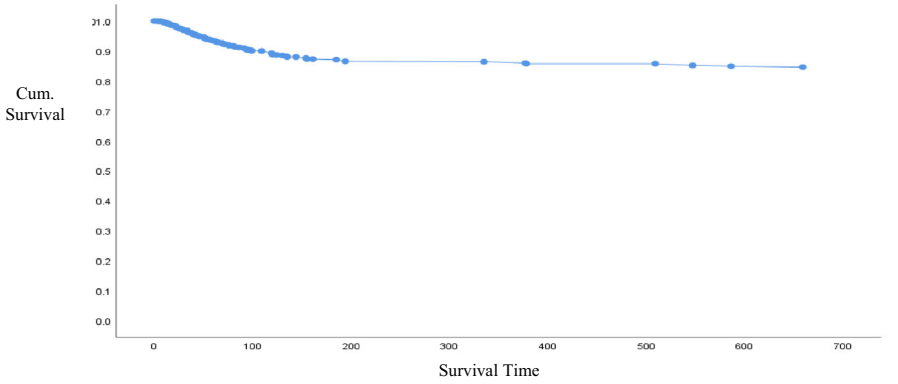

Fig. 3 Survival curve including censoring for the Kurdish data from the Hewa unadjusted data for 700 days

because only a small number of patients remain in the study after this time. Since we see a small number of patients remaining for several thousand days, the data after this period was removed as shown in Fig. 3.

2.3 Survival Analysis for the German Data

The data on breast cancer from a German hospital is examined in this section. Patient records from a German Breast Cancer Study Group (GBSG) at the University of Freiburg's W. Sauerbrei Institute for Medical Biometry and Informatics are among the data collected. 720 patients with node-positive breast cancer participated in the 1984–1989 study. Complete predictive variable data were available for 686 of those individuals (Royston and Altman, 21). First, the cumulative survival function for both the entire data set and two different age groups is calculated using the Cox regression equation. The cumulative survival function for the German data is displayed in Fig. 4.

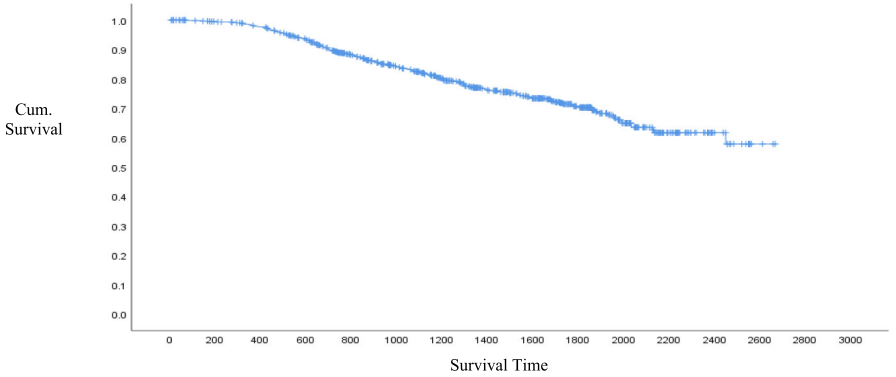

Fig. 4 Cumulative survival function curve for German data

2.4 Adjusted Data for Nanakaly and Hewa Hospitals

Censoring occurs when an individual is not followed up until occurrence of the event of interest. There is loss of information due to this incomplete observation because their different experience would lead to bias in the study. It is caused by failure to follow-up, withdrawal from the study, study termination when subjects had different dates of enrolment, or death due to a competing risk. Censored patients contribute to the analysis until the time of censoring (see [22]). Allowing for censoring may not be straightforward, for example, if individuals do not get censored at a constant rate, the censoring time will follow a distribution other than an exponential distribution, which arises from a natural Markov model for censoring as we describe below. Potentially more significant problems result from the lack of knowledge of the times of death. In our models deaths are assumed to follow a Markov process from the time of diagnosis and there are two main sources of error. Firstly, the rate (denoted in our second model below by z) of this process is unknown and had to be estimated and we have thus considered a range of values. Secondly, this may not be a Markov process, which would also affect the shape of the survival curve.

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 and deaths. For a detailed rationale of the modelling approach and explanations of parameter choices, see [11]. Figure 5 represents the original structure of the Markov survival model with no censoring [11].

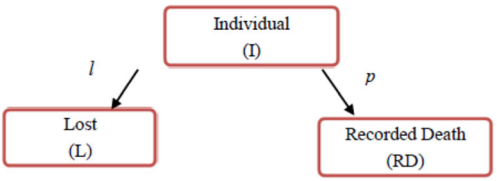

Fig. 5 The Markov model without censoring for Nanakaly data

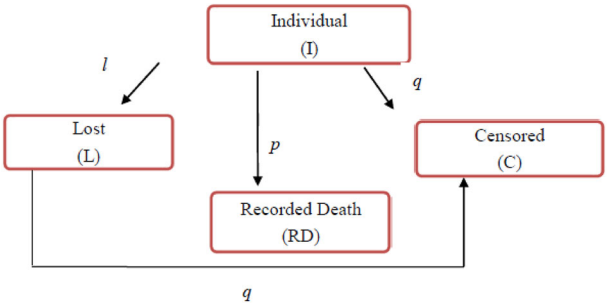

Fig. 6 The Markov model with censoring for the Nanakaly data

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.

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 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. For a detailed rationale of the modelling approach and explanations of parameter choices in this case, see Raza & Broom [12]. 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.

Here for the Hewa data set there were thus two problems with the data, the problem of "lost" individuals, but also the problem of the absence of definitive times of deaths. 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.

Figure 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; $\Omega = [\mathrm{I}, \mathrm{L}, \mathrm{RD}, \mathrm{D}]$.

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. 5 for the model without censoring and Fig. 6 for the model with censoring. For the second Markov chain structure model for the Hewa data (Figs. 7 and 8) we need

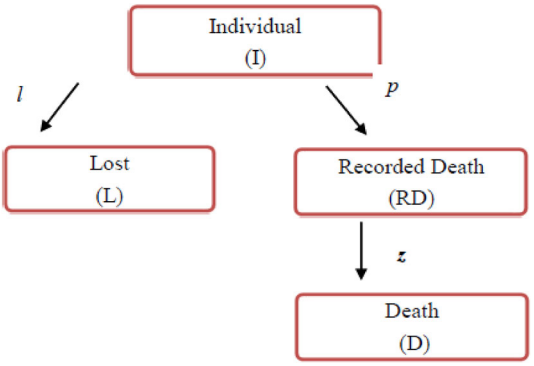

Fig. 7 The Markov model without censoring for the Hewa data

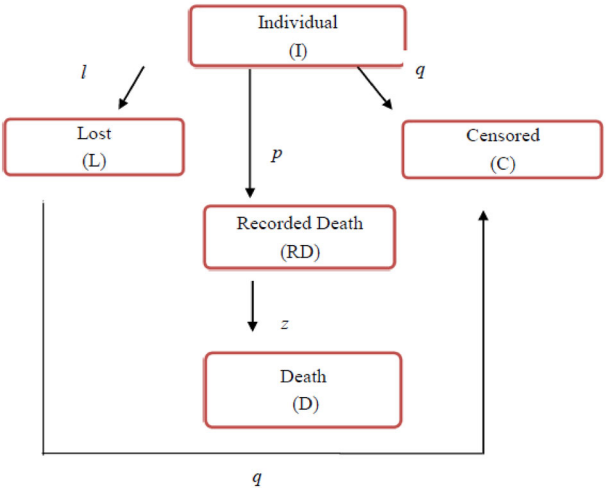

Fig. 8 The Markov model with censoring for the Hewa data

to estimate crucial parameter values and for appropriately chosen estimates we get realistic survival function curves [12].

The adjusted data plots in Sect. 3 use the models from Figs. 6 and 8 to correct for the problems outlined above.

3 Comparisons Between Data

We now compare the data using both the original unadjusted survival curves, and the new adjusted survival curves, where the adjustments follow the methodology described above. In Sect. 3.1 this will be a descriptive analysis, which will then be followed by statistical analyses in Sect. 3.2.

3.1 Survival Curve Comparisons

In this section we plot a series of survival curves for the Kurdish unadjusted and adjusted data, as well as the German data, which needed no adjustment. The outcomes of our investigation are shown for the adjusted data and summarised in Table 1.

3.1.1 Comparisons Between German and Nanakaly Data for Survival Analysis

Here we perform further data analysis for comparison purposes and apply it to the German patients and the Kurdish Hospital patients from Nanakaly. There are 713 patients at Nanakaly Hospital and 686 patients in the German data. The cumulative survival functions for the German and unadjusted Nanakaly data are shown in Fig. 9. In this case, the green line represents the Nanakaly data and the blue line represents the German data, indicating that the German patients had a higher chance of surviving than the Nanakaly patients. Beginning at day zero, it is evident that the German patients have a higher survival rate than the Nanakaly patients until day 2000. However, from days 2000–2500, the Nanakaly patients' apparent chances of surviving are very similar. This is due to hidden censoring, where patients have left the study without this being recorded, so that their deaths would not be observed. The overall real rate of death is then much more than that which is observed.

When comparing the blue line from Fig. 9, the cumulative survival function for the German data and Fig. 10, the adjusted cumulative survival curve for the Nanakaly data from the first model as shown in [11], it is clear that the cumulative survival curve for the German data is higher than that for the Nanakaly data, likely largely due to the superior German health care system. For German data the value of the cumulative survival function up to 100 days inclusive is equal to 0.999 while for the adjusted Nanakaly data it is equal to 0.965. Specifically in the German data there were 15 censored individuals and 1 death during the aforementioned period whereas for the adjusted Nanakaly data there are 20 censored individuals and 24 deaths. The cumulative survival function on day 500 is 0.956 and the adjusted Nanakaly counterpart is 0.679. The number of censoring events and deaths of patients in Germany from day 101 to day 500 inclusive is 21 and 28 patients while in the adjusted Nanakaly data for the same period there are 104 censored individuals and 152 deaths. For the last section, the cumulative survival function on day 1000 is 0.841 for the German data and 0.333 for the adjusted Nanakaly data. There were a further 105 censored individuals and 70 deaths for the German data and 109 censored individuals and 63 deaths in the adjusted Nanakaly data from day 501 to day 1000.

Figure 11, the German data is represented by the blue line, while the Nanakaly data is represented by the green line. From zero to 1000 days, the cumulative survival of the patients in the German data was higher than that of the Nanakaly adjusted data; the cumulative survival values were 0.84 and 0.33, respectively.

3.1.2 Comparing the German and Hewa Data for Survival Analysis

Now we compare the German and Hewa data. There are 686 and 1163 patients in the German and Hewa data sets, respectively. Figure 12 shows the cumulative survival

Table 1 Figure summary for the adjusted data outcomes

Figures	Figure content	Figure explanation
Figure 10	Nanakaly data	The cumulative survival function value at 100 days is 0.96. At day 500, the value is 0.67. Cumulative survival at day 1000 is 0.33
Figure 11	German and Nanakaly data	The green line represents the Nanakaly data and the blue line represents the German data. At 100 days the cumulative survival value for the German data is 0.99, with that for the Nanakaly data in same period being 0.97. Survival at 500 days for the German and Nanakaly data is 0.96 and 0.69, respectively. Finally at 1000 days the survival values for the German and Nanakaly data are 0.81 and 0.33, respectively
Figure 13	Hewa data	The cumulative survival function value at 100 days is 0.97. At day 500, the value is 0.86. Cumulative survival at day 700 is 0.70
Figure 14	German and Hewa data	The blue line represents the German data and the green line represents the Hewa data. At 500 days the cumulative survival values for the German and Hewa data are 0.98 and 0.81, respectively. At 800 days, survival for German and Hewa data is 0.89 and 0.72, respectively
Figure 16	German, Nanakaly and Hewa data	The red line represents Hewa patients, the green line represents Nanakaly patients and the blue line represents German patients. Survival for the German patients was higher than for the two Kurdish groups; cumulative survival values at time 500 for the German, Hewa and Nanakaly data are equal to 0.98, 0.81 and 0.71, respectively. Survival at 800 days is 0.89, 0.72 and 0.61 respectively
Figure 18	Nanakaly data: age classes	The cumulative survival values at 300 days for age classes less than or equal to 48 years and greater than 48 years are 0.87 and 0.81, respectively. Survival at 1000 days for those aged less than or equal to 48 years and greater than 48 years is 0.49 and 0.10, respectively
Figure 19	German data: age classes	The cumulative survival function at 700 days for both age classes is 0.90 and survival at 2600 days for those aged less than or equal to 48 and for those aged greater than 48 years old is 0.69 and 0.51, respectively
Figure 21	Hewa data: age classes	The cumulative survival function at 200 days for both age classes is equal to 0.90, but at 600 days survival for those aged less than or equal 48 years is higher than for those greater than 48 years old, at 0.85 and 0.80 respectively

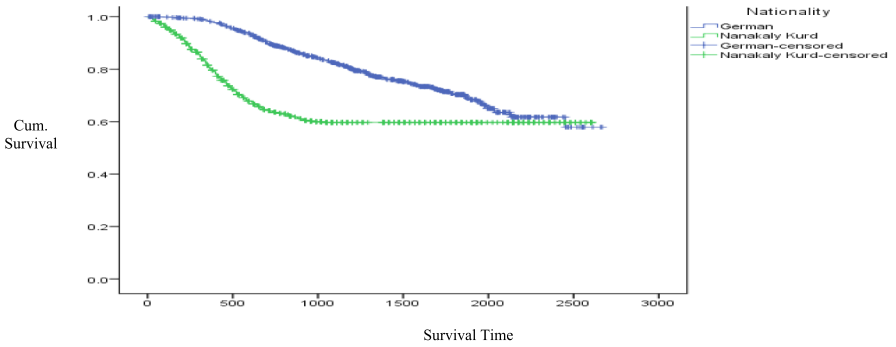

Fig. 9 Cumulative survival function curves for German and Nanakaly unadjusted data

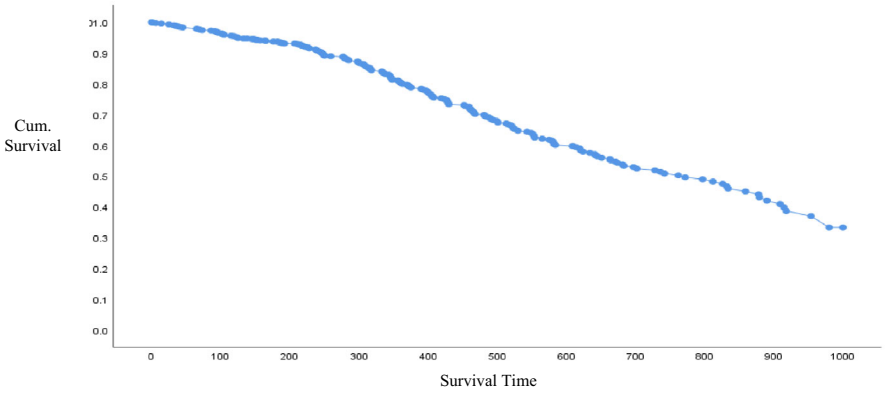

Fig. 10 Cumulative survival function curve for the Nanakaly adjusted data

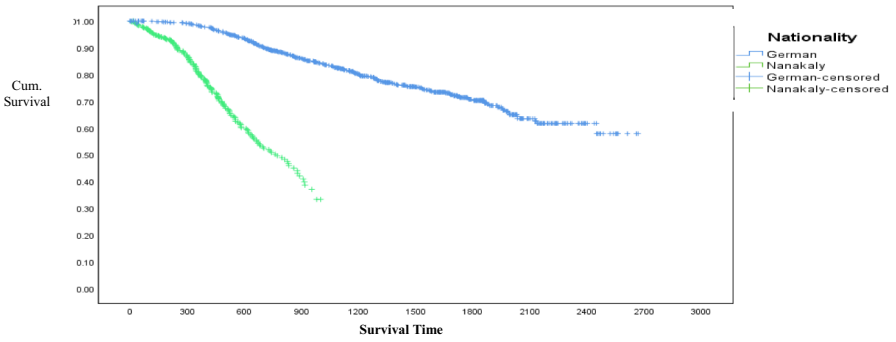

Fig. 11 Cumulative survival function curve for the German and Nanakaly adjusted data

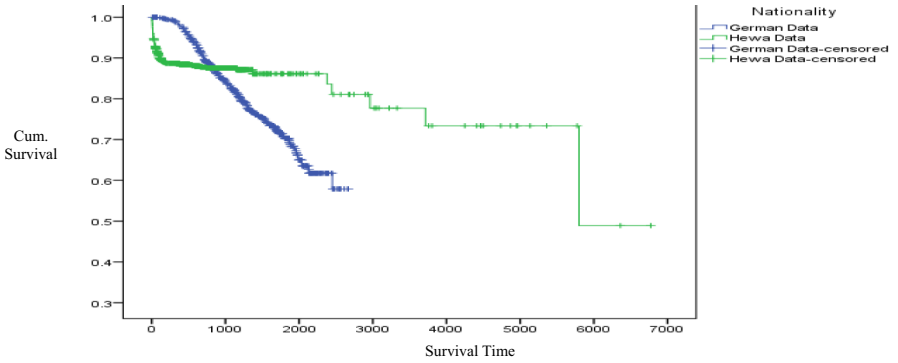

Fig. 12 Cumulative survival function curve for German and Hewa unadjusted data

functions for the German and the unadjusted Hewa data.

The green line, which represents the Hewa unadjusted patient data, intersects with the blue line, which represents the German patients, at 1000 days. The blue line gradually decreases from zero to 3000 days. The patients in both institutions apparently had an equal chance of surviving, as indicated by the cumulative survival function being almost equal to 0.9. However, above 1000 days, the cumulative survival function for the green line is higher than the blue, indicating that the Hewa patients apparently had a higher chance of surviving than the German patients. This figure, however, gives a very misleading impression of survival probabilities for the Hewa patients because of the hidden censoring commented upon previously. Actual death rates were not this low; the number of observed deaths greatly underestimated the total number of real deaths from patients who had mostly left the study. We see a truer picture in Fig. 14 below.

In general the cumulative survival function for the Hewa data is lower than for the German data as illustrated in Figs. 4 and 13. The value of the cumulative survival

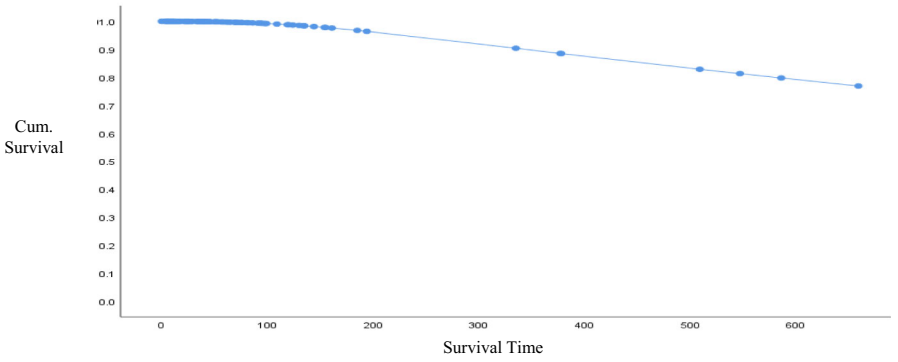

Fig. 13 Cumulative survival function curve for the Hewa adjusted data

function up to 100 days for the German data is equal to 0.999 while for the adjusted Hewa data it is equal to 0.971. At day 500 the cumulative survival function value is 0.956 for the German data and 0.866 for the adjusted Hewa data. Lastly the cumulative survival rate at day 700 is 0.902 for the German data and 0.707 for the adjusted Hewa data. The first adjusted survival curve model for the Hewa data is discussed in [12].

Figures 4, 10 and 13 represent the cumulative survival function for the German data with the adjusted Nanakaly data and the adjusted Hewa data. The value of the cumulative survival function at 100 days for the German data is equal to 0.999 while for the adjusted Nanakaly data it is equal to 0.965 and for the adjusted Hewa data it is equal to 0.971. At day 500 the cumulative survival function value is 0.956 for the German data, 0.680 for the adjusted Nanakaly data and 0.866 for the adjusted Hewa data. The cumulative survival probability on day 700 is 0.902 for the German, 0.524 for the adjusted Nanakaly data and 0.707 for the adjusted Hewa data. In general the survival curve for the German data is higher than for both the adjusted Nanakaly and adjusted Hewa data. On the other hand the survival function curve for the Hewa data is higher than that for the Nanakaly data. This is likely to be because the data from Hewa is not reliable, due to the fact that we do not have the real time of death, rather than representing a real large difference. Note that the first adjusted survival curve model for the Nanakaly data is discussed in [11] and the model for the Hewa data is discussed in [12].

The cumulative survival function for the German and Hewa-adjusted data is displayed in Fig. 14. From 0 to 1000 days, the blue line, representing the German data, indicates that patients had a higher chance of surviving than the Hewa patients (the green line). For the Hewa-adjusted data, the cumulative survival function values are equal to 0.90 and 0.70, respectively. Here, in contrast to Fig. 12, we can see that the survival rates for the Hewa patients are consistently below those of the German patients.

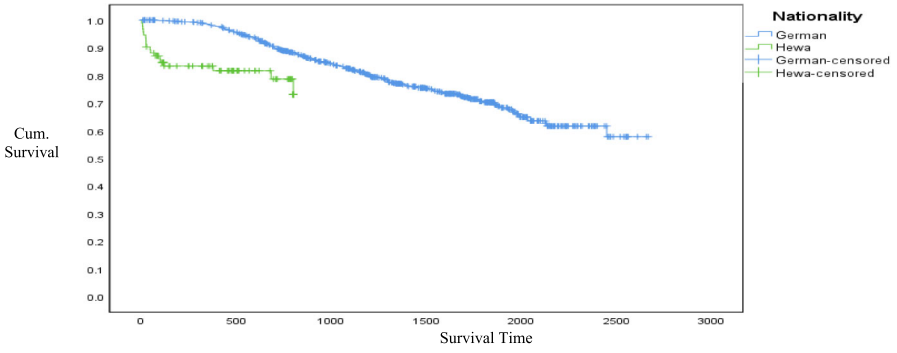

Fig. 14 Cumulative survival function curve for the German and Hewa adjusted data

3.1.3 Age-Based Analysis: Connections Between German, Nanakaly and Hewa Data for Survival Analysis

Finally, we consider a more detailed comparison between the three sets of data, German, Nanakaly and Hewa. There were 686, 713 and 1163 patients in the German, Nanakaly and Hewa hospitals, respectively. Here, we also make use of information we have on the age of participants in the different studies (for the Hewa and German datasets there were many categories, but age was the only additional one for the Nanakaly data). Figure 15 illustrates the cumulative survival function for three data sets, where the Nanakaly and Hewa data are unadjusted. The German data is represented by the blue line, the Nanakaly unadjusted data by the green line, and the Hewa unadjusted data by the red line. The cumulative survival function for the German patients is higher than that of the Nanakaly unadjusted patients from 0 to 2000 days, but after 2500 days, the cumulative survival function for Nanakaly unadjusted patients is marginally higher than that of German patients. However, the apparent cumulative survival function for Hewa unadjusted data, represented by the red line, had a lower value from zero to less than 1000 days than the blue line, but subsequently had a higher value. The values for higher times for the Nanakaly, and especially the Hewa, data are not reliable due to hidden censoring, as we discussed above.

The cumulative survival function for German, Nanakaly, and Hewa adjusted data is shown in Fig. 16; German patients are represented by the blue line, Nanakaly adjusted patients by the green line and Hewa adjusted patients are represented by the red line. Compared to the Nanakaly and Hewa patients, the German patients' had a higher survival probability at 1000 days; the German, Hewa and Nanakaly probabilities are 0.90, 0.75 and 0.40 respectively.

The following Figs. 18 and 19 illustrate the cumulative survival functions for the age class less than or equal to 48 years old and greater than 48 years for the patients in the adjusted Nanakaly and German data respectively (the Nanakaly unadjusted data is shown in Fig. 17). For the age class of less than or equal to 48 years old, the difference between German and adjusted Nanakaly cumulative survival functions is 0.074. On day 600, the German survival function is higher than the adjusted Nanakaly function

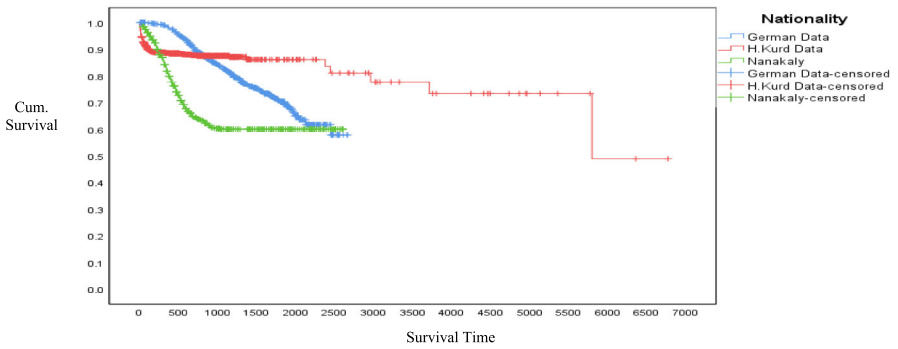

Fig. 15 Cumulative survival function curves for the German, Nanakaly and Hewa unadjusted data

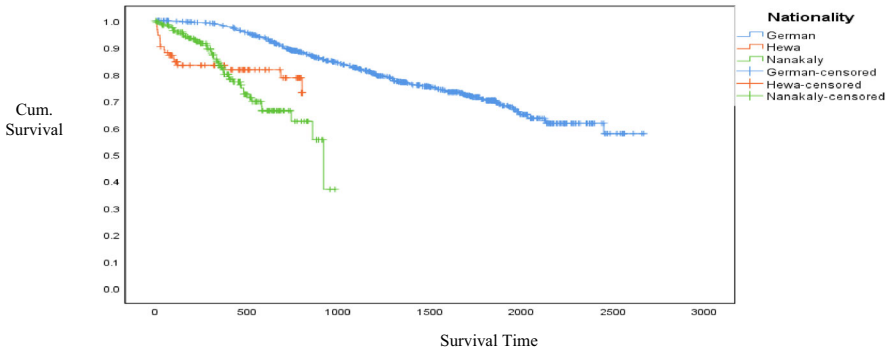

Fig. 16 Cumulative survival function curves for the German, Nanakaly and Hewa adjusted data

by 0.297 for the same age class. Finally, up to 800 days, the cumulative survival function for the German data exceeds its adjusted Nanakaly counterpart by 0.333. Whilst when comparing the age class greater than 48 years old, for the same time points, the difference between German data and the adjusted Nanakaly data is 0.088, 0.409 and 0.497 respectively, with the German patients again on the higher side.

On the other hand, for the unadjusted Hewa data the older age group (greater than 48 years of age) has a longer survival time, though as we see from Table 2, the p-value is 0.935, so not significant. We also recall that from before that the Hewa data is not fully reliable, as we do not have the real time of death and there is no real reason to believe that the pattern in patients in this data set should be markedly different from that in either the Nanakaly or German data sets. The cumulative survival function for the unadjusted Hewa data for the two age classes is shown in Fig. 20, again split into the same two groups.

Figures 18, 19 and 21 display the cumulative survival functions for the age class less than or equal to 48 years old and greater than 48 years old for the patients in the German, adjusted Nanakaly and Hewa data respectively. For the age class of less than

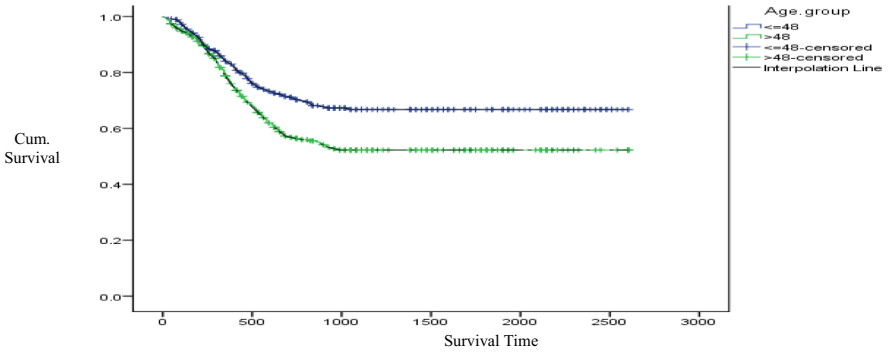

Fig. 17 Cumulative survival function for age less than equal and greater than 48 years for the unadjusted Nanakaly data

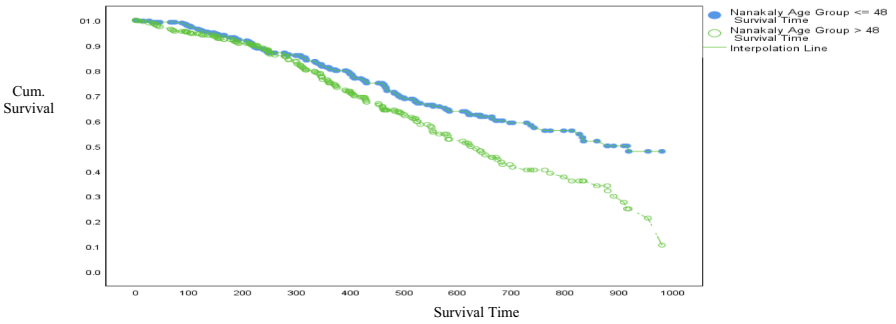

Fig. 18 Cumulative survival function for age less than or equal to and greater than 48 years for the adjusted Nanakaly data

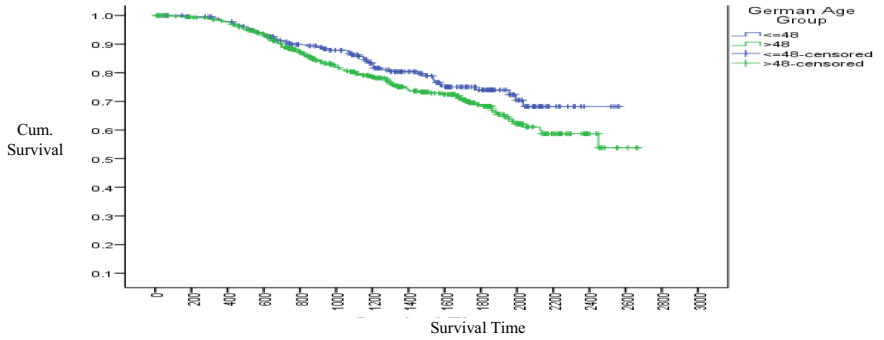

Fig. 19 Cumulative survival function for age less than or equal to and greater than 48 years for the German data

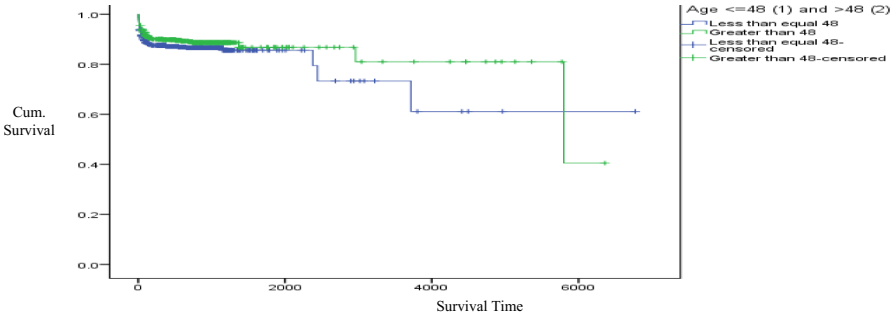

Fig. 20 Cumulative survival function for age less than or equal to and greater than 48 years for the Hewa data for the unadjusted data

Table 2 Cox regression model for age using unadjusted and adjusted Nanakaly and Hewa data and German data

Ages	Hazard Rate (B)	SE	Wald	Df	*P*. value	Exp(B)	Lower 95% CI for Exp(B)	Upper 95% CI for Exp(B)
Unadjusted Nanakaly	0.032	0.005	33.25	1	0.001	1.032	1.021	1.043
Unadjusted Hewa	0.001	0.013	0.07	1	0.935	1.001	0.976	1.027
German	0.010	0.008	1.663	1	0.197	1.010	0.995	1.026
Adjusted Nanakaly	0.029	0.005	30.741	1	0.000	1.029	1.019	1.040
Adjusted Hewa	− 0.006	0.007	0.749	1	0.387	0.994	0.980	1.008

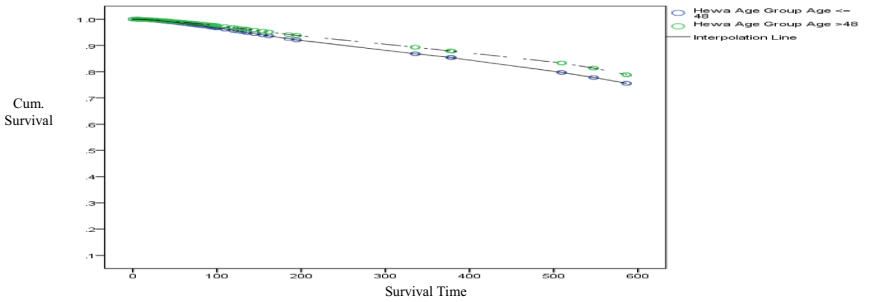

Fig. 21 Cumulative survival function for age less than or equal to 48 years and greater than 48 years for the Hewa adjusted data

or equal to 48 years old, the cumulative survival probabilities for the German, and both the adjusted Nanakaly and Hewa data are given for three different periods and they are; at 250 days where there are equal to 0.996, 0.871 and 0.921 for German, Nanakaly and Hewa data respectively. At 500 days the corresponding values are equal to 0.961, 0.693 and 0.853. Finally, at 700 days the cumulative survival probabilities are 0.908, 0.592 and 0.691 respectively. Comparing the age class greater than 48 years old, for the same time points the cumulative survival probability for 250 days shows 0.993, 0.865 and 0.938 respectively. The values for 500 days are 0.953, 0.626 and 0.878 respectively and the final values for 700 days are 0.896, 0.427 and 0.645 respectively.

3.2 Statistical Analyses

In this section we consider some analytical statistical approaches to the data. Whilst a wide range of methods could be applied to the complete German data, there are issues

with even the simplest approaches for the Iraqi data, and we have limited ourselves to those which are simple and relatively robust. The existing methodology that we developed can create a survival curve, but the typical analytical methods that underlie a direct comparison are not available to us. When comparing more than one dataset, the standard log rank test assumes that censoring is not correlated with the category of the data, for example, and this is far from true in our case.

3.2.1 Parametric Versus Non-parametric Estimation in Survival Analyses with Missing Data

In this paper we have concentrated on non-parametric methods for survival analysis. Typically, approaches are more adaptable and reliable than parametric approaches when the underlying distribution of survival times is unclear or complicated and when there is missing or censored data, which is a critical feature of our data. Results using parametric approaches may be skewed if certain assumptions, including about the particular distribution concerned, are broken. Since non-parametric techniques, such as the Kaplan-Meier estimator, do not make these kinds of assumptions, they are usually more trustworthy when handling missing data and possible departures from presumptive distributions.

Censoring, in which the precise survival duration is unknown (for example, a patient is still alive at the end of the research), is a common feature of survival analysis. While parametric approaches necessitate adjustments to their distribution assumptions, non-parametric approaches deal with censoring more naturally. Without making firm assumptions about the mechanism underlying missing data, they can be used in conjunction with other techniques. We are dealing with a situation where both the standard parametric and non-parametric methods fail, which is why in earlier work we developed methods appropriate to our data. Due to the above advantages of non-parametric methods, the models we developed were non-parametric in nature.

We acknowledge that at the extreme parts of our data sets, where the majority of data is effectively missing, there might be some advantage in considering parametric methods that could use the bulk of the data to predict the parameter values. However, no methods exist that would be appropriate to the data that we consider. In particular, any parametric estimates would be most sensitive to error at precisely this part of the survival curve. Developing parametric versions of our approach would be more complex than the non-parametric ones and we have not as yet attempted to do this.

3.2.2 Cox regression Analysis for Single data sets

In this section we consider the effect of a patient's age on their survival for our different data sets (bearing in mind again that the unadjusted values for the Nanakaly and Hewa data are not reliable). Here each data set is considered individually.

In Table 2, B is the hazard rate in the Cox regression model's variable parameters, SE is the standard error, Wald is the statistic that determines whether a given variable's beta (B) coefficient deviates statistically significantly from 0, Df is the degree of freedom, P is the significant value, $Exp(B)$ is the estimated parameter. When it comes to the

estimated parameters at a 95% confidence interval, the lower 95% *CI* for *Exp(B)* is the lower bound, and the higher 95% *CI* for *Exp(B)* is the upper bound.

The outcome indicates that the age variable for both the Unadjusted Nanakaly and Adjusted Nanakaly data models is significant, with a p-value of less than 0.05 compared to the other models. The age variable for the unadjusted Nanakaly data has highly statistically significant coefficients, according to the Wald statistic. For the unadjusted Nanakaly patient data, the hazard ratio may be expressed as follows, for example:

$\varphi(x) = \exp(0.032\,x)$

where x represents an individual's age. The positive value of the hazard ratio indicates that there is a greater risk with higher age. For example $\exp(0.032 * 21.66) = 2$ means that an age difference of years doubles the risk.

As we see in Table 2, the estimates and confidence intervals of B for the five data sets are:

Unadjusted Nanakaly: 0.032 and (0.021, 0.042);

Unadjusted Hewa: 0.001 and (− 0.024, 1.027);

German: 0.010 and (− 0.005, 0.026);

Adjusted Nanakaly: 0.029 and (0.019, 0.040);

Adjusted Hewa: − 0.006 and (− 0.020, 0.008).

We consider only the adjusted data below, as these are the estimates that are the appropriate ones following our methodology. We obtain estimates and confidence intervals for the differences between B values for the different data sets as follows:

Nanakaly-German: 0.019 and (0.000, 0.037);

Hewa-German: − 0.016 and (− 0.037, 0.005);

Nanakaly-Hewa: 0.035 and (0.017, 0.052).

We see from the above individual analyses that only the Nanakaly data (both adjusted and unadjusted) display a significant element of age risk and that this is borne out by the comparisons where the difference between the Nanakaly and Hewa hazard rates in clearly significant, Nanakaly versus German is borderline significant and Hewa versus German is not significant. As postulated elsewhere, this is likely due to there being greater age risks in Iraqi hospitals compared to that in German hospitals combined with reliability issues with the Hewa data, which were more profound and difficult to adjust for than for the Nanakaly data.

In the figures preceding this section we considered a split in the data between individuals above and below 48 years of age. This was chosen because it is the median of the ages in both the Nanakaly and Hewa studies. For example, Fig. 17 shows the cumulative survival function for the unadjusted Nanakaly data for the age classes less than or equal to 48 years old and greater than 48 years old. To determine which age group has a longer survival time, the two survival curves were compared. We can see that the results observed for these figures, with differences for the Nanakaly data sets of larger order than for the others, is consistent with the results above.

3.2.3 Cox regression Analysis for Combined Data Sets

In this section we now consider combined datasets. We do this by using a Cox regression model for the four possible combinations of the three data sets (using the adjusted

Table 3 Cox regression model for age using Nanakaly and Hewa data and German data combinations

							95.0% CI for Exp(B)	
Age	(B)	SE	Wald	df	Sig	Exp(B)	Lower	Upper
Nanakaly, Hewa and German	0.008	0.004	4.329	1	0.037	1.008	1.000	1.015
Hewa and German	− 0.002	0.005	0.178	1	0.673	0.998	0.988	1.008
Nanakaly and German	0.015	0.005	11.411	1	0.001	1.015	1.006	1.024
Nanakaly and Hewa	0.013	0.004	9.605	1	0.002	1.013	1.005	1.0

data only for Nanakaly and Hewa hospitals). The results for each of these are shown in Table 3.

In Table 3 we can see that the combined estimated value of our hazard parameter B is significantly different from zero in all three of the four combinations that contain the Nanakaly data. This is consistent with the earlier results where the Nanakaly data showed a significant age-related effect, but the other two data sets did not.

4 Discussion

In this paper we considered the comparison of breast cancer survival data between hospitals in the Kurdistan region of Iraq, Nanakaly and Hewa, and those from a German hospital. The models in this paper are built on those from our earlier work, [11, 12] 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 earlier model there was 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, as this was a problem present in the Hewa dataset. The models produced adjusted survival curves to take the data problems into account and give a better prediction of true survival rates. The original and new works both consider the variable of patient age. The purpose of this study is to compare patient survival between the age variable data sets from the Nanakaly and Hewa hospitals, using the adjusted models, with the original age variable data from the German hospital.

For the three hospitals we considered the age variable in two different ways. Firstly, in a qualitative analysis we plotted the survival functions for patients younger and older than a cut-off age of 48. This age was chosen as it was the median value of the age for both the Hewa and the Nanakaly data sets. We can see that, as you might expect, patients aged greater than 48 years have lower survival rates from breast cancer than those aged less than or equal 48 years for both the Nanakaly and German data sets,

as shown in Figs. 18 and 19, respectively, with the difference larger for the Nanakaly data.

After adjusting Nanakaly and Hewa data the finding shows the age group of less than or equal to 48 years old, the German cumulative survival probabilities, and the adjusted Nanakaly and Hewa data are provided for three distinct time periods: at 250 days, the German, Nanakaly, and Hewa data, respectively, are equal to 0.996, 0.871, and 0.921. The equivalent values are 0.961, 0.693, and 0.853 at 500 days. The cumulative survival probabilities at 700 days are, in order, 0.908, 0.592, and 0.691. The cumulative survival probability for 250 days for the age class older than 48 is 0.993, 0.865 and 0.938, for 500 days the cumulative survival probability for the same age is 0.953, 0.626 and 0.878 respectively and finally the values for 700 days are 0.896, 0.427 and 0.645 respectively.

Interestingly, for Hewa hospital the apparent survival rates are almost identical, but with the patients over 48 having a slightly higher survival curve than those less than or equal to 48 years old, as we see in Fig. 21. We have two comments to make on this surprising outcome. Firstly, as we see later, the differences are too slight to be significant, as opposed to for the other two data sets. Secondly, and perhaps more importantly, the survival curve for the Hewa data is more unreliable than that for the other data sets. The German data has an accurate unadjusted survival curve, and the adjustment process for the Nanakaly date is more robust given that times of death at least are accurately reported. Thus, whilst our adjustment method works to make the results more accurate, there is still some caution needed when interpreting the results.

To estimate the hazard ratio for differences in the age parameter a Cox's proportional hazards model applied for the adjusted Nanakaly and Hewa data is shown in Table 2. The hazard parameter B was positive and significant for the Nanakaly data, positive but not significant for the German data and (just) negative and not significant for the Hewa data. These results are, not surprisingly, consistent with the discussion above on the three different models.

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 [23]. Early detection is therefore required to improve the outcome of breast cancer and survival 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. At present, WHO together with support from the Komen Foundation is conducting a 5-year breast cancer cost-effectiveness study in 10 low and middle-income countries. The project includes a program-costing tool to assess affordability (WHO, 24). 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 [25]. 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 [26].

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., [27] 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 (Ministry of Health and Environment, [28]). 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 [29, 30]. This combination of trends makes the reported mortality rate one of the highest in the world, according to recent comparative literature [23].

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 recent government initiatives have been launched in an attempt to reverse the tide and improve certain key indicators [30], health outcomes have broadly deteriorated [31] & [32]).

The breast cancer early detection (BCED) programme has been 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 [33]. At the same time, the proportion of advanced stage (III & 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 tracked heavily with socioeconomic status, suggesting the necessity of further outreach programs targeting poorer and less literate communities [34].

In conclusion, despite challenges to related to missing data, we have managed to obtain a picture of breast cancer survival rates in the Kurdistan Region of Iraq in comparison to those in the west, in particular Germany. We have also been able to highlight the differences in survival rates between younger and older patients. We note that we would be able to make more definitive statements with more accurate data, but our methodology nonetheless highlights some important features of the data.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2024). Global cancer observatory: cancer today. Lyon, france: international agency for research on cancer. Available from: https://gco.iarc.who.int/today, accessed
2. Internet (2021). cited on 18 May 2021. https://www.who.int/news-room/fact-sheets/detail/cancer.index.html
3. Breastcancer.org (2025), Breast Cancer Facts and Statistics. Via internet resited on 3 July 2025. https://www.breastcancer.org/facts-statistics
4. Emmerson J, Brown JM (2021) Editorial understanding survival analysis in clinical trials, *Clinical Oncology* 33 12–14, Leeds, UK)
5. Tableman M, Kim JS (2004) Survival analysis using S. Chapman & Hall/CRC, New York
6. Majid R, Mohammed H, Hassan H, Abdulmahdi W, Rashid R, Hughson M (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, Vol. 12*, No. 16, pp. 1–10. Accessed from http://www.biomedcentral.com/1472-6874/12/16
7. Compton CC, Byrd DR, Garcia-Aguilar J, Kurtzman S, Olawaiye HA and Washington MK (2012). Cancer Survival Analysis. 21 September 2013 from http://www.springer.com/978-1-4614-2079-8
8. Robins J (1992) Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. Biometrika 79(2):321–334. https://doi.org/10.1093/biomet/79.2.321
9. Ziaei J, Zohreh S, Iraj A, Saeed D, Ali P, Jalil V (2013) Survival analysis of breast cancer patients in Northwest Iran. Asian Pac J Cancer Prev 14(1):39–42
10. Lambert PC, Royston P (2009) Further development of flexible parametric models for survival analysis. Stata J 9(2):265–290
11. Raza MS, Broom M (2016) Survival analysis modeling with hidden censoring. J Stat Theory Pract 10(2):375–388
12. Raza MS, Broom M (2023) The use of survival analysis modelling with incomplete data with application to breast cancer. Asian J Prob Stat 25(ISSUE, 3.):45–69
13. Hamdan H, Garibaldi JM (2009) Modelling survival prediction in medical data. *Intelligent Modelling and Analysis (IMA) Research Group.* University of Nottingham:UK
14. Miller LD, Johanna S, Joshy G, Vinsensius B, Liza V, Alexander P, Yudi P, Per Hall S, Edison T, Jonas B (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 102(38):13550–13555
15. Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C et al (2011) Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the international cancer benchmarking partnership): an analysis of population-based cancer registry data. Lancet 377:127–138
16. Corazziari I, Quinn M, Capocaccia R (2004) Standard cancer patient population for age standardising survival ratios. *Eur J Cancer*. Vol. 40, No. 15, pp. 2307–2316. https://doi.org/10.1016/j.ejca.2004.07.002

17. Lambert PC, Dickman PW, Rutherford MJ (2015) Comparison of different approaches to estimating age standardized net survival. BMC Med Res Methodol. https://doi.org/10.1186/s12874-015-0057-3
18. Pohar Perme M, Stare J, Estève J (2012) On estimation in relative survival. Biometrics 68:113–120. https://doi.org/10.1111/j.1541-0420.2011.01640.x
19. Lambert PC, Dickman PW, Nelson CP, Royston P (2010) Estimating the crude probability of death due to cancer and other causes using relative survival models. Stat Med 29:885–895
20. Dickman PW, Sloggett A, Hills M, Hakulinen T (2004) Regression models for relative survival. *Stat Med*. Vol. 23, No. 1, pp. 51–64. https://doi.org/10.1002/sim.1597
21. Royston P, Altman DG (2013). External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. Mar 6;13:33. https://doi.org/10.1186/1471-2288-13-33. PMID: 23496923; PMCID: PMC3667097
22. Leung KM, Elashoff RM, Afifi AA (1997) Censoring issues in survival analysis. Annu Rev Public Health 18:83–104
23. IARC (2020b) Estimated age-standardized mortality rates (World) in 2020, breast, females, all ages. CANCER TODAY. Lyon: International Agency for Research on Cancer. Retrieved July 28, 2022, from: https://gco.iarc.fr/today
24. World Health Organization (WHO) (2008). The global burden of disease: 2004 update. From www.who.int/evidence/bod ISBN 9789241563710
25. Lyons G, Sankaranarayanan R, Millar A, Slama S (2018) Scaling up cancer care in the WHO Eastern Mediterranean Region. East Mediterr Health J 24(01):104–110
26. Duggan C, Dvaladze A, Rositch AF, Ginsburg O, Yip C, Horton S, Camacho Rodriguez R, Eniu A, Mutebi M, Bourque J, Masood S, Unger-Saldaña K, Cabanes A, Carlson RW, Gralow JR, Anderson BO (2020). The breast health global initiative 2018 global summit on improving breast healthcare through resource-stratified phased implementation: methods and overview cancer. Vol. 126, No. S10, pp. 2339–2352
27. Mahdi H, Mula-Hussain L, Ramzi ZS, Tolba M, Abdel-Rahman O, Abu-Gheida I, Khorshid O, Al Sukhun S, Siddiqi NP, Al Mandhari Z, Al Hussaini M (2022) Cancer burden among Arab-World females in 2020: working toward improving outcomes. JCO Glob Oncol 8:e2100415. https://doi.org/10.1200/GO.21.00415
28. Ministry of Health and Environment, Republic of Iraq (2021). Annual Report Iraqi Cancer Registry 2020. Iraqi Cancer Board. Retrieved on 25 August, 2022 from: https://moh.gov.iq/upload/2991322580.pdf
29. Alwan N (2022) General Oncology care in Iraq. Al-Shamsi H, Abu-Gheida I, Iqbal F. Springer, Al-Awadhi A. Cancer in the Arab World. Singapore, pp 353–362
30. Alwan N, Kerr D (2018) Cancer control in war-torn Iraq. Lancet Oncol 19(3):291–292
31. World Health Organization (WHO) (2022) The Global Health Observatory, Health workforce. Retrieved 18 July, 2022 from: https://www.who.int/data/gho/data/themes/topics/health-workforce
32. Physicians for Human Rights. (2021) Challenges Faced by the Iraqi Health Sector in Responding to COVID-19. Retrieved July 28, 2022, from https://phr.org/our-work/resources/challenges-faced-by-the-iraqi-health-sector-in-responding-to-covid-19/
33. Alwan NAS, Tawfeeq FN, Maallah MH, Sattar SA, Saleh WA (2017) The stage of breast cancer at the time of diagnosis: correlation with the clinicopathological findings among Iraqi patients. J Neoplasm 2(3):22
34. Husamaldien L, Dauod A (2016) Knowledge, awareness and practices about breast and cervical cancer in a group of women in Erbil city-Iraq. Tikrit Med J 21(2):54–66