Int J Clin Exp Med 2011;4(2):148-156
www.ijcem.com /ISSN:1940-5901/IJCEM1102003

Original Article

Application of hazard models for patients with breast cancer in Cuba

Anet García Alfonso[1], Néstor Arcia Montes de Oca[2]

[1]National Center Coordinators of Clinical Trials, [2]National Statistics Office, Cuba.

Received February 15, 2010; accepted March 29, 2011; Epub April 6, 2011; published May 15, 2011

Abstract: There has been a rapid development in hazard models and survival analysis in the last decade. This article aims to assess the overall survival time of breast cancer in Cuba, as well as to determine plausible factors that may have a significant impact in the survival time. The data are obtained from the National Cancer Register of Cuba. The data set used in this study relates to 6381 patients diagnosed with breast cancer between January 2000 and December 2002. Follow-up data are available until the end of December 2007, by which time 2167 (33.9%) had died and 4214 (66.1%) were still alive. The adequacy of six parametric models is assessed by using their Akaike information criterion values. Five of the six parametric models (Exponential, Weibull, Log-logistic, Lognormal, and Generalized Gamma) are parameterized by using the accelerated failure-time metric, and the Gompertz model is parameterized by using the proportional hazard metric. The main result in terms of survival is found for the different categories of the clinical stage covariate. The survival time among patients who have been diagnosed at early stage of breast cancer is about 60% higher than the one among patients diagnosed at more advanced stage of the disease. Differences among provinces have not been found. The age is another significant factor, but there is no important difference between patient ages.

Keywords: Survival analysis, Akaike information criterion, accelerated failure-time metric, proportional hazard metric

Introduction

Cancer constitutes an important health problem which affects to an increasing number of people throughout the world. In developed countries the malignant tumors are the second cause of death, whereas in developing countries they are among the first five causes of death. Specifically, in Cuba, the malignant tumors in women were the second cause of death with a rate of 162.6 per 100 000 women in 2009 [1].

The breast cancer occupies the first place among the cancer in women in the world. In Cuba, 44.3 per 100 000 women were diagnosed with breast cancer in 2006 and it has kept soaring over time. The mortality rate of this disease also tends to increase as time passed by, for example, in 1970 and 1980 the rates were of 10.2 and 13.3 per 100 000 women respectively, whereas this rate is considerably increased up to 24.2 in 2008 [1]. These numbers demonstrate by themselves the importance of studying such a disease within the Cuban context.

Cuba represents an important alternative example where modest infrastructure investments combined with a well-developed public health strategy have generated health status measures comparable with those of industrialized countries. In Cuba, the aforementioned increasing of mortality rate of breast cancer over time may be explained by two main risk factors: on the one hand, the aging process which behaves similar to developed countries, and, on the other hand, the more comprehensive and earlier detection of breast cancer in the last decade which means that an increasing number of individuals are detected at early clinical stages making possible more effective treatments and therefore higher survival times.

This article aims to assess the overall survival

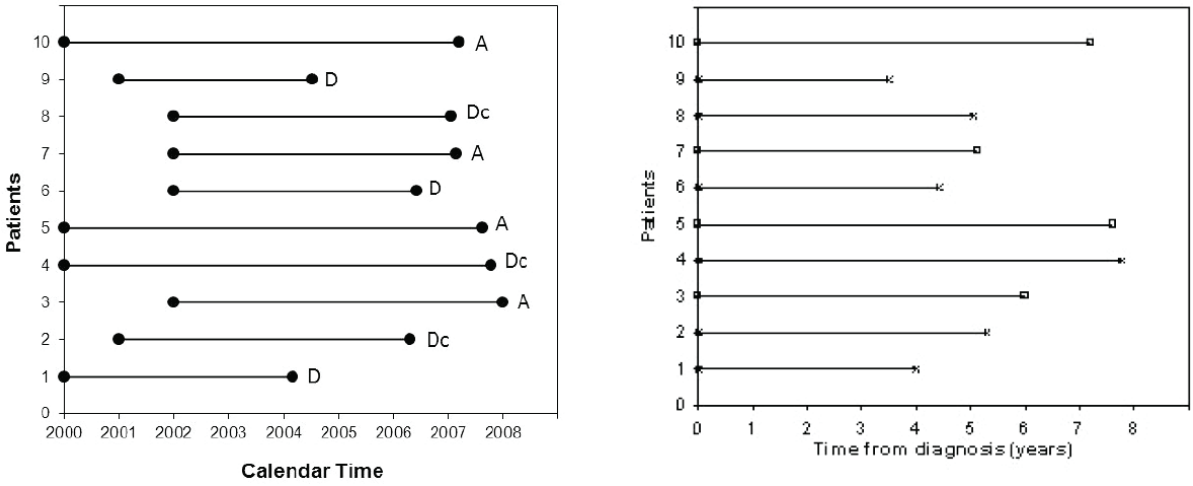

Figure 1. Adapting calendar time in the breast cancer study to a survival analysis format (D=death from breast cancer; Dc=death from other cause; A= still alive; X=death from any cause; □=censored).

time of breast cancer in Cuba, as well as to determine plausible factors that may have a significant impact in the survival time.

The article is outlined as follows. Section 2 describes the survival analysis methods used in this article. Section 3 presents the main results in two different sections: Section 3.1 and section 3.2. Section 3.1 assesses the overall survival time of breast cancer in Cuba and it also describes the survival with respect to two different factors separately: clinical stage and age. Section 3.2 determines the prognostic ability of various factors on overall survival. Finally, section 4 describes some final remarks and suggestions for further research.

Materials and methods

The data set used in this study relates to 6381 patients diagnosed with breast cancer between January 2000 and December 2002. Among them, the diagnostic of 890 (13.9%) patients were unknown. Follow-up data were available up to the end of December 2007, by which time 2167 (33.9%) had died and 4214 (66.1%) were still alive. A patient within the latter situation is often called *right censoring*. The data were obtained from the National Cancer Register of Cuba.

Figure 1 shows data from 10 patients diagnosed in the early 2000s and illustrates how patient profiles in calendar time are converted to time to event (death) data. **Figure 1** (left) shows that three of ten patients have died from breast cancer (patients 1, 6 and 9), three died from other cause different to breast cancer (2, 4, and 8), and four patients (3, 5, 7, and 10) are still alive. In the right-side figure, the data are presented in format for a survival analysis where all-cause mortality is the event of interest. In general, it is a good practice to choose an end-point that cannot be misclassified. All-cause mortality is a more robust end-point that a specific cause of death [2]. Patients 1, 2, 4, 6, 8 and 9 have died from any cause, whereas patients 3, 5, 7, and 10 are still alive by the end of the follow-up study (censored patients). It is important to note that censored patients are assumed to be those who have not been reported as dead by the end of the follow-up study in the mortality data set released by the Ministry of Public Health.

Many statistical methods have been historically used to respond the afore-mentioned objectives, for instance the Kaplan-Meier (KM) plots, logrank tests, and two models for adjusting survival functions for the effects of covariates: the accelerated failure-time (AFT) and the proportional hazard (PH) rate model.

The overall survival probability of breast cancer in Cuba can be estimated nonparametrically from observed survival times by using the KM method [3]. This method allows each patient to contribute information to the calculations for as

long as they are alive. The KM survival curve provides a useful summary of the data that can be used to estimate measures such as median survival time.

Another way to describe and model survival is in term of hazard. The hazard is the probability that an individual who is under observation at a time *t* has an event at that time. In this study, the cumulative hazard function is used to derive the hazard function by applying a kernel smoother to the increments [4]. This function is estimated by using the Nelson-Aalen estimator [5].

In this study, it is also useful to analyse the survival curves for different patient groups and to introduce several tests to investigate differences between them. Specifically, different survival curves for different patient groups have been plotted to verify that the proportionality assumption holds, which means that the survival probability of a specific individual profile with regard to the probability of another one does not change over time. The proportionality assumption is assessed for two variables: the patient age and the patient clinical stage. The patient clinical stage covariate, which is a variable that measures to some extent the clinical degree of the breast cancer diagnosed at the beginning of the patient follow-up, may be an important indicator to investigate the behaviour of the survival curves. This variable has been classified according to the Tumor Node Metastasis (TNM) staging classification for breast cancer [6]. There are five different categories within this covariate; the first one (stage I) corresponds to early stage disease and the stage IV corresponds to the most advanced disease, whereas the stage V means that the patient clinical stage is unknown. Note that all substages, defined in the TNM staging classification, within each clinical stage have been grouped and a new stage has also been incorporated for patients who have an unknown diagnosis.

The test of proportional hazards based on the generalization by Grambsch and Therneau [7] is implemented to analyse proportionality for each covariates. Furthermore, the Wilcoxon test of Breslow [8] based on Wilcoxon [9] is also performed to testing the equality of survivor curve across different groups. This test is appropriate when hazard functions are thought to vary in ways other than proportionally as it is in our application.

This study also aims to determine plausible factors that may have a significant impact in the survival time. For that reason, the adequacy of six parametric models is assessed by using the Cox-Snell residuals and also using their Akaike information criterion (AIC) values [10]. Five of the six parametric models (Exponential, Weibull, Log-logistic, Lognormal, and Generalized Gamma) are parameterized by using the AFT metric, and the Gompertz model is parameterized by using the PH metric. The knowledge of which metric has been used is extremely important to make an appropriate interpretation of the results.

For each model, we include three covariates: province where the patient lives, patient age, and patient clinical stage. A backward elimination procedure [11] was implemented to determine the final number of significant covariates.

Note that, in the accelerated failure-time model, the natural algorithm of the survival time *ln t* is expressed as a linear function of the covariates yielding the linear model: $In\ t_j = x_j\beta + e_j$, where x_j is a vector of covariates, β is a vector of regression coefficients, and e_j is the error with a specific distribution. The distributional form of the error term determines the regression model. The effect of the AFT model is to change the time scale by a factor of $exp(-x_j\beta)$.

Thus, the survival times can be seen to be multiplied by a constant effect under the model specification, and the exponentiated coefficients, $exp(-x_j\beta)$ are referred to as time ratios. A time ratio above 1 for the covariate implies that this prolongs the time to the event, while the time ratio below 1 indicates than earlier event is more likely [12].

In the PH rate model, the covariates have a multiplicative effect on the hazard function $h(t_j) = h_0(t)\ exp(x_j\beta)$, where $h_0(t)$ is a baseline hazard function. A number of different parametric PH models may be derived by choosing different hazard models. The models commonly applied are the Exponential, Weibull or Gompertz models and they take their names from the distribution that the survival times are assumed to follow.

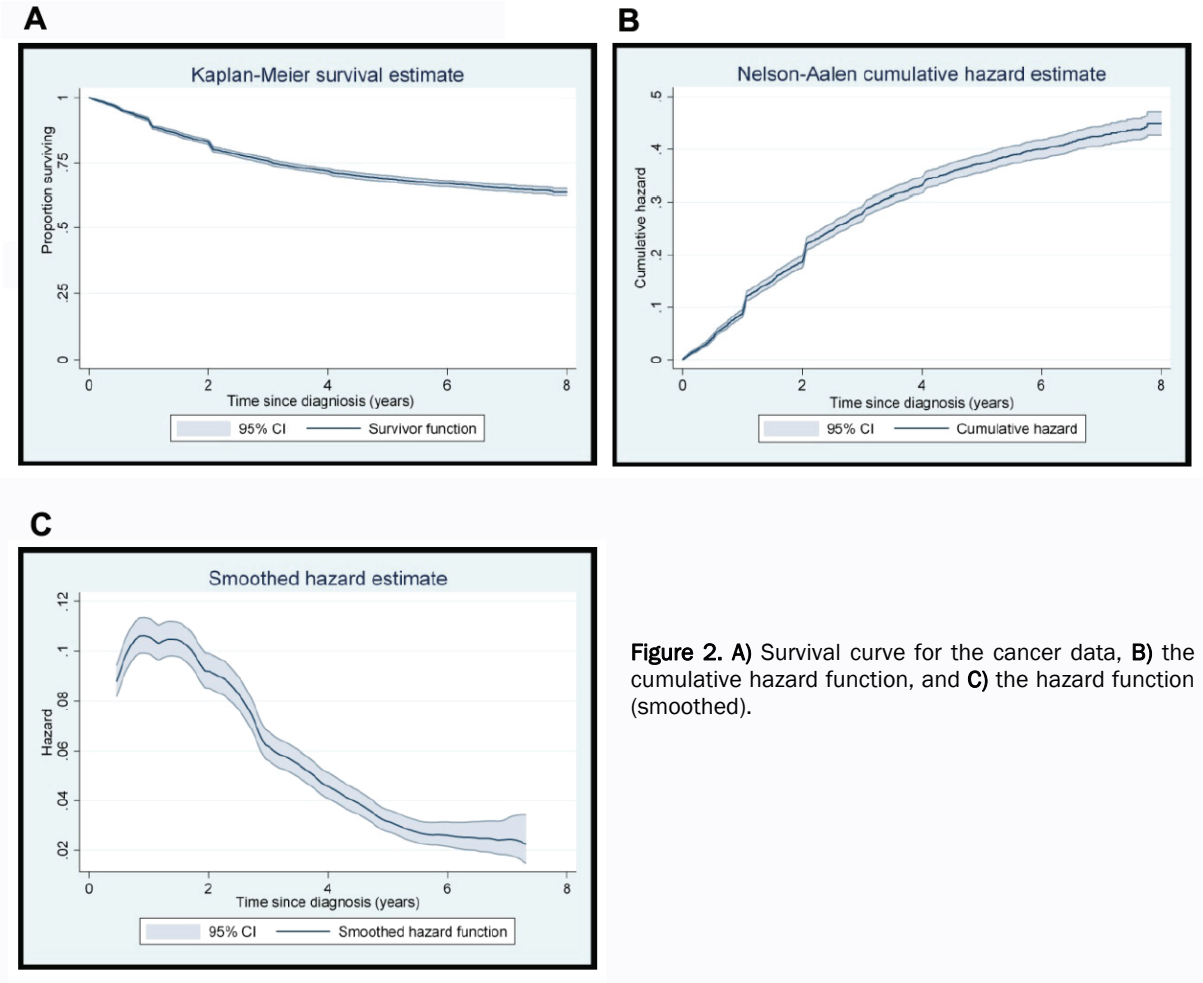

Figure 2. A) Survival curve for the cancer data, **B)** the cumulative hazard function, and **C)** the hazard function (smoothed).

Results

Survival analysis of the breast cancer data

The KM survival curve is shown in **Figure 2A**. Overall, patients have a predicted median survival time of approximately 5 years and 8 months. As expected the curve decreases as long as time increases. The survival probability is nearly 85% for a patient with 1 year of survival time, whilst this probability decreases by roughly 10% for a patient with 5 years of survival time.

The slight steeper decline in earlier years might indicate poor prognosis from the disease. This is also indicated by changes in the cumulative number of deaths and number at risk. Specifically, of the total deaths as recorded by the last date of follow-up (2167 women), about a half had died within the first two years since diagnosis (1101 women), while only about 3% died within the last two years.

The 95% confidence intervals of the survival function are also shown. Such confidence intervals are wide at the tail of the curve because there are patients alive at the end of follow-up. Specifically, of the 6381 woman diagnosed with breast cancer, about 66% were alive at the end of follow-up (right censored). This fact may make meaningful interpretations difficult, although, in our study, such intervals are not too large due mainly to there is still enough sample size to obtain accurate estimates.

Figure 2B shows the cumulative hazard function estimated by using the Nelson-Aalen estimator. **Figure 2C** shows the hazard function which appears to reach a peak in the first year after diag-

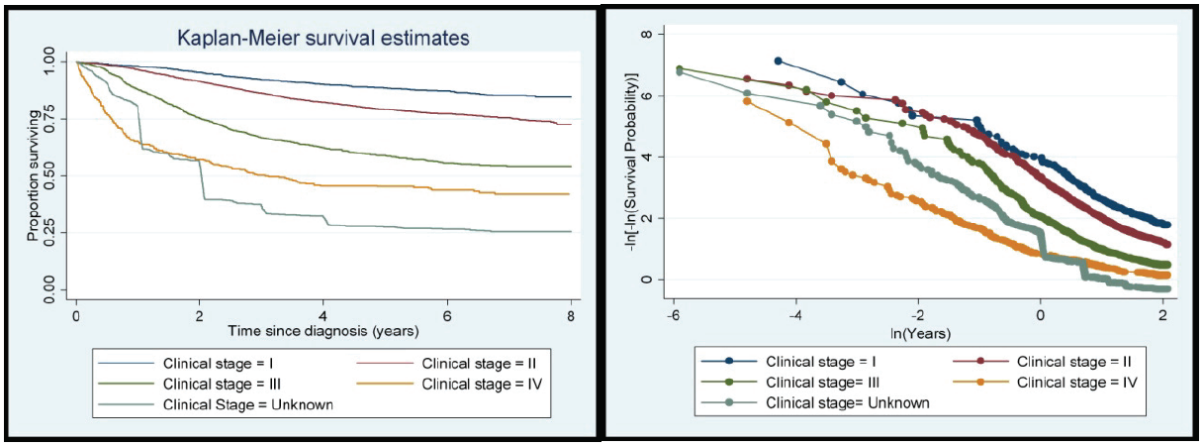

Figure 3. The survival curves (left-side) and the Log-log plots (right-side) by clinical stages.

Table 1. Test of proportional-hazards assumption

Covariates	rho	chi2	Df	Prob>chi2
Clinical stage II	-0.00429	0.04	1	0.8424
Clinical stage III	-0.07623	12.46	1	0.0004
Clinical stage IV	-0.20441	89.09	1	0.0000
Unknown	-0.09911	20.71	1	0.0000
Age	-0.06224	8.55	1	0.0035
Global test		180.80	5	0.0000

nosis and decreases afterward. It is evident that the hazard function is not constant over time, thus an Exponential distribution of the survival time should be ruled out. The shape of the hazard function (**Figure 2C**) suggests that either the log-normal or the Generalised Gamma distributions for the survival time might be preferable to be used [10].

As we pointed out in section 2, it is also useful to analyse the survival curves for different patient groups and to introduce several tests to investigate differences between them. **Figure 3** shows the survival and the log-log plots by clinical stage covariate. Clearly the patients with the least degree of severity of the clinical stage are more likely to survive than any other patient, whereas patients whose clinical stage is unknown are the most likely to die. The percentages of patients who have died within the clinical stage I and II are 4.8% and 8.7% respectively, whereas these percentages increase up to 43.6% and 43.9% for the clinical stage IV and the unknown clinical stage respectively.

It is also obvious that the proportionality assumption does not hold because the survival curves are not parallel (**Figure 3**). This is statistically confirmed by implementing the test of proportional hazards based on the generalization by Grambsch and Therneau [7]. Clearly three dummy variables *"Clinical stage III", "Clinical stage IV", and "Unknown"* created by using the clinical stage covariate and taking as reference category the clinical stage I, violate the proportional hazard assumptions. For age, such an assumption does not hold. Whereas the global test (on 4 degrees of freedom) also violates such an assumption (**Table 1**).

The Wilcoxon test shows that the null hypothesis of equal survival curves is rejected (Pr>chi2 = 0.0000). Thus, we have statistically tested that the survival curves for the clinical stages variable are different and the proportionality assumption does not hold.

Survival analysis adjusting for covariates

One of the aforementioned objectives at the introductory section is to determine the prognostic ability of various factors on overall sur-

Table 2. Akaike Information Criterion (AIC) of six different distributions fitted to the full model

Model	Log Likelihood (LL)	No. of covariates (c)	No. of ancillary parameters (p)	AIC
Exponential	-3155.571	21	0	6355.094
Weibull	-1315.2288	21	1	2676.458
Gompertz	-1782.9864	21	1	3611.973
Log-logistic	-1687.6147	21	1	3421.229
Lognormal	-2237.3019	21	1	4520.302
Generalized Gamma	-921.08302	21	2	1890.166

AIC=-2LL+2(c+p+1)

vival. For example, in our study, the patients with the clinical stage I and II might be surviving longer because of either lower age or the province where they live. In this case, the clinical stage effect is confounded by either the effect of age or the residence place. For that reason, a multivariate analysis for adjusting survival functions is implemented here by taking into account different covariate effects. Specifically, three covariates that were all known at baseline or entry to the study have been considered in this case: 1) clinical stage that it is diagnosed to a patient at the beginning of the study, 2) age, and 3) residence place (province). It is likely that the use of more factors, for example, tumor size, histology type, treatment type, family history [13], strogen receptor [14], oncogene and anti-oncogene [15], may also affect the survival time of a patient. However, all these factors have not been considered in this study because there were not either data available or the gathered information was not reliable.

In general, two models have been frequently used for adjusting survival functions for the effects of covariates: the AFT model and the PH model. As pointed out early, the adequacy of six parametric models (each with all covariates included) are assessed and present their AIC values in **Table 2.** Despite the backward elimination procedure implemented takes the clinical stage and age as significant covariates, we also decided to put into the model the province where the patient lives. This decision was taken because two main reasons: 1) from the clinical infrastructure perspective, we suspect that there is a clear difference between provinces, and 2) from the statistical perspective, the model that includes the province as an additional covariate shows less AIC than the model without this covariate. In addition, as neither the clinical stage covariate nor age holds the proportionality assumption (see Table 1 above), two interactions between covariates have also been included into the model: a linear interaction of time and age, and time and the clinical stage covariate. These interactions allow the effect of these two covariates to change with time and they are also a way of accommodating non-proportional hazards.

Table 2 confirms the suggestion we pointed out early in which the Generalized Gamma distribution for the survival time might be preferable to be used in this case. The Generalized Gamma model has effectively a higher log-likelihood than the other models and a lower AIC, indicating that this distribution may be the most accurate.

The Cox-Snell residuals are also useful in assessing the overall model fit. If the model fits the data, then the plot of the cumulative hazard, based for example on the Kaplan-Meier survival estimates or the Aalen-Nelson estimator, versus the Cox-Snell residuals should be a straight line with slope 1. **Figure 4** indicates that the Generalized Gamma and Gompertz models fit the data best, and the exponential, Weibull, log-logistic, and log-normal fit poorly. The result for the Generalized Gamma model is consistent with our previous result based on Akaike's information criterion. Therefore, the Generalized Gamma model seems to be the best model for our data.

The multivariate effect sizes obtained from the Generalized Gamma model are presented in **Table 3**. As one can be seen, there is not a significant variation of the survival time between provinces which may be a direct consequence of the fact that the public health in Cuba is accessible and free of charge for everyone. The main differences are explained by the clinical stage covariate. For example, the survival time is about 40% shorter among patients who have

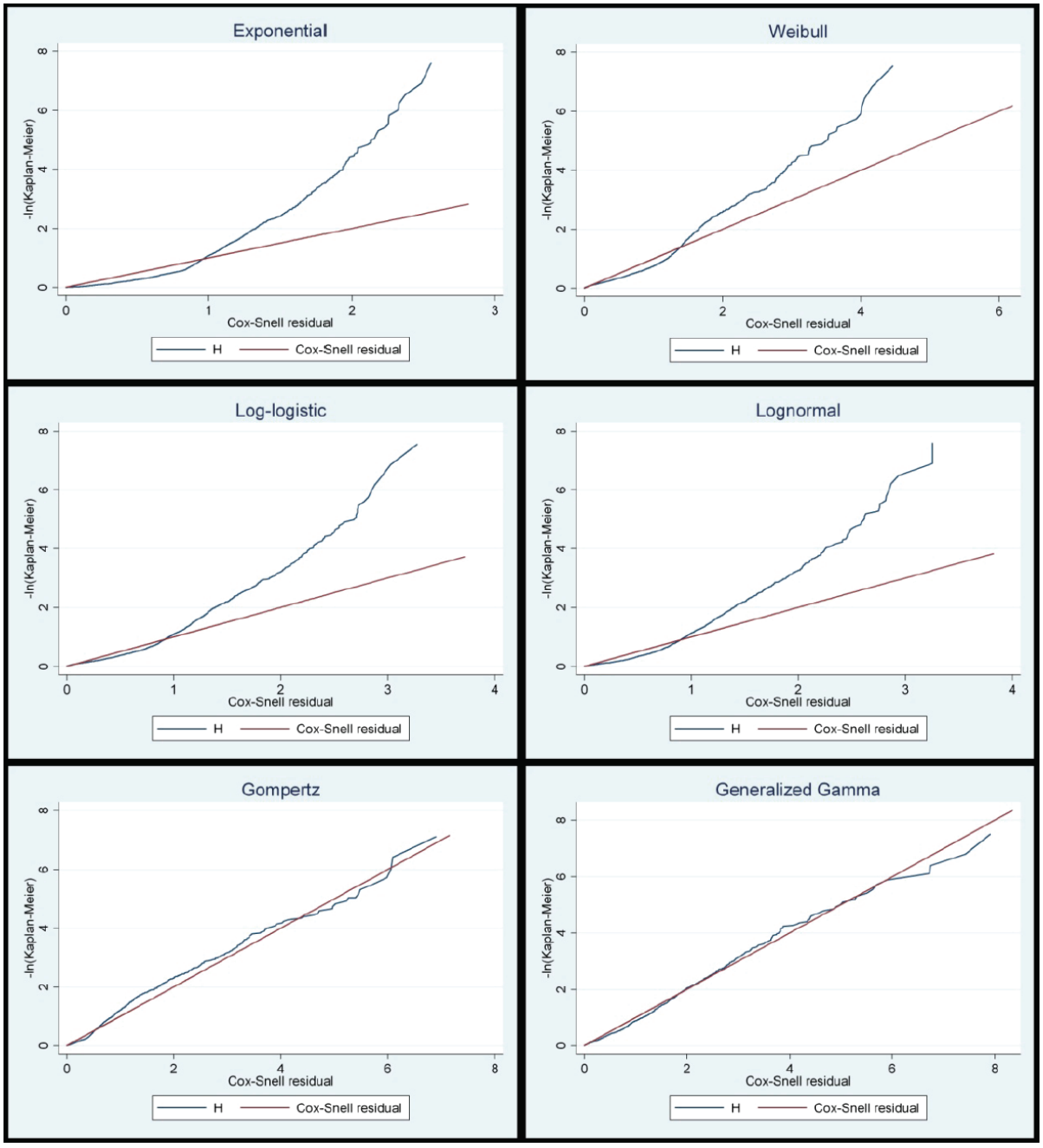

Figure 4. Cox-Snell residuals to evaluate model fit of six regression models.

been diagnosed at the second clinical stage in comparison with those at the first clinical stage, whereas this difference is increased up to about 50% and 60% for patients within the third and fourth clinical stage respectively. Overall, patients within the unknown clinical stage are in the worst situation with a survival time of about 65% shorter than patients within the first clinical stage. However, within one and a half years since diagnosis, patients at the fourth clinical stage have a predicted median survival time of approximately 6 months which is less than the 8 months for patients at the unknown clinical stage. Therefore, as a remarkable finding, one can conclude that the survival time among patients who have been diagnosed at early stage of breast cancer is much higher than the one among patients diagnosed at more advanced

Table 3. Time ratios from the generalized gamma AFT model for the breast cancer data

Covariates	Time ratio	Std. Err.	P-value	[95% Conf. Interval]	
Pinar del Rio +	**1.000**				
Habana	0.988	0.011	0.309	0.966	1.011
C. Habana	0.990	0.009	0.307	0.973	1.008
Matanzas	0.974	0.011	0.015	0.953	0.995
Villa Clara	1.018	0.010	0.063	0.998	1.039
Cienfuegos	1.005	0.013	0.671	0.979	1.033
Sancti Spiritus	0.999	0.012	0.956	0.975	1.024
Ciego de Avila	0.990	0.014	0.484	0.964	1.017
Camaguey	0.992	0.010	0.449	0.972	1.012
Las Tunas	1.001	0.012	0.956	0.977	1.024
Holguin	0.981	0.012	0.060	0.961	1.000
Granma	1.005	0.010	0.631	0.984	1.025
Santiago de Cuba	1.011	0.010	0.254	0.992	1.031
Guantanamo	1.004	0.010	0.704	0.981	1.028
Isla de la Juventud	0.973	0.012	0.254	0.929	1.019
Clinical Stage I +	**1.000**				
Clinical Stage II	0.599	0.019	0.000	0.562	0.639
Clinical Stage III	0.471	0.020	0.000	0.432	0.513
Clinical Stage IV	0.394	0.019	0.000	0.358	0.424
Unknown	0.361	0.020	0.000	0.324	0.403
Age	0.992	0.0005	0.000	0.991	0.993
Age x time	1.004	0.0002	0.000	1.004	1.005
Clinical Stage x time	1.105	0.006	0.000	1.093	1.117
ln_sig	-2.631	0.102	0.000	-2.832	-2.428
kappa	4.425	0.473	0.000	3.501	5.356
sigma	0.720	0.007		0.058	0.088

+: reference category

stage of the disease. Furthermore, patients within one and a half years from diagnosis at the unknown clinical stage have higher survival probability than those within one and a half years from diagnosis at the clinical stage IV. This fact confirms the above result shown in **Figure 3** (left-side) where the survival probabilities within one and a half years from diagnosis for patients at the unknown clinical stage are higher than those at the fourth clinical stage.

The age is also another significant factor, but there is no important difference between patient ages (time ratio is close to 1). Such differences might have been more important if the age variable would have been categorized in different age groups, however we decide to use the age itself to take advantage of valuable individual information.

Discussion

In this article three patient-related factors have been used to assess the survival time of a patient. The province where the patient lives was not a significant factor, whereas the age and the clinical stage of the patient were both statistically significant. The biggest differences in terms of survival were found for the different categories of the clinical stage covariate. The survival time among patients who have been diagnosed at early stage of the disease is about 60% higher than patients diagnosed at more advanced stage of the disease. For that reason, it is extremely important to detect the cancer at early stage to prolong the survival time.

Other important finding is that the diagnostic of the clinical stage must be more precise in order to decrease the number of patients with unknown clinical stages. This fact affects the survival and the quality of life of the patient.

This study has been only focused on comparing prognostic groups in terms of survival. However, further research should be also carried out to compare the impact of different treatments in fighting cancer. For example, the potential effi-

cacy in fighting cancer of a new treatment should be assessed by including the treatment effect as a new covariate into the model used in section 3.2.

It is also worth clarifying that if the model were to be used for the purpose of predicting future survival patterns, it is appropriate to ensure that the effect sizes are robust [10]. That is if the scenarios change, the regression estimates are still near to those obtained from the original data. One approach is to use bootstrap sampling, which involves randomly resampling the data and fitting the model to these modified datasets [16].

Finally, a class of parametric PH models and AFT models have been used in this study; nevertheless other approaches might have also been employed for survival analysis. Within these approaches, we can mention the stratified survival analysis method [12], the Aalen's additive model [17], the classification trees method [18], and the artificial neural networks [19].

Address correspondence to: Nestor Arcia, PhD, National Statistics Office, Cuba. E-mail: nestor@one.cu

References

[1] Health Statistics Bureau. Annual Health Statistics Report. Ministry of Public Health. National Medical Records and Health Statistics Bureau. Cuba. 2008.

[2] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis Part I: Basic concepts and first analyses. Br. J. Cancer. 2003; 89: 232-238.

[3] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Assoc. 1958; 53: 457-481.

[4] Ramlau-Hansen H. Smoothing counting process intensities by means of kernel functions. Ann Statist. 1983; 11: 453-466.

[5] Hosmer DW, Lemeshow S: Applied Survival Analysis: Regression modelling of Time to Event data. New York, Wiley, 1999.

[6] Pories SE. Tumor node metastasis (TNM) staging classification for breast cancer. www.uptodate.com. 2010.

[7] Grambsch PM, Therneau TM. Proportional Hazards tests and diagnostics based on weighted residuals. Biometrika. 1994; 81: 515-526.

[8] Breslow NE. A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. Biometrika. 1970; 57: 579-594.

[9] Wilcoxon F. Individual comparisons by ranking methods. Biometrics. 1945; 1: 80-83.

[10] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis Part III: multivariate data analysis – choosing a model and assessing its adequacy and fit. Br. J. Cancer. 2003; 89: 605-611.

[11] Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate data analysis. Edited by Prentice Hall International, Inc., 1998.

[12] Bradburn MJ, Clark TG, Love S, Altman DG. Survival analysis Part II: multivariate data analysis – an introduction to concepts and methods. Br. J. Cancer. 2003; 89: 431-436.

[13] Fitzgibbons PL, Page DL, Weaver D, Thor AD, Allred DC, Clark GM, Ruby SG, O'Malley F, Simpson JF, Connoly JL, Hayes DF, Edge SB, Lichter A, Schnist SJ. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med July 2000; 124:966-978.

[14] Pinto AE, Andre S, Pereira T, Nobrega S, Soares J. C-erB-2 oncoprotein overesxpression identifies a subgroup of strogen receptor positive (ER +) breast cancer patients with poor prognosis. Ann Oncol 2001 Apr; 12(4): 525-533.

[15] Slamon D, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987, 235: 177-182.

[16] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case-study. J Clin Epidemiol. 2002; 56: 28-37.

[17] Aalen OO. A linear regression model for the analysis of life times. Stat Med. 1989; 8: 907-925.

[18] Lausen B, Sauerbrei W, Schumacher M. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In Computational Statistics, Dirschedl P, Osermann R (eds). Heidelberg/ New York: Physica-Verlag. 1994.

[19] Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet. 1995; 346: 1075-1079.