Survival Curves Projection and Benefit Time Points Estimation using a New Statistical Method

Toni Monleón-Getino*

Section of Statistics, Department of Genetics, Microbiology and Statistics. University of Barcelona, Spain

Abstract: Survival analysis concerns the analysis of time-to-event data and it is essential to study in fields such as oncology, the survival function, *S*(*t*), calculation is usually used, but in the presence of competing risks (presence of competing events), is necessary introduce other statistical concepts and methods, as is the Cumulative incidence function *CI*(*t*). This is defined as the proportion of subjects with an event time less than or equal to. The present study describe a methodology that enables to obtain numerically a shape of *CI*(*t*) curves and estimate the benefit time points (BTP) as the time (t) when a 90, 95 or 99% is reached for the maximum value of *CI*(*t*). Once you get the numerical function of *CI*(*t*), it can be projected for an infinite time, with all the limitations that it entails. To do this task the R function Weibull.cumulative.incidence() is proposed. In a first step these function transforms the survival function (*S*(*t*)) obtained using the Kaplan–Meier method to *CI*(*t*). In a second step the best fit function of *CI*(*t*) is calculated in order to estimate BTP using two procedures, 1) Parametric function: estimates a Weibull growth curve of 4 parameters by means a non-linear regression (nls) procedure or 2) Non parametric method: using Local Polynomial Regression (LPR) or LOESS fitting. Two examples are presented and developed using Weibull.cumulative.incidence() function in order to present the method. The methodology presented will be useful for performing better tracking of the evolution of the diseases (especially in the case of the presence of competitive risks), project time to infinity and it is possible that this methodology can help identify the causes of current trends in diseases like cancer. We think that BTP points can be important in large diseases like cardiac illness or cancer to seek the inflection point of the disease, treatment associate or speculate how is the course of the disease and change the treatments at those points. These points can be important to take medical decisions furthermore.

Keywords: Survival function, projection, Weibull growth curve, non linear regression.

1. INTRODUCTION

A branch of statistics that deals with analysis of time duration until one or more events happen, such as death in biological organisms and failure in mechanical systems [1] is known as survival analysis [2]. Survival analysis concerns the analysis of time-to-event data and it is essential to study in fields such as oncology, the survival function, $S(t)$, calculation is usually used.

In survival data, subjects experience only one type of event over follow-up, such as death from a disease (e.g. cancer). Unfortunately, life is very complex, and sometimes, subjects can potentially experience more than one type of a certain event (e.g. senior patients at an oncology department, could possibly die from heart attack or breast cancer, or even traffic accident). When only one of these different types of event can occur, we refer to these events as “competing events” [3]. In this case, one competing event compete with each other to deliver the event of interest (e.g. death due to illness), and the occurrence of one type of event will prevent the occurrence of the others. As a result, we call the probability of these events as “competing risks” [4, 5], in a sense that the probability of each competing event is somehow regulated by the other competing events, which has an interpretation suitable to describe the survival process determined by multiple types of event [3].

In the presence of competing risks (presence of competing events), is necessary introduce statistical concepts and methods for the analysis of survival data. Cumulative incidence $CI(t)$ is defined as the proportion of subjects with an event time less than or equal to t [4].

In this field the Cumulative incidence function, $CI(t)$, is defined as the probability that a particular event related with time, such as occurrence of a particular disease, has occurred before a given time. It is equivalent to the incidence, calculated using a period of time during which all of the individuals in the population are considered to be at risk for the outcome. It is sometimes also referred to as the incidence proportion, but in function of the evolution of the disease [6] not all the events occur at the same moment or with the same speed, so it would be of interest assess a possible benefit time points (BTP) when the disease could be stable or change.

1.1. Survival Analysis

The survival function $S(t)$ analyses the "time to event outcome variable".

*Address correspondence to this author at the GRBIO (Research Group in Biostatistics and Bioinformatics). BIOST3. Section of Statistics, Department of Genetics, Microbiology and Statistics. University of Barcelona, Spain; Tel: 003494.402.15.60; E-mail: amonleong@ub.edu

A time to event variable, t, reflects the time until a participant has an event of interest (e.g., heart attack, goes into cancer remission, death, curation, etc). Statistical analysis of time to event variables requires special techniques [1] than those described thus far for other types of outcomes because of the unique features of time to event variables. Statistical analysis of these variables is called time to event analysis or survival analysis [6] even though the outcome is not always death. What we mean by "survival" in this context is remaining free of a particular outcome over time.

The survival function, $S(t)$, of an individual is the probability that they survive until at least time t.

$$S(t) = Pr(T > t) \tag{1}$$

where t is a time of interest and T is the time of event.

The survival curve of $S(t)$ is non-increasing (the event may not reoccur for an individual) and is limited within [0,1]. Note that the event might not happen within our period of study and we call this right-censoring (See Figure **1**).

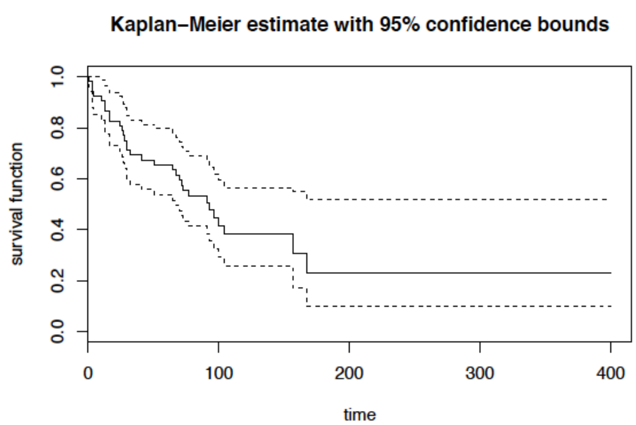

Figure 1: Survival function $S(t)$ for the first example (Tongue cancer using Kaplan–Meier).

The questions of interest in survival analysis are questions like: What is the probability that a participant survives 10 or 20 years? Are there differences in survival between groups (e.g., between those assigned to a new versus a standard drug in a clinical trial)? How do certain personal, behavioural or clinical characteristics affect participants' chances of survival? [1]

1.2. Survival Time t as a Random Variable

The survival analysis of a random variable of time study the T variables "time until an event or event" is known as survival analysis. This analysis contemplates a specific methodology since T measurements occur frequently before the event and patients do not enter the study at the same time [7].

The event considered is not whether or not death occurs, for example, but death related to the disease. If a death unrelated to the disease is considered, an information bias occurs, so the patient died for a cause that is not related to the event of interest should be considered as censored and compute their follow-up time as incomplete or lost.

The event or event studied must also be perfectly defined in order to determine exactly the date of the event. This event is almost always associated with the death of the patient but it does not have to be so, since it can also refer to the discharge date, the date of remission of the disease, the date a clinical event occurs (example: Cardiovascular), [6] the date of relapse, the date of relapse or failure, etc.

From the clinical point of view, survival can be defined as:

- Disease-free survival: Time during which the patient is free of any evidence of illness. It is applicable to patients who undergo radical root treatment and disappear the moment a relapse occurs. If the patient presents advanced disease, the concept of disease-free survival is not applicable, but the duration of the response. Event-free survival:

- Global survival: Life time from the start of study treatment to death or to the last known data, in case of abandonment or loss of follow-up.

One of the objectives of these techniques is to infer the relationship between T and the explanatory variables of the model X that are known and controlled by the researcher in the study. The variable T does not belong to a normal population and can be distributed according to exponential function, Weibull, log-normal or log-logistic.

The differences between the factors studied by the survival analysis can be performed using parametric and non-parametric techniques. A summary is:

- Parametric:
  - Exponential Distribution.
  - Weibull Distribution.
  - Log-Normal Distribution.

- Non-parametric:
  - Kaplan-Meier.
  - Log-rank.
  - Cox Regression (Semi-Parametric method)

We're measuring time-to-event in the real world and so there's practical constraints on the period of study and how to treat individuals that fall outside that period. Censoring is when the event of interest (death, relapse, curing, failure, etc) occurs outside the study period, and truncation is due to the study design.

It is sometimes unknown if the patient has presented the event studied (death, relapse, etc.) or not. These data are known as the censored data [7, 8]. There are several types of censorship such as:

- Censorship type I: is the most common. The study has a limited time. If the time until the event occurs in the patient is less than the time set, the time obtained is taken, otherwise the time until the end of the study.
- Type II Censorship: The study ends when the event has occurred in a given number of individuals.
- Random censorship: The time until the event is observed less than or equal to a constant in censorship I. In this case it is not a constant but a random variable d, which takes into account the causes not considered in the experiment and that cause the censorship. The failure time is observed when $T < d$.

The survival function $S(t)$ (Figure **1**) is defined as the probability of a patient surviving a time t, if T is the survival time variable. It is a decreasing function that satisfies:

$$S(t) \geq 0, S(0) = 1, S(+\infty) = 0 \tag{2}$$

1.3. Kaplan-Meier Method

It is a non-parametric method widely used to estimate survival function [9, 10] (it does not assume any probability function) that uses maximum likelihood, maximising the sample likelihood function. We allow for right-censoring (but not truncation). We start with a random sample of size n, drawn from a population, it will be formed by $k(k \in n)$ times $t_1 < t_2 < ... < t_k$ in which events are observed. At each time, there are no "individuals at risk" and d_i events are observed [1]

This model gives us a maximum-likelihood estimate of the survival function $S(t)$ with Kaplan-Meier product-limit [10] estimator defined as,

$$\widehat{S}(t) = \prod_{t_i < t} (1 - \frac{d_i}{n_i}) \tag{3}$$

where d is the frequency of interest events (e.g. deaths, curing, etc) and n the individuals at risk at time t. The cumulative product (equation 3) gives us a non-increasing curve of survival $S(t)$, at any times t during the study, the estimated probability of survival from the start to that time t. A good survival estimator is the median of survival time (half-life), used frequently.

Kaplan-Meier method calculates survival every time t and an event of interest is presented:

$$S(t) = \left\{ \begin{matrix} 1, 0 \leq t \leq t_1 \\ \prod_{j \leq t} \frac{\mathrm{n_j - d}}{\mathrm{n_j}}, t \geq t_1 \end{matrix} \right\} \tag{4}$$

Where $t_1 < t_2 < ... < t_k$ is defined at times where the event of interest occurs and n_j is the number of survivors before t_j and d_j is the number of individuals presenting the event at time t_j.

This function $S(t)$ is usually represented on a graph as such as the example in Figure **1**.

Survival table [8] is another possibility to compute $S(t)$, in it you can present the values of proportion of survivors, time, number of individuals, cumulative survival rate, probability density, and risk ratio, number of abandonment, number of individuals exposed to risk, number of terminal events, proportion of individuals who have completed, etc.

In a general sense, $S(t)$ is the survival density function, which indicates the time of the greatest number of events T. $H(t)$ is the instantaneous failure rate or risk function and represents the probability that an individual remains alive between the moment t and the $T + \delta t$, previously knowing that it has arrived alive at time t.

There are many methods associated with survival curves, used to compare survival when different levels of a factor associated with an experimental design are available. The log-rank test is used to compare the survival functions according to the assigned treatments or some relevant factor.

In order to be able to construct an explanatory model of the survival function and to explain the relationship between the survival time and the independent variables of the model (sex, age, treatment, stage of disease, tumour marker, etc.), we can use the Cox regression . This methodology [1, 7] allows us to more accurately estimate the survival function $S(t)$ and to determine which variables best explain patients' survival. The Cox regression is represented by a risk function:

$$H(t, x_1, ..., x_n) = h_o(t)e^{\beta_1 X_1 + .. + \beta_n X_n} \quad (5)$$

Where $h_o(t)$ is the baseline risk and $e^{\beta_1 X_1 + .. + \beta_n X_n}$ depends on the independent or explanatory variables (weight, age, treatment, concomitant factors, etc.).

In the Cox model, the coefficients β are determined first and by the Wald test or by the logarithm of maximum likelihood it will be determined whether or not they are significant for the model. Subsequently it is estimated $h_o(t)$.

1.4. A First Example of Survival Analysis with R

The first working example is the study of survival time of a set of patients affected by two variants of the tongue cancer and its survival function is going to be estimated using the Kaplan–Meier method previously exposed, this example will be used later for the purpose of this article. This data set comes from the R package KMsurv. See R Documentation [11] for more information.

The tongue data frame has 80 rows and 3 columns and this data frame contains the following columns:

- Tumor DNA profile (1=Aneuploid Tumor, 2=Diploid Tumor)
- Time to death or on-study time, weeks
- Death indicator (0=alive, 1=dead)

Source was obtained from Klein and Moeschberger (1997) [8].

The survival data estimated was presented in Table **1** and in Figure **1**.

The life table with the survival function estimation and CI95 was represented at Table **2**.

Figure **1**, represents $S(t)$ with survival decreasing from 100% to 20% over 400 weeks is shown with lines above and below that indicate the 95% confidence limits for the survival estimates.

1.5. Cumulative Incidence Curves $CI(t)$ and Competitive Risk (CR)

Competing risks (CR) are present in many medical articles dealing with survival analysis: about half of the Kaplan-Meier analyses in medical journals are susceptible to CR. The issue may become even more relevant in the future, e.g. for elderly patients who are more likely to experience several potential disease endpoints, i.e. the occurrence of competing events increases [12]. The Kaplan-Meier method is applied to estimate the cumulative incidence of an event, using Cumulative Incidence Curves $CI(t)$, computed as 1-(Kaplan-Meier) estimator. This method is appropriate for endpoints such as overall survival, but also for composite endpoints such as progression-free survival [12].

So, complementary to the estimate of $S(t)$ and frequently, the researchers prefer to generate $CI(t)$ (See Figure **2**), as opposed to survival curves $S(t)$ which show the cumulative probabilities of experiencing

Table 1: Survial Data in Tongue Cancer Data-Set

1	3	3	4	10	13	13	16	16	24	26	27	28	30
30	32	41	51	65	67	70	72	73	77	91	93	96	100
104	157	167	61+	74+	79+	80+	81+	87+	87+	88+	89+	93+	97+
101+	104+	108+	109+	120+	131+	150+	231+	240+	400+	1	3	4	5
5	8	12	13	18	23	26	27	30	42	56	62	69	104
104	112	129	181	8+	67+	76+	104+	176+	231+				

The median survival is 93 (67 and NA) weeks is

n	events	median	0.95LCL	0.95UCL
52	31	93	67	NA

Table 2: Life Table with the Survival Function Estimation and CI95 using Kaplan-Meier

Time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95%CI
1	52	1	0.981	0.0190	0.944	1.000
3	51	2	0.942	0.0323	0.881	1.000
4	49	1	0.923	0.0370	0.853	0.998
10	48	1	0.904	0.0409	0.827	0.988
13	47	2	0.865	0.0473	0.777	0.963
16	45	2	0.827	0.0525	0.730	0.936
24	43	1	0.808	0.0547	0.707	0.922
26	42	1	0.788	0.0566	0.685	0.908
27	41	1	0.769	0.0584	0.663	0.893
28	40	1	0.750	0.0600	0.641	0.877
30	39	2	0.712	0.0628	0.598	0.846
32	37	1	0.692	0.0640	0.578	0.830
41	36	1	0.673	0.0651	0.557	0.813
51	35	1	0.654	0.0660	0.537	0.797
65	33	1	0.634	0.0669	0.516	0.780
67	32	1	0.614	0.0677	0.495	0.762
70	31	1	0.594	0.0683	0.475	0.745
72	30	1	0.575	0.0689	0.454	0.727
73	29	1	0.555	0.0693	0.434	0.709
77	27	1	0.534	0.0697	0.414	0.690
91	19	1	0.506	0.0715	0.384	0.667
93	18	1	0.478	0.0728	0.355	0.644
96	16	1	0.448	0.0741	0.324	0.620
100	14	1	0.416	0.0754	0.292	0.594
104	12	1	0.381	0.0767	0.257	0.566
157	5	1	0.305	0.0918	0.169	0.550
167	4	1	0.229	0.0954	0.101	0.518

the event of interest. Cumulative incidence, or cumulative failure probability, is computed as *1-S(t)*, and can be computed easily from the life table using the Kaplan-Meier approach. The cumulative incidence function, also referred to as the cause-specific failure probability [12], can be interpreted as the cumulative probability that a failure of type k occurs on or before time t [13]. The cumulative incidence function helps to determine patterns of failure and to assess the extent to which each component contributes to overall failure. For competing risks data one often wishes to estimate the cumulative incidence probability of failure of a specific cause, k, at time t, that is [9]:

$$P_t(k) = P(T_j \le t, \varepsilon_j = k) = \int_0^1 \lambda_k(s) S(s) du \qquad (6)$$

where ε_j indicates the cause of type of failure, $S(s)$ is the overall survival probability, and $\lambda_k(s)$ is the cause-specific hazards for cause k [14]. For more information about calculation cumulative incidence curve $CI(t)$ see [9]. The cumulative incidence estimator can be expressed in terms of the Kaplan–Meier estimator as,

$$CI(t) = \sum_{t_i < t} \frac{\mathrm{d_i}}{\mathrm{n_i}} K(t_i) \qquad (7)$$

where, t_i is the distinct ordered observed times, n_i is the number of patients who at risk beyond t_i, d_i is the number of events of interest at t_i, $K(t_i)$ is the Kaplan–Meier estimate of the probability of the free of all events at the time t_i.

Furthermore competing risks are events that occur instead of the failure event of interest, and we cannot treat these as censored [15]. When you have competing events, you want to focus on cause-specific hazards rather than standard hazards. When we have competing events, we want to focus on the cumulative incidence function ($CI(t)$) rather than the survival function $S(t)$, Cox regression is fine for cause-specific hazards, but for $CI(t)$ you need to go through a lot of work competing-risks regression by the method of [16] is a possibility.

2. METHOD PROPOSED TO PROJECT CUMULATIVE INCIDENCE CURVE

Byung Mook Weona and Jung Ho Jeb [17] describe a methodology that enables us to obtain separate measurements of scale and shape variances in survival curves and these authors demonstrated that they will be useful for performing better tracking of ageing statistics and it is possible that this methodology can help identify the causes of current trends in human ageing. Also, in this work it is desired to find a method that generalizes this process to diseases where objectives such as survival analysis, such as cancer and heart disease, are used.

2.1. The Cumulative Incidence Curve and its Shaping

We propose the use of a parametric method based on the Weibull growth function or Weibull sigmoid model inspired in a previous research model [18]. We think this method can estimate 90, 95 and 99 maximum percent of ($CI(t)$), X-axis (time) points of great clinical interest known as benefit time points (BTP).

The Weibull distribution of four parameters is an asymptotic growth function and can be expressed as,

$$W(x) = a - be^{-cx^m} \tag{8}$$

where $W(x)$ represents an approximation to ($CI(t)$) being expressed at each time (x). a, b, c and m are parameters to be estimated and e is the base of the natural logarithms.

Parameter a is the upper asymptote of limiting value of the response variable (W):

$$\lim_{x\to\infty} W(x) = a \tag{9}$$

which represents the maximum cumulative survival modelled $1-S(t)$. b is the lower asymptote, c is the parameter governing the rate at which the response variable approaches its potential maximum a or growth rate. Finally, m is a parameter that controls the x-ordinate (time) for the point of inflection (allometric constant). The four parameter Weibull growth model can be easily transformed in a 3, 2 and 1 parameter Weibull model to adapt the relation between dependent ($CI(t)$) at each time "x".

When $m=1$ the Weibull model is a simple exponential growth curve.

Finally, once a good estimate of the function parameters (a, b, c and m) is obtained, it is possible to calculate the desired points on the X axis using the inverse function ($CI(t)^{-1}$). If the Weibull curve correctly represents the function, it has the advantage that it can be projected over time and its immediate application is to know whether or not the curve ($CI(t)$) has reached saturation and when it will reach this maximum limit.

2.2. Representing $CI(t)$ using the Proposed Weibull Model and its Calculation

Here, we present the calculation of survival cumulative incidence $CI(t)$ from S(t) using Kaplan-Meir method. To do this task we have create the function **Cumulative.Incidence.Curves()** , see Appendix I.

Cumulative.Incidence.Curves() function transforms the function $S(t)$ into $CI(t)$: 1-S(t), the results for the case study is presented on Table **3** and drawn in Figure **2**.

In order to detect some relevant clinical points over $CI(t)$ we propose fit it using a Weibull growth model ($W(x)$) of 4 parameters, before commented:

$$W(x) = a - be^{-cx^m} \tag{10}$$

where a is the upper asymptote of limiting value of the response variable (W) , b is the lower asymptote, c is the growth rate and m is the point of inflection. We will estimate these in order to characterise the function $CI(t)$ and trying to determine a benefit time points (BTP).

The specific function **Weibull.cumulative. incidence()** has been developed to be able to estimate the parameters of this model by non-linear regression using the R function, **nls2()**. Fundamentally this function adjust a weibull grown model with 4 parameters (equation 10) using a non-linear regression procedure.

Table 3: Cumulative Indicedence Curve ($CI(t)$) Calculate from the Survival Table

	myFit.time	myFit.surv	myFit.std.err	myFit.lower	myFit.upper
1	1	0.9807692	0.01941839	0.9441432	1.0000000
2	3	0.9423077	0.03431318	0.8810191	1.0000000
3	4	0.9230769	0.04003204	0.8534195	0.9984199
4	10	0.9038462	0.04523081	0.8271685	0.9876318
5	13	0.8653846	0.05469418	0.7774159	0.9633075
6	16	0.8269231	0.06344324	0.7302341	0.9364144
7	24	0.8076923	0.06766650	0.7073724	0.9222396
8	26	0.7884615	0.07182948	0.6849189	0.9076572
9	27	0.7692308	0.07595545	0.6628317	0.8927093
10	28	0.7500000	0.08006408	0.6410776	0.8774289
11	30	0.7115385	0.08829642	0.5984672	0.8459729
12	32	0.6923077	0.09245003	0.5775712	0.8298369
13	41	0.6730769	0.09664709	0.5569274	0.8134500
14	51	0.6538462	0.10090092	0.5365233	0.7968243
15	61	0.6538462	0.10090092	0.5365233	0.7968243
16	65	0.6340326	0.10548917	0.5156073	0.7796580
17	67	0.6142191	0.11016365	0.4949392	0.7622454
18	70	0.5944056	0.11494041	0.4745101	0.7445954
19	72	0.5745921	0.11983624	0.4543127	0.7267155
20	73	0.5547786	0.12486893	0.4343412	0.7086117
21	74	0.5547786	0.12486893	0.4343412	0.7086117
22	77	0.5342312	0.13044827	0.4137057	0.6898696
23	79	0.5342312	0.13044827	0.4137057	0.6898696
24	80	0.5342312	0.13044827	0.4137057	0.6898696
25	81	0.5342312	0.13044827	0.4137057	0.6898696
26	87	0.5342312	0.13044827	0.4137057	0.6898696
27	88	0.5342312	0.13044827	0.4137057	0.6898696
28	89	0.5342312	0.13044827	0.4137057	0.6898696
29	91	0.5061138	0.14121165	0.3837502	0.6674945
30	93	0.4779963	0.15234403	0.3546085	0.6443177
31	96	0.4481216	0.16545504	0.3240114	0.6197713
32	97	0.4481216	0.16545504	0.3240114	0.6197713
33	100	0.4161129	0.18130051	0.2916674	0.5936554
34	101	0.4161129	0.18130051	0.2916674	0.5936554
35	104	0.3814368	0.20111100	0.2571797	0.5657292
36	108	0.3814368	0.20111100	0.2571797	0.5657292
37	109	0.3814368	0.20111100	0.2571797	0.5657292
38	120	0.3814368	0.20111100	0.2571797	0.5657292
39	131	0.3814368	0.20111100	0.2571797	0.5657292
40	150	0.3814368	0.20111100	0.2571797	0.5657292
41	157	0.3051494	0.30074180	0.1692469	0.5501796
42	167	0.2288621	0.41686804	0.1010963	0.5180988
43	231	0.2288621	0.41686804	0.1010963	0.5180988
44	240	0.2288621	0.41686804	0.1010963	0.5180988
45	400	0.2288621	0.41686804	0.1010963	0.5180988

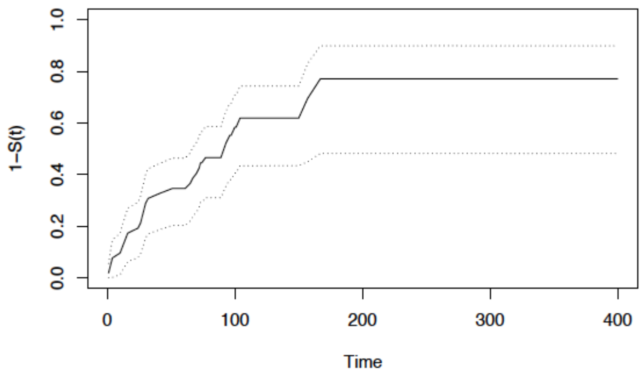

Figure 2: Cumulative survival function $CI(t)$.

Weibull.cumulative.incidence() has been encapsulated within the library BDSbiost3 [Machine learning and advanced statistical methods for omic, categorical analysis and others] [19] that has been developed by the author and is located in Github https://github.com/amonleong/BDSbiost3. The model accuracy (goodness of fit) was tested using Efron's pseudo R^2, Min.max.accuracy (for minimum, maximum accuracy, more substantial indicates a better fit, and a perfect fit is equal to 1) and root mean square error (RMSE) which has the same units as the predicted values. The Weibull sigmoid model obtained best scores and was selected as a good function that fits and extrapolates curve.

The **Weibull.cumulative.incidence()** function also allows to represent the estimated function, its CI95% of prediction and the BTP points of interest of 90, 95 and 99% of the asymptote.

In Table **4** is shown the parameters obtained in the estimation of the Weibull curve using: **Weibull.cumulative.incidence().**

The value of the estimation of the parameters of our case of use is:

- a = 0.81314 (upper asymptote or Asym)
- b = 0.75735 (lower asymptote or Drop)
- c = -5.18548 (growth rate or lrc)
- m = 1.13684 (point of inflection or pwr)

The goodness of test computed for the model is:

- Efron's pseudo r-squared = 0.974
- Min.max.accuracy = 0.923
- RMSE = 0.032

In Figure **3** is presented the curve fitted for $CI(t)$, and the Weibull model estimation is shown (red line with the CI95%).

Also we can calculate in Figure **3** where are the possible benefit time points (BTP) as the time to reach the 90%, 95% and 99% of the upper asymptote, a, using the inverse of the Weibull growth function $W(x)^{-1}$. So, we can obtain the value of time to reach some value of the upper asymptote, we will consider this a possible benefit time points (BTP).

Finally as a result of the function **Weibull.cumulative.incidence()** we have estimated the presumable BTP:

183.9 time units to reach the 90% of maximum value of $CI(t)$ 246.0 time units to reach the 95% of maximum value of $CI(t)$ 361.6 time units to reach the 99% of maximum value of $CI(t)$

2.3. Local Polynomial Regression (LPR)

Another possibility to make a projection of the CI curve is to use a non-parametric method. After trying several it has been chosen Local Polynomial Regression (LPR) fitting.

LPR is a family of flexible and robust non-parametric regression methods that allow fitting smooth curves between two or more independent variables. This type of methods combine multiple regression models in a k-nearest-neighbor neighbour-based meta-

Table 4: Estimation of the 4 Parameters Model Weibull for the Cumulative.Survival.Data.Frame

Parameters for the $CI(t)$:

	Estimate	Std. Error	t value	Pr(>|t|)	
Asym	0.81314	0.02989	27.20	< 2e-16	***
Drop	0.75735	0.04132	18.33	< 2e-16	***
lrc	-5.18548	0.44124	-11.75	1.05e-14	***
pwr	1.13684	0.10191	11.15	5.39e-14	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 model. The method used in the study was LOESS (locally scatter-plot smoothing) as the generalization of LOWESS (locally weighted scatterplot smoothing). The mathematical description of LPR and LOESS can be found at [20].

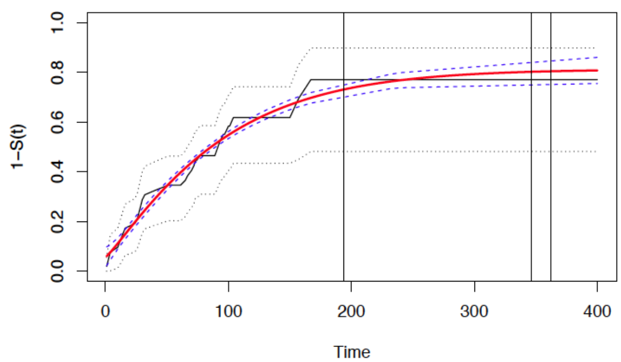

Figure 3: Cumulative survival function *CI*(*t*) (grey line) and the Weibull growth curve *W*(*x*) estimate (red line with CI95% in blue) with the 3 benefit time points obtained.

This method will be used for the following example and it has also been incorporated into the function **Weibull.cumulative.incidence()**.

Is possible to get an R-squared for a LOESS fit using:

$$pseudo - R^2 = 1 - SS.resid \,/\, SS.variable \qquad (11)$$

where SS.resid is the sum of square for the residuals and SS.variable is the sum of square of the variable of interest (Y).

3. AN ALTERNATIVE CASE OF USE FOR WEIBULL.CUMULATIVE.INCIDENCE()

Here, We present other examples of the use of the function **Weibull.cumulative.incidence()**, now using LPR.

The second example (Figures **4**, **5** and **6**) studied belong to the baboon data-set of the package KMSurv. This dataset contains the survival time of few patients during 2000 days of clinical treatment (Cancer) [8]

Figure **4** shows the survival function for the data in this example.

Figure **5** shows the $CI(t)$ function for the data in this example and the estimation of the 4-parameter Weibull function, along with the estimated BTPs.

Figure **6** presents the function $CI(t)$ for the data in this example and the estimation of the non-parametric function LOESS (), together with the estimated BTP.

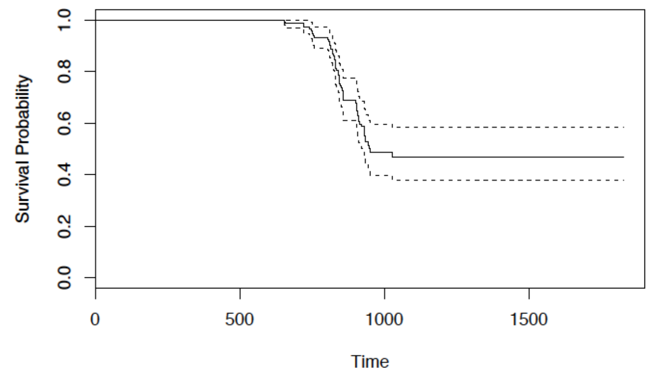

Figure 4: *S*(*l*) for the example of baboon data-set (Klein and Moeschberger (1997) [8].

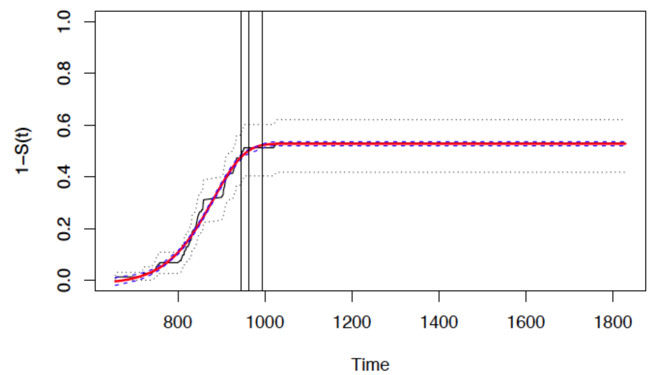

Figure 5: Cumulative survival function *CI*(*t*) (grey line) and the Weibull growth curve *W*(*x*) estimate (red line with CI95% in blue) with the 3 benefit time points obtained in the case of cardiovascular events.

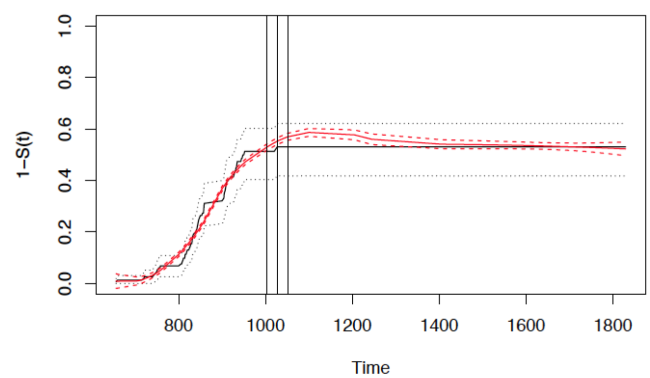

Figure 6: Cumulative survival function *CI*(*t*) (grey line) and the LOESS() growth curve *L*(*x*) estimate (red line with CI95% in blue) with the 3 benefit time points obtained in the case of cardiovascular events.

The goodness of test computed for the Weibull model (parametric) is:

- Efron's pseudo r-squared = 0.882
- Min.max.accuracy = 0.998
- RMSE = 0.0215

The goodness of test computed for the LOESS method (non parametric) is:

- Pseudo r-squared = 0.981704

4. CONCLUSIONS

In this work we present the use of the cumulative survival function, $1-S(t)$, as a form to calculate a benefit time points (BTP) as a time when a 90, 95 or 99% is reach for the maximum $1-S(t)$. To do this task we propose the use of the R function: **Weibull.cumulative.incidence()**

These function transforms the cumulative survival curve $1-S(t)$ ($CI(t)$) obtained using the Kaplan–Meier method and transform $S(t)$ to $1-S(t)$.

In a second step, **Weibull.cumulative.incidence()** estimates a Weibull growth curve of 4 parameters to characterise the best fit function for the *CI*(*t*) and its inverse $CI(t)^{-1}$ to estimate the BCP is a best way.

Alternativelly, **Weibull.cumulative.incidence()** can also estimate BTP using a non parametric LOESS() growth curve.

BTP can be important in diseases like cardiac illness or cancer to seek the inflection point of the disease, treatment associate or speculate how is the course of the disease and change the treatments at those points.

Finally, this model has many possibilities and applications for those situations in which there is great uncertainty and it is necessary to make temporal projections, such as microbiological growth models, or epidemiological models, as in the world of coronavirus pandemic.

APPENDIX: FUNCTION TO COMPUTE CUMULATIVE INCIDENCE FUNCTION $CI(t)$

This function allows calculate Cumulative incidence function $CI(t)$ using estimations of $S(t)$. Is written in R language:

```
#Example in R

#function

Cumulative.Incidence.Curves <- function(survival.data.frame) {#1-S(t)

survival.data.frame$surv_1 <- 1- survival.data.frame$surv

survival.data.frame$lower_1 <- 1- survival.data.frame$lower

survival.data.frame$upper_1 <- 1- survival.data.frame$upper

survival.data.frame$upper_1

survival.data.frame$lower_1

survival.data.frame$surv_1

plot(survival.data.frame$time, survival.data.frame$surv_1,

ylim=c(0,1), ylab="1-S(t)", xlab="Time", lty=1, type="l")

lines(survival.data.frame$time, survival.data.frame$lower_1,

ylim=c(0,1), ylab="1-S(t)", xlab="Time", lty=3, type="l")

lines(survival.data.frame$time, survival.data.frame$upper_1,

ylim=c(0,1), ylab="1-S(t)", xlab="Time", lty=3, type="l")

#return the new data-frame with the 1-S(t)
```

```
return(survival.data.frame)

}

#Example of use

#' #Example with a tongue cancer of the survival package from library KMsurv (Time to death or on-study time, weeks)

#=====> 1. Load libraries: survival and KMsurv <=====#

#Esample with the data-set tongue.

# R packages : survival & KMsurv

library(survival)

library(KMsurv)

#=====> 2. Data-set tongue from KMsurv <=====#

data(tongue)

attach(tongue)

#see the events done in the data-frame (see +)

mySurvObject <- Surv(time, delta)

mySurvObject

detach(tongue)

#=====> 3. Kaplan-Meier estimate and pointwise bounds <=====#

data(tongue)

attach(tongue)

mySurv <- Surv(time[type==1], delta[type==1])

#Estimate the survival mean and CI95

(myFit <- survfit(mySurv ~ 1))

#the life table with the survival funtion estimation and CI95 summary(myFit)

#outputs of the Kaplan-Meier estimate:

myFit$surv # outputs the Kaplan-Meier estimate at each t_i

myFit$time # t_i

myFit$n.risk # Y_i

myFit$n.event # d_i

myFit$std.err # standard error of the K-M estimate at t_i
```

```
myFit$lower # lower pointwise estimates (alternatively, $upper)

# plot Kaplan-Meier estimate

plot(myFit, main="Kaplan-Meier estimate with 95 xlab="time", ylab="survival function")

#use the new function

#1.First:store the life table in a new data-frame attach(myFit)

myfit.data.frame <- data.frame(time,surv,std.err, lower, upper)

myfit.data.frame

#2.Second: Calculate 1-S(t) and plot a graph

res <- Cumulative.Incidence.Curves(myfit.data.frame)

# funtion to transform Survival Cumulative Incidence Curves

#example of use the function Weibull.cumulative.incidence()

#use the new function Weibull.cumulative.incidence()

#1.First:store the life table in a new data-frame attach(myFit)

myfit.data.frame <- data.frame(time,surv,std.err, lower, upper)

myfit.data.frame

#2.Second: Calculate 1-S(t) and plot a graph

res1 <- Cumulative.Incidence.Curves(myfit.data.frame)

res1

#download and install in R the library(BDSbiost) in: https://github.com/amonleong/BDSbiost3

#3.Calculate a Weibull function and analyse 1-S(t) curve and its parameters

a<-Weibull.cumulative.incidence(res1,type = F) #WEIBULL 4 PARAMETERS METHOD

#plot the lines obtained before (90,95 and 99% of maximum survival)

abline(v=c(193.9024, 245.9908, 361.7547),lwd=0.5,lty=2,col="green")

#4.Calculate a Weibull function and analyse 1-S(t) curve and its parameters

a<-Weibull.cumulative.incidence(res1,type = T) #USING A loess METHOD

#plot the lines obtained before (90,95 and 99% of maximum survival)

abline(v=c(158.1066, 176.8298, 193.1702),lwd=0.5,lty=2,col="pink")
```

REFERENCES

[1] Kalbfleisch JD, Prentice RL. Competing Risks and Multistate Models. The Statistical Analysis of Failure Time Data. Second Edition. New York: Hoboken, N.J.: J. Wiley,; 2002.

[2] Dodge Y. The Oxford Dictionary of Statistical Terms. OUP Oxford 2003.

[3] Dutza A, Löck S. Competing risks in survival data analysis. Radiotherapy and Oncology. 2019;130:185–189.

[4] Columbia-University-MSPH. Competing Risk Analysis Tutorial from Population Health Methods. Population Health

Methods; 2020. Available from: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis.

[5] Prentice RL, Kalbfleisch JD, Peterson AV, Jr NF, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978;34(4):541–554.

[6] Armitage P, Berry G, Matthews J. Clinical trials.Statistical methods in medical research. 4th ed. Oxford (UK): Blackwell Science; 2001.

[7] Lee ET, Wang JW. Statistical methods for survival data analysis. Wiley; 2003.

[8] Klein JP, Moeschberger ML. Survival Analysis Techniques for Censored and truncated data. Springer; 1997.

[9] Bintu NS. A comparison of Kaplan–Meier and cumulative incidence estimate in the presence or absence of competing risk in breast cancer. University of Pittsburgh, 2004 . Thesis.; 2004.

[10] Kaplan E, Meier P. Non-parametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53:457–481.

[11] R-Core-Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.; 2013. Available from: http://www.R-project.org/.

[12] Gaynor JJ, Feuer EJ. On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples From Clinical Oncology Data. Journal of the American Statistical Association. 1993;88(422):400–409.

[13] Bryant J, Dignam JJ. Semiparametric models for cumulative incidence functions. Biometrics. 2004;60(1):182–190.

[14] Scheike TH, Zhang MJ. Flexible competing risks regression modeling and goodness-of-fit. Lifetime Data Anal. 2008;14(4):464–483.

[15] Coviello V, Boggess M. Cumulative incidence estimation in the presence of competing risks. The Stata Journal. 2004;4:103–112.

[16] Fine J, Gray R. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association. 1999;94:496–509.

[17] Weona BM, Jeb JH. Trends in scale and shape of survival curves. Sci Rep. 2012;2:504.

[18] Mendez J, Monleon-Getino A, Jofre J, Lucena F. Use of non-linear mixed-effects modelling and regression analysis to predict the number of somatic coliphages by plaque enumeration after 3 hours of incubation. Journal of water and health. 2017;15(5):706–717.

[19] Monleón-Getino T. BDSbiost3: Machine learning and advanced statistical methods for omic, categorical analysis and others; 2020. Available from: https://github.com/amonleong/BDSbiost3.

[20] Fox J. Nonparametric Regression in R: An Appendix to An R Companion to Applied Regression 2nd edition; 2010.

Received on 26-03-2020 Accepted on 23-04-2020 Published on 09-05-2020

https://doi.org/10.6000/1929-6029.2020.09.04