MECAI 2021 I Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Cancer Risk Assessment Tool: A new general model to estimate cure-rate fraction in patients under tumor therapy

Diego C. Nascimento, Pedro L. Ramos[a], Oilson A. Gonzatto Junior[a], Gabriel G. Ferreira[a], Patrícia P.M. de Castro[a], Renan S. Barbosa[a], Vinicius O. Boen[a], Vinicius H. Valentim[a], Luiz G. Silva[a], Mariana M. Gomes[a], Gleice S. C. Perdoná[b] and Francisco Louzada[a]

[a]Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil
[b]Department of Social Medicine, School of Medicine - FMRP, University of São Paulo, Ribeirão Preto, Brazil

1 Introduction

Cancer is a disease that inflicts a significant portion of the world population. According to Cancer Research UK, there were 17 million new cancer cases in the year 2018 and 9.6 million deaths caused by cancer in the same year [1]. Lung cancer alone is responsible for the most significant number of deaths (1.8 million deaths, 18% of the total), according to the World Health Organization [2]. There has been a tremendous governmental and research institutes effort to minimize the causes of cancer and understand the efficiency of the treatments for the disease.

In this work, we introduced a new cure rate model, named long-term generalized weighted Lindley (LGWL) distribution, to describe and estimate the cure fraction in a study of lung cancer. This distribution considers as a baseline the generalized weighted Lindley distribution (Ramos and Louzada [3, 4]), a critical model that unified different generalizations of the Lindley distributions such as the weighted Lindley and power Lindley. For the proposed model, essential mathematical functions are presented. The inferential procedure is conducted under the maximum likelihood estimators. The likelihood equations are derived and can be used to achieve the estimates of the parameters. A This new distribution, LGWL, can be used to describe the life expectancy of patients with lung cancer.

2 Long-term survival Model

The probability sub-density function (PSDF) of the long-term generalized weighted Lindley (LGWL) distribution is given by

$$f(t|\boldsymbol{\theta}) = \frac{(1-\pi)\alpha\lambda^{\alpha\phi}}{(\lambda+\phi)\Gamma(\phi)} t^{\alpha\phi-1}(\lambda+(\lambda t)^{\alpha})e^{-(\lambda t)^{\alpha}}, \tag{1}$$

where $\boldsymbol{\theta} = (\phi, \lambda, \alpha, \pi)$ denotes the parameter vector. The improper survival function of the LGWL distribution is obtained:

$$S(t|\boldsymbol{\theta}) = \frac{\pi(\lambda+\phi)\Gamma(\phi) + (1-\pi)\left(\Gamma\left[\phi, (\lambda t)^{\alpha}\right](\lambda+\phi) + (\lambda t)^{\alpha\phi} e^{-(\lambda t)^{\alpha}}\right)}{(\lambda+\phi)\Gamma(\phi)}.$$

3 The Cancer Genome Atlas Data

A sample from the Cancer Genome Atlas data (also known as TCGA) is considered in the present work. The TCGA set came from a 10-year project, and it can be accessed at GDC Data Portal, the original data contains more than 11,000 patients with 33 different types of cancer (where the top 3 are *breast invasive carcinoma, kidney clear cell carcinoma*, and *lung adenocarcinoma*). However, the focus herein is to analyze lung adenocarcinoma (a subtype of non-small cell lung cancer), which is related to the lifetime (in years) of 629 patients (5.72% of the set).

Thus, adjusting some survival models, considering the long-term, an interactive online tool was implemented to unravel the gain of the proposed model, by visualizing the applicability of this class of models in the medical field. The URL `https://cemeai.shinyapps.io/cancer_tool` provides this tool to explore better the potential of some survival methods with the cure-rate fraction.

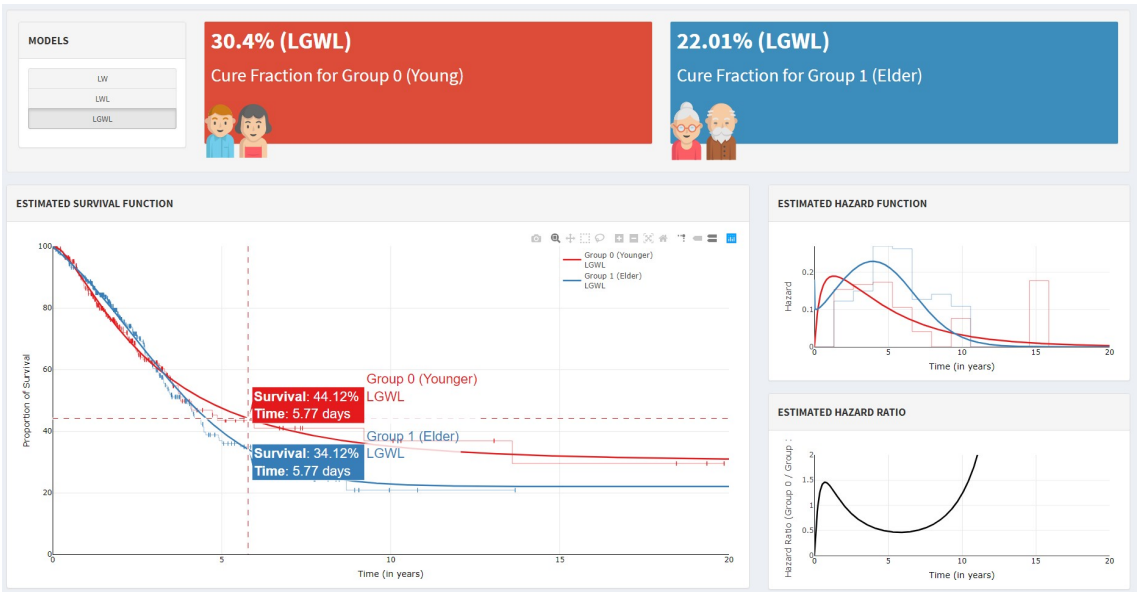

Figura 1: Cancer Risk Assessment Framework, adopting the cure-rate fraction general models. On the Top left, the user can select the models, and the top center and right shows the estimated cure-rate fraction for both groups (younger vs. elder). The left-hand chart represents the adjustment of the selected model based on the empirical function (Kaplan-Meier). The right-hand top chart presents the instantaneous failure rate and Right-hand bottom chart the odds rate among the groups.

The developed framework has mainly three elements: i) the survival model selection, ii) cards displaying information about the estimated cure fraction, and iii) three graphical reports regarding the adjusted survival curve, the instant risk and the risk ratio between the groups under study, as shown in Figure 1.

The patients were divided into two age groups to determine if their age influences on the survival time. This covariable was created to check if there are significant differences in survival time between patients who had cancer up to 69 and those who had the disease at 70 years of age and more. Patients up to 69 years old were classified as "age = 0,"and those aged 70 and older were labeled "age = 1". Table 1 exhibits the ML estimates, the standard errors (SE) and 95% confidence intervals (95% CI) for ages 0 and 1.

Tabela 1: ML estimates, SE and 95% CI for the parameters of the LGWL distribution, considering the TCGA set for ages 0 and 1

	Group 0 (Younger)			Group 1 (Elder)		
Parameter	$\hat{\boldsymbol{\theta}}$	SE	95%CI	$\hat{\boldsymbol{\theta}}$	SE	95%CI
$\hat{\phi}$	5.902	0.1672	(5.101 ; 6.703)	0.647	0.0304	(0.304 ; 0.988)
$\hat{\lambda}$	23.208	3.8835	(19.346 ; 27.070)	0.316	0.0023	(0.221 ; 0.410)
$\hat{\alpha}$	0.427	0.0008	(0.370 ; 0.484)	1.491	0.0878	(0.910 ; 2.072)
$\hat{\pi}$	0.304	0.0058	(0.154 ; 0.454)	0.220	0.0027	(0.117 ; 0.322)

As can be seen in the previous table, the results show that it is possible to estimate the cure fraction ($\hat{\pi}$) related to each group. Since the cure fraction was different for each age group, it is possible to assume that as the patient's age progresses, the cure-rate fraction increases, meaning that younger patients have a higher probability of surviving lung cancer than older patients.

Table 3 presents the Akaike information criterion (AIC) values for different long-term lifetime models using Groups 0 and 1. It can be observed from this table that, although subtle, the LGWL distribution provides a better fit to these data since the adjusted distribution has the lowest value. An essential result of the present work is that we were able to observe such unimodal hazard behavior while the current long-term Weibull distribution cannot fit this type of data. Hence, the new model is an efficient tool to estimate the cure fraction in a clinical trial.

Additionally, from the estimated parameter values, we observe that the hazard rate has unimodal behaviors: for the younger group, the most dangerous period is close to 1.3 years after the pathological diagnosis. For the older group, the higher risk period is close to 4 years after the diagnosis.

Tabela 2: AIC values for the fitted distributions for group 0 (younger patients) and group 1 (elder patients).

	Weibull	LW	LWL	LGWL
Group 0	547.80	533.10	533.20	531.37
Group 1	840.52	838.90	837.40	836.77

4 Discussion

In this article, we proposed a flexible extension of the generalized weighted Lindley. This distribution takes into account individuals who are long-term survivors of lung cancer (event of interest). The Long-term generalized weighted Lindley (LGWL) distribution can estimate the cure-rate fraction of the patients, even considering the posterior period of the study (named long-term). Using the LGWL distribution, medical doctors will reference survival estimation based on the development and spread of cancer as an inverse problem. In the complete paper, we have considered the maximum likelihood estimator to obtain the estimates for the parameters.

Additionally, we have presented a simulation study that shows that our estimation methods return consistent and efficient estimators. While the patient has the disease, a different survival rate must be associated with each period, since the patient's chances of survival should decrease over the years. For instance, in the second year of the development of lung cancer, both groups of patients (younger and older age) present an 80% cure-rate fraction. However, in the tenth year, the group of patients who are younger than 70 presents a 40% rate, whereas the group of elders presents a 20% cure fraction.

Based on the estimated cure fractions, after 14 years of the development/spread of lung cancer, the group of patients under 70 had a cure-rate fraction of 0.322 (which means that around 32% of these patients will survive). In contrast, the group of elderly patients presented a rate of 0.2204 or 22%, whereas those estimations using the long-term Weibull distribution are 31% and 17%, respectively (underestimating the cure-rate fraction elder group).

Referências

[1] Worldwide cancer statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer`. Accessed: 2020-01-13.

[2] Organization, W. H. et al. (2018). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018.

[3] Ramos, P. and F. Louzada (2016). The generalized weighted lindley distribution: Properties, estimation, and applications. *Cogent Mathematics 3*(1), 1256022.

[4] Ramos, P., D. Nascimento, and F. Louzada (2017). The long term fréchet distribution: Estimation, properties and its application. *Biom Biostat Int J 6*(3), 00170.