Paper AS06

Methods for survival analysis with competing risks

Laura Marquis, Chrestos Concept, Essen, Germany
Leonie Wagner, Chrestos Concept, Essen, Germany

ABSTRACT

Survival analysis is a major part of clinical trials, especially in cancer studies. The definition of an event varies for different endpoints. An often focused event is death through cancer as a specific cause. In reality a patient is always at risk to die in more than one way. The competing risk survival analysis takes this fact into consideration and can estimate the probability of surviving the cancer if one assumes that it is impossible to die by other means. The situations in which it is useful to consider the competing risks will be discussed. The theoretical options, such as cumulative incidence function (CIF), cause-specific hazards (CSH) and sub-distribution hazards (SDH) will be presented. An example of oncological survival data will show the benefits of considering competing risks. The focus is set on carrying out the analysis with SAS 9.4.

INTRODUCTION

In clinical trials survival analysis is a major part and can differ along with many endpoints. There are many events, which can be considered as endpoints. In traditional survival analysis only one type of event is taken into consideration. This paper covers what to do if there is a second competitive event. For example the event of interest is death trough a specific cause. In reality death can also happen through other causes, which will prevent the primary event (death from cancer) from ever happening. The events are not independent from each other as the first event (death from another cause) prevents the second one from ever happening. If one event happens a subject will stay in this state of death and cannot change anymore.

SURVIVAL ANALYSIS

The survival-times describe how long an individual is observed until an event of interest occurs. Frequently in oncology studies, overall survival is analyzed, but also time to metastasis, time to tumor progression or time to cure are analyzed. Every individual is observed until the event of interest is detected for the first time. For n subjects up to n events can be observed in the study.
In clinical research every individual designated to the study, is assigned a start date. In most cases the event of interest will not be observed in all individuals. For example contact to the individual can be discontinued, if the subject quits the study or the end of the observation time has been reached without any incident. In such cases an observation will be censored.
In general n_i is the number of subjects in the risk set at time point t_i for $i = 1, \dots, m$ for m times of events. At the beginning of the study all individuals are at risk and therefore $n_0 = n$. If one event happens before time point t_i, the individual is excluded from the risk set $R(t_i)$ for time point t_i.
If two different treatments in a study are to be compared in relation to survival times, this can be done using survival analysis methods. Some of these methods will be mentioned below.

SURVIVAL FUNCTION AND KAPLAN MEIER ESTIMATE

The survival-function $S(t)$ describes the conditional probability of the time one random subject has not experienced an event before, $S(t) = P(T \geq t)$. This means, that the target variable for the i-th observation is $T_i > t, i \in \{1, \dots, n\}$. T is a non-negative random variable. The general assumption is $S(0) = 1$. The function $S(t)$ is monotone decreasing. If $f(t)$ is the continuous density of the random variable T, then the survival-function can be defined as:

$$S(t) = \int_t^\infty f(u)du$$

If $f(t)$ is discretely distributed, then S is a step function

$$S(t) = \sum_{t_j > t} p(t_j)$$

whereby $p(t_j) = P(T = t_j)$ $(j = 1,...,m)$ is the probability that an event is going to happen at time point t_j (D. G. Kleinbaum, 1996, p.9). The estimator $\hat{S}(t)$ of the survival function is calculated using the Kaplan-Meier estimator. In order to enable the above calculation, the data needs to be sorted as in Table 1 below.

Table 1: Data structure of event times for Kaplan-Meier estimator

Time of event	Number of risk set	Count of events
t_1	n_1	d_1
t_2	n_2	d_2
:	:	:
t_m	n_m	d_m

In Table 1 the event times are arranged as follows $t_1 < t_2 < \cdots < t_m$. The times of the events $t_1, \ldots, t_m$ indicate m different times at which one or more events happened. The observations are sorted in increasing order. The number of the subjects at risk n_i is often the number of individuals who were event free just before t_i. If an event has happened or a censoring is present then the individual is removed from the risk set and cannot be added again. The censoring times are not directly used in the calculation of the Kaplan-Meier estimator, but the information is used for the risk set. In the case of non existing incident of the event at each time point, the conditional probability that none is going to happen at this time point in time, is derived. In general there will be a step function for time point t:

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{d_i}{n_i}\right)$$

Due to a premature termination of observation and censored observations it may well happen that $\hat{S}(t)$ will not end at zero but higher. This is going to happen if the last observation is censored.
The failure rate describes the probability for an event up to a specific time point. This rate can be estimated by $\hat{F}(t) = 1 - \hat{S}(t)$ (M. Schumacher und G. Schulgen, 2008, p.80f).

HAZARD FUNCTION
The probability for an event at a specific time point t can be calculated using the hazard function. The condition is that no event happened before, $T \geq t$. The hazard function is as follows:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Whereby Δt is preferably small and the length approaches zero. The function $h(t)$ can realize values between 0 and ∞. The hazard function is a conditional density. $H(t)$ is the cumulative distribution function of $h(t)$ and is called the cumulative hazard function. There is a direct link between the cumulative hazard function and the survival function:

$$S(t) = \exp\{-H(t)\}, H(t) = -\log(S(t))$$

(J. Hedderich and L. Sachs, 2016, p.841).
The cumulative hazard function can be estimated with the Nelson-Aalen estimator:

$$\hat{H}(t) = \sum_{t_i < t} \frac{d_i}{n_i}$$

For every time point t the amount of events will be divided by the amount of subjects in the risk set.
The cox regression model is used to describe the impact of covariates on surviving data. This is done with the help of the hazard function and the hazard rates for each individual i at time point t_i are derived (J. Beyersmann et

al., 2012, p. 14).

COMPETING RISK METHODS

Up to this point only methods were described with one event or censored observation per subject. It might also be possible that at least one competing event can occur. Competing risks are defined when subjects are at risk in two or more events, e.g. different causes of death. If one of the events happens, the other event cannot happen anymore.
This means in a multi stage model, that at time point $t = 0$ every subject starts is at stage zero. If the main event happens, the subjects stage changes to one and this cannot be changed anymore. If some other event interferes, the subjects stage changes from zero to two, which is definite as well. The events are not independent of each other. There still can be censored observations as well (J. Beyersmann et al., 2012, p. 42).
If competing risks are present but not taken into consideration in the analysis, this can lead to improper conclusions. With K different events and corresponding probable event times $M_1, \dots, M_K$ the observed time is $T = \min(M_1, \dots, M_K, C)$ where C indicates the time for censoring. The indicator variable $\delta = j$ if $T = M_j$ shows which event was realized (N. Balakrishnan und C. R. Rao, 2004 , p. 292).

CUMULATIVE INCIDENCE FUNCTION

A method for implementing the competing risks is the cumulative incidence function (CIF). This function shows the proportion of the event j that happened at time point t. It is possible that other events can occur, in a ‚realistic' scenario. Formulated as a probability it is $G_j(t) = P(T \leq t, \delta = j)$. This can be calculated as

$$G_j(t) = \int_0^t h_j(u)S(u)du$$

whereby h_j is the hazard function of the event j and S is the survival function. For the calculation of h_j all competing events will be censored, as previously described in the paragraph covering cause-specific hazards. The survival function is computed without taking the concept of competing risks into consideration, as only one event is possible. Due to this, the survival function depends on all events, which leads to a dependency of the CIF of the hazard function on the other events. For the time $t \to \infty$, G_j describes the probability for the event j, $G_j(\infty) = P(\delta = j)$ (N. Balakrishnan und C. R. Rao, 2004 , p. 292).
The estimation of the cumulative incidence function is given by

$$\hat{G}_j(t) = \sum_{t_k \leq t} \frac{d_{jk}}{n_k} \hat{S}(t_{k-1})$$

whereby d_{jk} is the count of events j which happen at time point t_k and n_k is the count of subjects which are at risk just before time t_k. The Kaplan-Meier estimator at time point k-1 is indicated by $\hat{S}(t_{k-1})$ (SAS/STAT 14.1 User's Guide, 2015, p. 5166).

CAUSE-SPECIFIC HAZARD

The cause-specific hazards (CSH) h_j describes the direct risk for an event to happen under the condition that no event happened before.

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta = k | T \geq t)}{\Delta t}$$

Competing risks are considered as independent ‚censoring'. The assumption within this approach is, that it is only possible for one event to happen. The coherence between the CSH and the survival function can be explained using the following. The hazard-function can be written as the sum of the cause specific hazards $h(t) = h_1(t) + \cdots + h_L(t)$. The survival function is defined as $S(t) = \exp(-\int_0^t h(u)du)$ (N. Balakrishnan und C. R. Rao, 2004 , p.314).
The analysis is done by censoring these competing events. With this censoring a cox-regression can be modeled in this approach.

SUB-DISTRIBUTION HAZARD

In cause-specific hazards the effects of the covariates are not in direct correlation with the CIF. Due to this, Fine and Gray (1999) developed a new approach for implementing competing events. The effects of the covariates from the sub-distribution hazards (SDH) approach are in direct coherence with the cumulative incidence function. For this the sub-distribution hazard rate $\gamma_1(t)$ needs to fullfill the following formula for a specific event $\delta = 1$:
$P(T \leq t, \delta = 1) = 1 - \exp(-\int_0^t \gamma(u)du)$.
This results in

$$\gamma(t) \cdot dt = \frac{P(T \epsilon\, dt, \delta = 1)}{1 - P(T \leq t, \delta = 1)}$$

for the SDH rate, which describes the immediate risk for a specific event to happen at timepoint t if no other event has happened before.
The connection between the separate functions with two possible events is:

$$h_1(t) = \left(1 + \frac{P(t \leq t, \delta = 2)}{P(T > t)}\right) \cdot \gamma(t) = (1 + \frac{G_2(t)}{S(t)}) \cdot \gamma(t)$$

In this approach a model with covariates can be calculated, which entails the difficulty to find a proper risk set (J. Beyersmann et al., 2012, p. 91f) .
If competing risk events occur, they will not get censored in the SDH approach. It is also not possible for another event to happen. The risk set consists of all subjects without any event and the subjects with a competing event, which has happened before censoring. The real risk set is actually not known, because the censoring times of subjects with competing events are not given. According to Fine and Gray such a risk set can be determined. At the time point t the individual i is in the risk group with the weight

$$\hat{n}_{0;i}(t) = \mathbb{I}(C_i \geq \min(T_i, t)) \cdot \frac{\hat{S}(t-)}{\hat{S}(\min(T_i, C_i, t)-)} \cdot (\mathbb{I}(t < T_i) + \mathbb{I}(T_i < t, \delta \neq 1)).$$

In the above formula C_i is the time of censoring the individual i and $\hat{S}(t-)$ is the unbiased Kaplan-Meier estimator at the time point just before time t. The subjects having a competing event are taken into account in the calculation by using weights. The weights are dependent on the time t. The calculation of the cox regression can be made with this special risk set (J. Beyersmann et al., 2012, p. 128f). Since SAS 9.4 the CSH and related analysis is integrated. How SAS is used for the analysis will be shown in the following chapters. If there are no competing events present in the data, all regression methods will lead to the same results.

ANALYSIS

In the following chapter the analysis is shown using generated data. The methods described above are used to show the various results and how to interpret them. Additional figures are shown to give a descriptive overview.

DATA

The data in this paper comprises of 100 subjects, 50 for each of both treatments. 20 of the 100 subjects had the event death due to cancer and 20 died through other circumstances. The remaining 60 subjects were censored due to the loss to follow-up or were still alive at the end of the analysis. The time to event or censoring is randomly generated with the utilization of the uniform distribution. The number of days until an event occurs or censoring takes place is stored in the data as seen in Table 2 below.

Table 2: First 11 observations of the analysis data set

	Subject ID	Treatment group	Event	Death Flag	Day of death relative to treatment start	Time
1	10001001	1	1	Y	95	95
2	10001002	1	2	Y	106	106
3	10001003	1	1	Y	65	65
4	10001004	1	2	Y	430	430
5	10001005	1	0		.	273
6	10001006	1	0		.	509
7	10001007	1	1	Y	70	70
8	10001008	1	2	Y	136	136
9	10001009	1	0		.	520
10	10001010	1	0		.	264
11	10001011	1	2	Y	320	320

Table 2 shows the first eleven observations in the analysis data set. All further observations have the same structure. For each subject the event variable indicates wether an event happened or not. If the variable contains the value 1 then death happened due to cancer, for the value 2 death due to other circumstances was observed. Here the value 0 indicates a censored observation. The variable 'Day of death relative to treatment start' is only filled in if a death event occurred during the time of the study. If this is true also the death flag is equal to 'Y'. The

'Time' variable is filled in for every subject and contains the time until the death or the censoring event has happened.

A short summary of the data is shown in Table 3 below. The time is displayed in days.

Table 3: Summary of the analysis data set

Analysis Variable : time Time						
Treatment group	Event	N Obs	N	Mean	Minimum	Maximum
1	0	22	22	454.5000000	22.0000000	989.0000000
	1	14	14	48.5714286	5.0000000	95.0000000
	2	14	14	521.5000000	41.0000000	923.0000000
2	0	38	38	484.8947368	39.0000000	989.0000000
	1	6	6	633.3333333	77.0000000	979.0000000
	2	6	6	726.3333333	507.0000000	907.0000000

Table 3 shows the number of events 1 and 2 as well as the censoring in the two treatment groups 1 and 2. For the variable time in days the mean, minimum and maximum are displayed.

EVALUATION AND INTERPRETATION

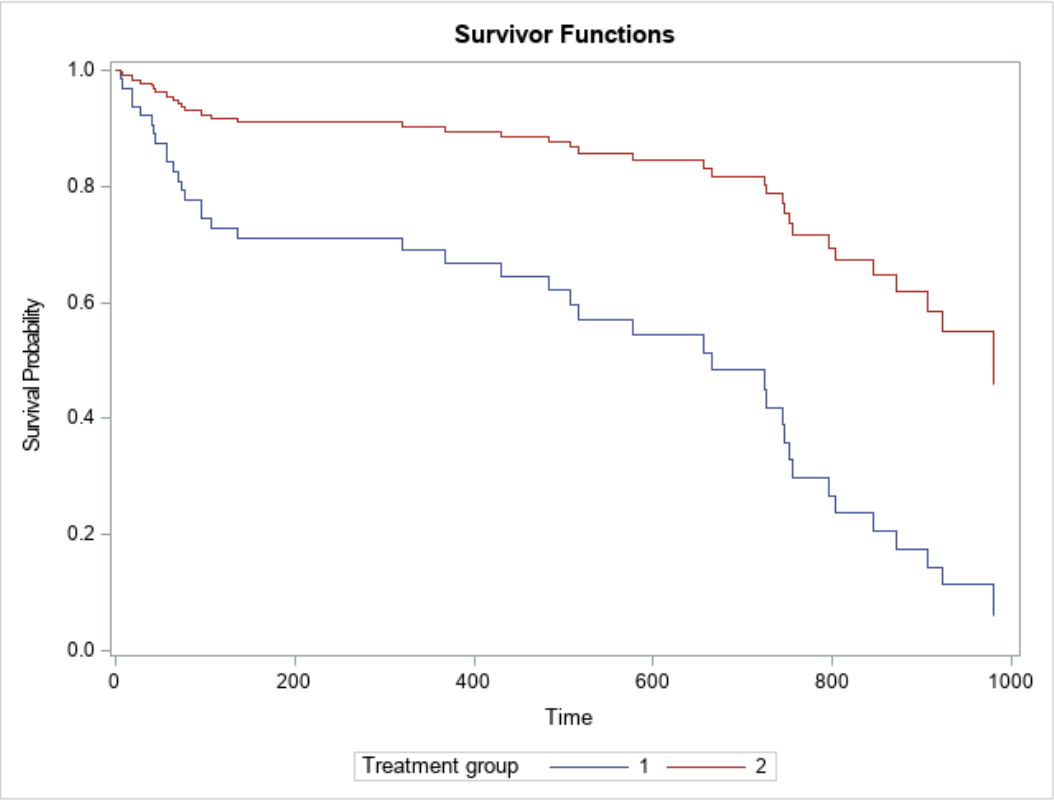
Because the information regarding the death dates is displayed in the above Table 3, the parameter Overall Survival can be analyzed. In general overall survival means time to death. Firstly the Kaplan-Meier curve is shown and the hazard ratio is computed. This means all death events are counted equally as an event. In using SAS 9.4 this can be done with the help of the following code:

```
/*standard cox model*/
proc phreg data=death plots(overlay=stratum)=surival;
      class trt;
      model time*event(0)=trt;
      hazardratio 'Hazard Ratio' trt / diff=pairwise;
      baseline covariates=risk out=_null_ / rowid=trt;
run;
```

Here the 'risk' data only consists of two observations and one variable. This variable is used as the treatment variable and contains either value 1 or 2 for any observation.

In order to keep it simple, no further covariates are considered here. The results of this computation are displayed below in Figure 1.

Figure 1: Kaplan-Meier curves for overall survival in treatment groups 1 and 2



In Figure 1, the curve for each treatment is displayed. The curves do not intersec and the red curve in treatment two remains above the other curve. This means that death events do occur more frequently later in treatment two than in treatment one. In the account of censored observations, the curves do not end at the survival probability of zero. The hazard ratio of comparing treatment one to treatment two is equal to 3.618. This means the risk is increased by about 3.6 times if the patient is receiving treatment one compared to treatment two in any death event.

If the event of death is due to other circumstances then the cancer should be considered as a competing event and a different method needs to be applied. As for the cause-specific hazards, the competing events will be censored at the time of the event. In SAS the competing events will be indicated as censored in the PROC PHREG statement.

```
/*cause specific hazards*/
proc phreg data=death;
      class trt;
      model time*event(0,2)=trt;
      hazardratio 'Cause-specific Hazards' trt / diff=pairwise;
      baseline covariates=risk out=_null_ / rowid=trt;
run;
```

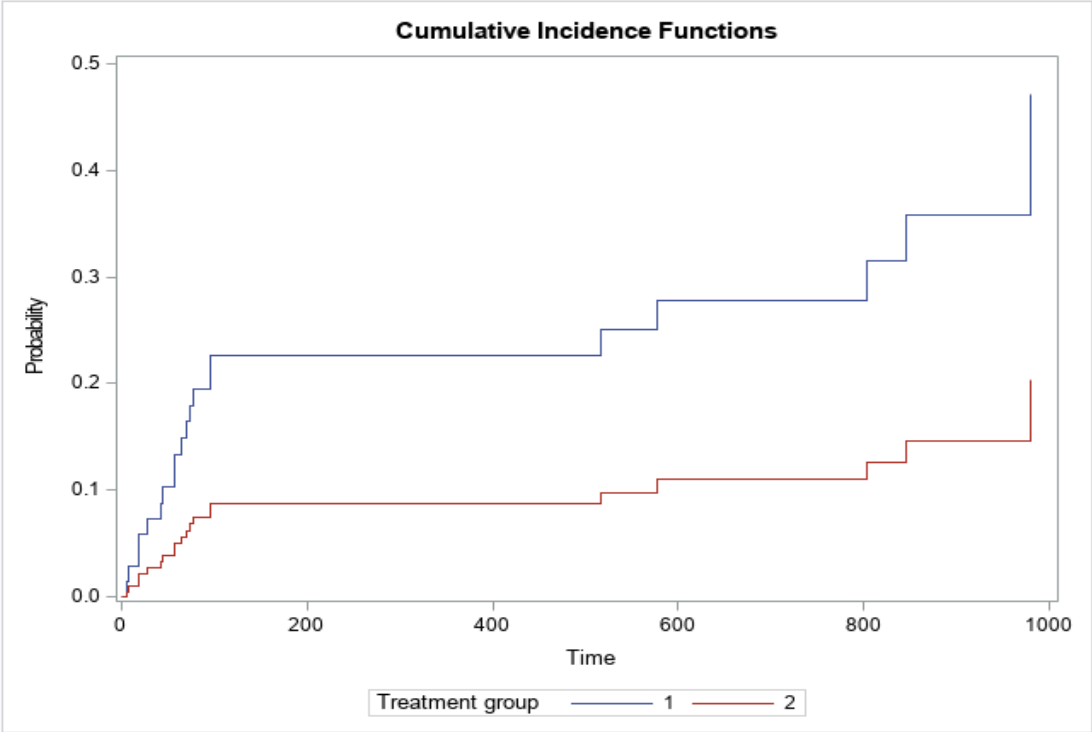
Here the assumption is that it is only possible to die of cancer but not due to any other condition. The cause-specific hazard ratio, treatment one compared to treatment two, is 3.2. This indicates the risk for the event death due to cancer is 3.2 times higher for patients under treatment one compared to patients under treatment two. The method is not commonly used because the effects of the covariates in a cox-regression are not in direct correspondence with the CIF. Furthermore this approach is not really realistic in the assumption that death can only occur due to cancer disease.

In the second method considering competing risk, sub-distribution hazards are derived. The risk set is utilized by applying the method defined by Fine and Gray. The competing events will be considered as such and in the case of an event a subject will stay in the risk set with a specific weight for each time point assigned. This can be directly applied in the PROC PHREG statement using SAS version 9.4 and higher. The program code needs to be modified explicitly for the model description.

```
/*sub-distribution hazards*/
proc phreg data=death plots(overlay=stratum)=cif;
      class trt;
      model time*event(0)=trt /eventcode=1;
      hazardratio 'Subdistribution Hazards' trt / diff=pairwise;
      baseline covariates=risk out=_null_ / rowid=trt;
run;
```

Using the above code the Cumulative Incidence function is also provided. In Figure 2 below one CIF curve for each treatment is displayed, like in the Kaplan-Meier plot.

Figure 2: Cumulative incidence functions for treatment groups 1 and 2



In contrast to the Kaplan-Meier curves a failure function is shown but not a survival function. The curve for treatment one is higher than the curve for treatment two at all times. This means that the probability for a death event due to cancer is higher for subjects receiving treatment one compared to subjects in treatment two. The sub-distribution hazard ratio for treatment one compared to treatment two or in other words, the risk for the event death due to cancer, is 2.8. The hazards for the covariates are in direct coherence with the CIF above. The method incorporates the event of death due to circumstances other than cancer and the information is provided by the risk set.

CONCLUSION

Comparing the three methods, all lead to very similar results. In all methods the subjects in treatment two have a lower probability for a death event. The hazard ratios for the methods differ as well. This does not mean that one method is wrong but the interpretation of the results is different in all methods. If an analysis is done where competing events are present without being considered, the results can be interpreted incorrectly. In this example the difference in risks comparing the two treatment groups is smaller if the competing death events are considered. The main difference between the methods is the definition and derivation of the risk sets.
The program codes in SAS do not differ a lot. It is a simple computation if the data is in the correct format and the procedure PROC PHREG in SAS 9.4 can be used. Prior to SAS 9.4 the calculation of the risk for the sub-distribution method was not provided by the PROC PHREG.

REFERENCES

Balakrishnan, N., Rao C. R. (2004*): Handbook of Statistics, Volume 23.* 1. Edition, Elsevier Amsterdam, The Netherlands.

Beyersmann, J., Schumacher M., Allignol A. (2012): *Competing Risks and Multistate Models with R*. 1. Edition, Springer New York.

Fine, J. P. and Gray, R. J., (1999). *A proportional hazards model for the subdistribution of a competing risk.* Journal of the American Statistical Association, 94: 496–509.

Hedderich, J., Sachs, L. (2016): *Angewandte Statistik. Methodensammlung mit R.* 15. Edition, Springer-Verlag, Berlin, Heidelberg.

Kleinbaum, D. G. (1996): *Survival Analysis: a self-learning text*. 1. Edition, Springer Science+Buisness Media, New York.

SAS®support(https://support.sas.com/en/support-home.htm)

SAS Institute Inc. (2015): *SAS/STAT® 14.1 User's Guide*. SAS Institute Inc, Cary, NC.

Schumacher, M., Schulgen, G. (2013): *Methodik klinischer Studien. Methodische Grundlagen der Planung, Durchführung und Auswertung.* 3. Edition, Springer-Verlag, Berlin, Heidelberg.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name	Laura Marquis
Company	CHRESTOS Concept GmbH & Co.KG
Address	Giradetstraße 1-5
City / Postcode	45131 Essen
Work Phone:	+49 151 74387620
Email:	laura.marquis@chrestos.de
Web:	www.chrestos.de