Paper PP03

CDISC's ADTTE domain: Time-to-event analysis from a programmer's point of view

Vitali Gering, Chrestos Concept, Essen, Germany
Frieder Wolff, Chrestos Concept, Essen, Germany

ABSTRACT

In later phase clinical trials, the comparison of time-to-event variables, so-called endpoints, is of paramount interest. According to CDISC's Analysis Data Model (ADaM version 2.1) everything needed for this endpoint analysis is stored in the ADTTE domain. Note that the event of interest does not necessarily have to occur in each patient. As a consequence, the observed times might have to be censored according to prespecified rules. As a statistical analyst or data manager it is an advantage to have knowledge of the most common approaches to analyze such data. In this paper, we focus on the Kaplan-Meier estimator and the resulting graphical curves. We use progression free survival (PFS) and overall survival (OS) as examples, as these endpoints are frequently used in oncological trials. Our goal is to show how the ADTTE domain is used as a general basis, for enabling a valid and well-structured evaluation.

INTRODUCTION

With approximately 14 million new cases in 2012, cancer is one of the leading causes of morbidity and mortality worldwide (Ferlay et al., 2013) and therefore of major interest for the pharmaceutical industry. According to the world preview (EvaluatePharma®, 2017), oncology is the largest major therapy segment in the pharmaceutical industry with drug & OTC sales of 93.7 $bn and no change in course is expected.

To study the occurrence and timing of events is of great importance in clinical trials. Examining the time span from a defined starting point to the time of event of interest is the basis of survival analysis. The first part of this paper covers useful background information on survival analysis including the concept of censoring as well as the Kaplan-Meier approach. The second part introduces the most important variables in the CDISC ADTTE domain. A sample study is used to apply the introduced methods to two of the most important endpoints in oncology.

The purpose of this paper is to assemble and explain what to keep in mind as a statistical analyst or data manager working on the CDISC ADTTE domain. Being familiar with the underlying data structure and the distinctions between the different time-to-event endpoints facilitates the creation of a high quality ADTTE domain. To understand how the statistical analysis of survival data is enhanced by a structured database is particularly important.

SURVIVAL ANALYSIS

In clinical science, particularly in oncologic trials, the evaluation of time-to-event data is very important. The methods applied in this context are termed as survival analysis. Time-to-event variables are defined as the time span from a starting point to the time when an event of interest occurs.

According to the FDA guidance document *Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics* (2007) overall survival (OS) is considered the most reliable cancer endpoint. Overall survival is defined as the time from randomization until death from any cause. In comparison to other endpoints the definition is quite simple, only the date of death needs to be assessed properly.

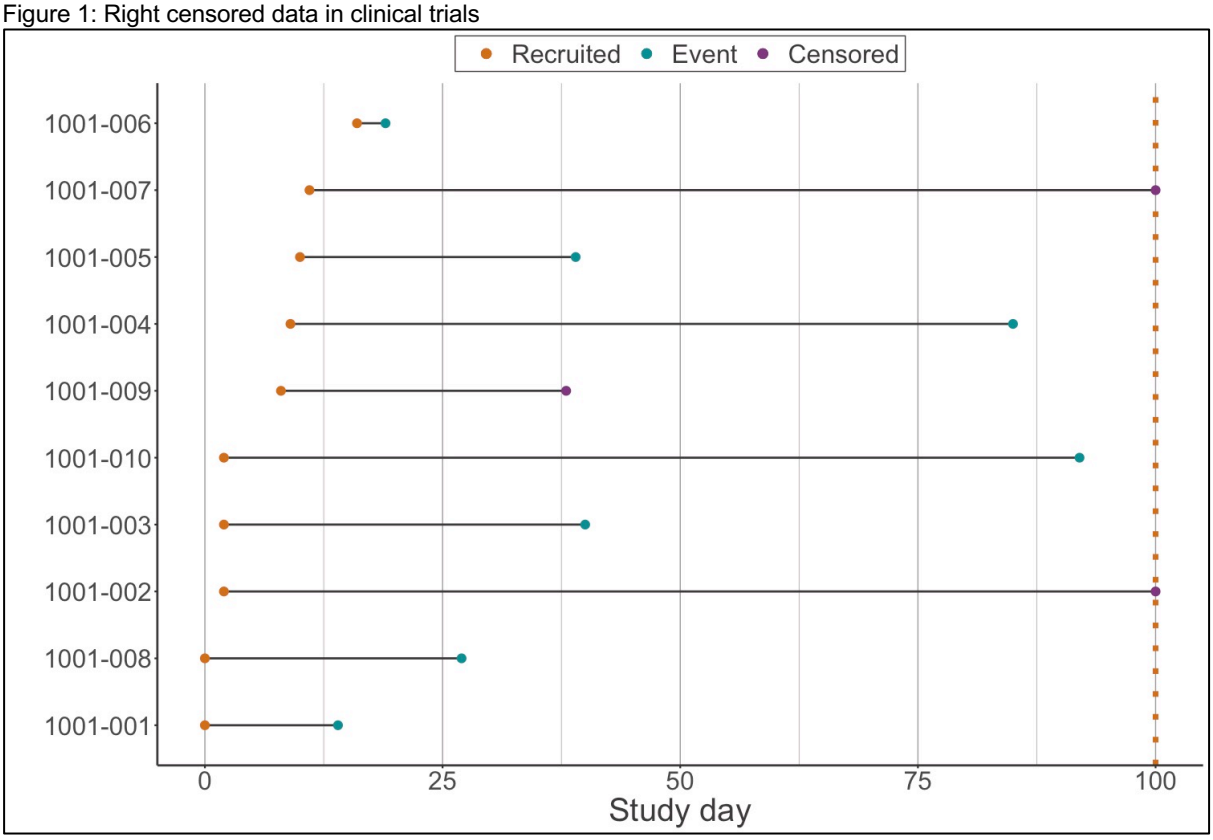
In addition, the guidance document defines endpoints based on tumor assessment. In this paper, progression free survival (PFS) is described exemplary for this category of endpoints. PFS is defined as the time from randomization until objective tumor progression or death. In contrast to OS there are two potential events of interest for PFS. It is very important that the definition of tumor progression is defined accurately in the study protocol. For example, tumor assessments usually have to be verified by central review. One common method to assess solid tumor burden is RECIST (Eisenhauer et al., 2009). There are appropriate criteria for other indications as well, e.g. the Revised Response Criteria for Malignant Lymphoma (Cheson et al., 2007).

CENSORED DATA

Unfortunately, there are limitations in clinical trials that may lead to incomplete data. For example, subjects might withdraw their informed consent, which means that it is not admissible to collect further data. Furthermore, there often are pre-defined time points for analysis of the data. This is because neither the pharma companies nor the authorities are willing to wait for every subject to finish the study in order to obtain valid results. Consequently, the event of interest may or may not have occurred during the observation period for each subject within the trial. If no event occurred for a subject the respective time-to-event is said to be censored. Censored data is the distinctive feature of survival data and needs to be addressed appropriately (CDISC ADaM Team, 2012).

Figure 1 illustrates right censored data in a typical clinical trial. Our example contains 10 subjects that were recruited successively. At the beginning of the trial only two subjects were included (see also table 3). For subjects 001, 003, 004, 005, 006, 008 and 010 an event was observed before the analysis time point (study day = 100). The time-to-event for subject 009 was censored before the analysis time point. For example, this subject prematurely discontinued the study before an event of interest occurred. Subjects 002 and 007 were censored at the analysis time point as no event occurred until then.

Figure 1: Right censored data in clinical trials



KAPLAN-MEIER APPROACH

One widely used method to deal with censored data is the product-limit (PL) estimator published by Kaplan and Meier (Kaplan and Meier, 1958). The PL estimator is a nonparametric method to estimate a survival function and is usually mentioned and recognized as the Kaplan-Meier (KM) estimator.

In general, the survival function describes the population probability of surviving beyond timepoint t:

$$S(t) = P\,(T\ > t)$$

It should be noted that the term survival function is used regardless of the nature of the event of interest. For example, a survival function could describe the probability of not having a disease progression until timepoint t. Even time-to-event variables with positive outcomes as response to a therapy could be described by survival functions.

Table 1 shows the cumulative structure of the KM estimator for the event timepoints $t_1, \cdots, t_m$. The numbers of subjects with an event at each of these timepoints are denoted by $d_1, \cdots, d_m$ and the numbers of subjects at risk at these timepoints are denoted by $n_1, \cdots, n_m$. A subject is considered at risk at a timepoint t if no event occurred until then and the time-to-event is not censored. In the last column the KM estimator for each timepoint is shown.

Table 1: cumulative structure of the KM estimator

Event timepoints	Number of subjects at risk	Number subjects with an event	KM estimate
t_1	n_1	d_1	$\hat{S}(t_1) = \frac{n_1 - d_1}{n_1}$
t_2	n_2	d_2	$\hat{S}(t_2) = \frac{n_1 - d_1}{n_1} \times \frac{n_2 - d_2}{n_2}$
t_3	n_3	d_3	$\hat{S}(t_3) = \frac{n_1 - d_1}{n_1} \times \frac{n_2 - d_2}{n_2} \times \frac{n_3 - d_3}{n_3}$
⋮	⋮	⋮	⋮
t_m	n_m	d_m	$\hat{S}(t_3) = \frac{n_1 - d_1}{n_1} \times \frac{n_2 - d_2}{n_2} \times \cdots \times \frac{n_m - d_m}{n_m}$

In each line of the table the product of the KM estimate is extended by one factor. Each factor the fraction of the number of subjects at risk minus the number of subjects with an event and the number of subjects at risk at the timepoint. The universal KM estimate for a timepoint t could be written as:

$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i})$$

Please note that the number of censored subjects is taken into account indirectly through the number of subjects at risk. The feature of using information from censored subjects as well is what makes the KM estimator a very useful method in survival analysis. Since the KM estimate only changes at timepoints where subjects have an event, the resulting function is a step function (see figure 2 and figure 3 as well).

CDISC's ADAM BASIC DATA STRUCTURE FOR TIME TO EVENT ANALYSIS

To support submission to a regulatory agency such as the United States Food and Drug Administration (FDA) and to facilitate the comparison between studies, the Clinical Data Interchange Consortium (CDISC) has developed multiple standard data models. The Analysis Data Model (ADaM) was designed to meet the agencies' as well as the industry's needs. Its fundamental principle is the requirement for transparency and a valid scientific review through regulatory agencies. ADaM is applicable in a wide range of drug development activities and provides a standard for transferring datasets between sponsors and contract research organizations (CROs). The ADaM Implementation Guide (ADaMIG) describes two standard data structures: The Subject-Level Analysis Dataset (ADSL) and the Basic Data Structure (BDS). ADSL contains variables such as population flags, treatment variables and demographics. It always includes only one record per subject, whereas a BDS dataset can consist of one or more records per subject. (ADaMIG_1.1.pdf, Kapitel 1.2)

According to the CDISC Analysis Data Model (ADaM, version 2.1, 2009) the ADaM TTE dataset should be used for time-to-event variables. In addition to the ADaM Implementation Guide (ADaM IG, version 1.1, 2016) CDISC published "The ADaM Basic Data Structure for Time-to-Event Analyses" document to describe variables that support a specific TTE analysis. The ADaM TTE is simply referred to as ADTTE.

IMPORTANT VARIABLES IN ADTTE

The following table shows important variables that one would commonly find in an ADTTE dataset. Besides the variable`s name and label, a third column contains a short description of each variable. All of the variables included in this table are used throughout this paper.

Table 2: Important variables in the CDISC ADTTE domain

Variable Name	Variable Label	CDISC Notes
USUBJID	Unique Subject Identifier	Must be identical to the ADSL variable.
PARAM	Parameter	Description of the analysis parameter. Examples include: "Time to Death (days)" and "Time to Progression Free Survival (days)". PARAM should be sufficient to describe clear contents of AVAL, including the unit of the measurement, if applicable.
PARAMCD	Paramter Code	Short name of the analysis parameter in PARAM. There must be a one-to-one mapping with PARAM. Examples include DEATH and PFS.
AVAL	Analysis Value	AVAL is the elapsed time to the event of interest from the origin. For example, if AVAL is measured in days, AVAL would be ADT – STARTDT + 1.
STARTDT	Time to Event Origin Date for Subject	Original date of risk for the time-to-event analysis. This is generally the time at which a subject is first at risk for the event of interest evaluation. This may be the randomization date or the date of first study therapy exposure.
ADT	Analysis Date	Analysis date of event or censoring associated with AVAL.
CNSR	Censor	CNSR is a required variable for a time-to-event analysis dataset, though it is a conditionally required variable with respect to the ADaM BDS. CNSR = 0 for event and CNSR > 0 for censored records.
EVNTDESC	Event or Censoring Description	Describe the event of interest or an event that warrants censoring.

EXAMPLE 1: OVERALL SURVIVAL (OS)

The first example in this paper covers the endpoint OS, which incorporates a binary censoring variable. This censoring variable has a value of 0 for subjects who died and a value of 1 for those who were censored. The reason for their censoring is either because they were lost to follow-up or completed the study assessment without experiencing the event of interest, which is death.

Consider a fictional study including ten subjects. Table 3 shows their time-to-event information. Each subject is identified by a unique subject identifier (variable USUBJID). As this example deals with time to death, PARAM (PARAMCD) would be filled with DEATH (Time to Death) accordingly. The study started on January 1st 2018 and variable STARTDT represents the origin of the time-to-event calculation. As in a real-life scenario, the subjects in this study were recruited successively and not all at once. However, the analysis timepoint of interest was April 11th 2018, i.e. the same for all of the subjects. In case of no occurrence of an event up to the analysis timepoint, ADT was set to 04/11/2018 (see subjects 002 or 007). If a subject died between their start date and the analysis time point, then ADT was populated with the subject's death date. One reason for censoring could be a preliminary withdrawal of informed consent prior to April 11th 2018. In such a case the date of the withdrawal was used to populate ADT (see subject 009).

Table 3: Example of the CDISC ADTTE domain (overall survival)

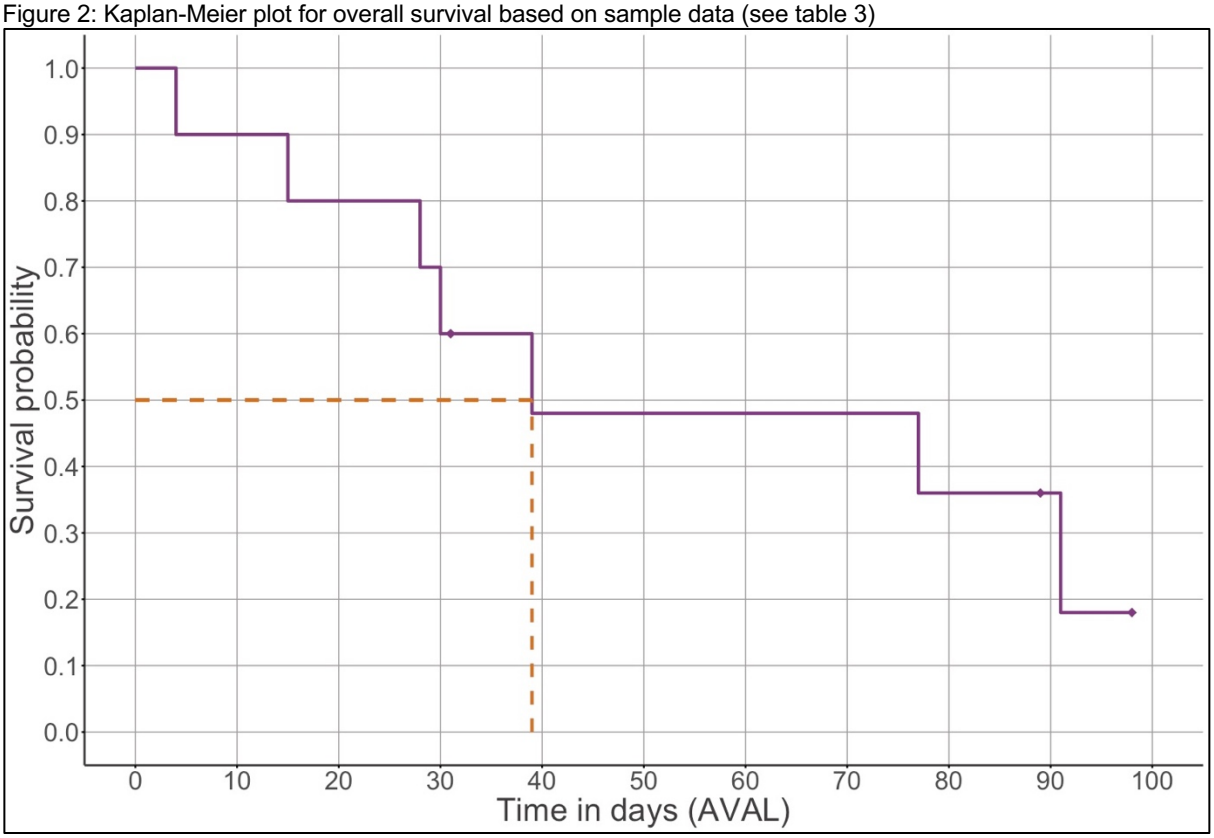
USUBJID	AVAL	STARTDT	ADT	CNSR	EVNTDESC
1001-001	15	01/01/18	01/15/18	0	DEATH
1001-002	99	01/03/18	04/11/18	1	NO EVENT UP TO ANALYSIS TIMEPOINT
1001-003	39	01/03/18	02/10/18	0	DEATH
1001-004	77	01/10/18	03/27/18	0	DEATH
1001-005	30	01/11/18	02/09/18	0	DEATH
1001-006	4	01/17/18	01/20/18	0	DEATH
1001-007	90	01/12/18	04/11/18	1	NO EVENT UP TO ANALYSIS TIMEPOINT
1001-008	28	01/01/18	01/28/18	0	DEATH
1001-009	31	01/09/18	02/08/18	1	WITHDRAWAL OF INFORMED CONSENT
1001-010	91	01/03/18	04/03/18	0	DEATH

Table 4 demonstrates the cumulative structure of the KM estimator as described in table 1 for this particular example. When calculating the KM estimates it is important to keep an eye on the censored subjects. On day 30, there are seven subjects at risk and although only one event occurred on that day, the number of subjects at risk on the event timepoint is only five. The reason for this is the censoring of subject 009 on day 31.

Table 4: Kaplan-Meier estimates for overall survival based on sample data (see table 3)

Event timepoint	Number of subjects at risk	Number of subjects with an event	KM estimate
4	10	1	$\hat{S}(4) = \frac{10-1}{10} = 0.9$
15	9	1	$\hat{S}(15) = 0.9 \times \frac{9-1}{9} = 0.8$
28	8	1	$\hat{S}(28) = 0.8 \times \frac{8-1}{8} = 0.7$
30	7	1	$\hat{S}(30) = 0.7 \times \frac{7-1}{7} = 0.6$
39	5	1	$\hat{S}(39) = 0.6 \times \frac{5-1}{5} = 0.48$
77	4	1	$\hat{S}(77) = 0.48 \times \frac{4-1}{4} = 0.36$
91	2	1	$\hat{S}(91) = 0.36 \times \frac{2-1}{2} = 0.18$

Figure 2: Kaplan-Meier plot for overall survival based on sample data (see table 3)



The KM plot can be used as a visual representation of the above table as shown in figure 2. On the horizontal axis the time in days (AVAL) is plotted. It ranges from 0 to the study day of the analysis timepoint, i.e. 100. The survival probability is plotted on the vertical axis. Starting at 1 (e.g. 100% survival), the so-called survival curve decreases stepwise over time. Each step represents the occurrence of an event. The median survival probability of 50% is crossed on day 39 with the death of subject 003, which is illustrates with the orange dotted line.

EXAMPLE 2: PROGRESSION FREE SURVIVAL (PFS)

The second example covers PFS, which is a composite endpoint. Table 5 below shows how PFS data with multiple event and censoring outcomes can be handled in ADTTE. In contrast to OS not only death but also a documented progression of disease according to a response evaluation criterion such as RECIST is considered as an event. The censoring variable is populated with different values > 0 depending on the reason of censoring. Because the time of last tumor assessment is used, the censoring date tends to be earlier than ADT for OS.

Table 5: Example of the CDISC ADTTE domain (progression free survival)

USUBJID	AVAL	STARTDT	ADT	CNSR	EVNTDESC
1001-001	15	01/01/18	01/15/18	0	DEATH
1001-002	95	01/03/18	04/08/18	1	NO EVENT UP TO ANALYSIS TIMEPOINT. CENSORED AT TIME OF LAST TUMOR ASSESSMENT.
1001-003	30	01/03/18	02/01/18	0	DOCUMENTED PROGRESSION
1001-004	76	01/10/18	03/26/18	0	DOCUMENTED PROGRESSION
1001-005	30	01/11/18	02/09/18	0	DEATH
1001-006	3	01/17/18	01/19/18	0	DOCUMENTED PROGRESSION
1001-007	72	01/12/18	03/24/18	0	DOCUMENTED PROGRESSION
1001-008	28	01/01/18	01/28/18	0	DEATH
1001-009	27	01/09/18	02/04/18	2	WITHDRAWAL OF INFORMED CONSENT. CENSORED AT TIME OF LAST TUMOR ASSESSMENT.
1001-010	91	01/03/18	04/03/18	0	DEATH

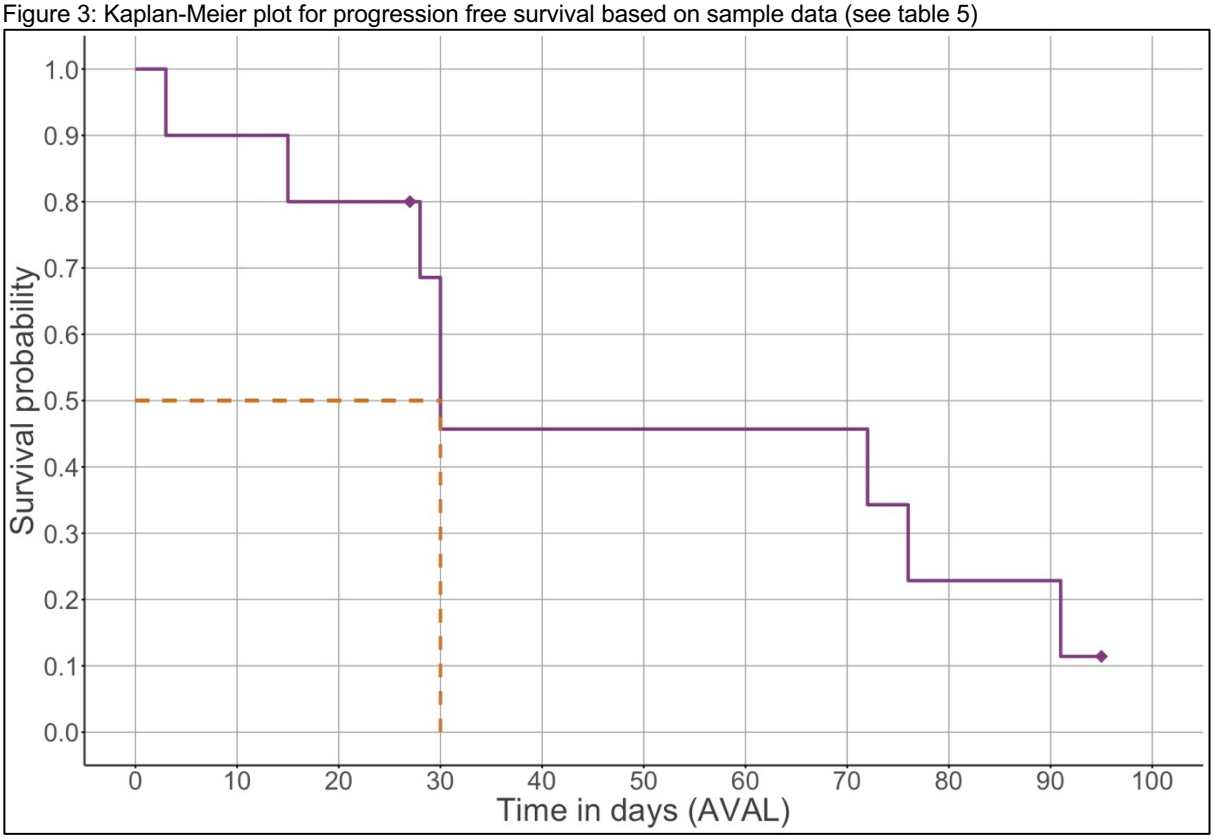
The KM estimates change in comparison to OS because of the different definition of an event. Table 6 shows the KM estimates for PFS.

Table 6: Kaplan-Meier estimates for progression free survival based on sample data (see table 5)

Event timepoint	Number of subjects at risk	Number of subjects with an event	KM estimate
3	10	1	$\hat{S}(3) = \frac{10-1}{10} = 0.9$
15	9	1	$\hat{S}(15) = 0.9 \times \frac{9-1}{9} = 0.8$
28	7	1	$\hat{S}(28) = 0.8 \times \frac{7-1}{7} = 0.686$
30	6	2	$\hat{S}(30) = 0.686 \times \frac{6-2}{6} = 0.457$
72	4	1	$\hat{S}(72) = 0.457 \times \frac{4-1}{4} = 0.343$
76	3	1	$\hat{S}(76) = 0.343 \times \frac{3-1}{3} = 0.229$
91	2	1	$\hat{S}(91) = 0.229 \times \frac{2-1}{2} = 0.114$

The corresponding KM plot for PFS is shown in figure 3. In comparison to OS the survival curve tends to be shifted to the left. The median survival of 50% is already reached by day 30.

Figure 3: Kaplan-Meier plot for progression free survival based on sample data (see table 5)



CONCLUSION

Working as a statistical programmer in the therapeutic area of oncology, it is important to know the purpose and the structure of the CDISC ADTTE domain. In this paper we aim to explain how time-to-event data need to be stored in the ADTTE to enable a valid survival analysis. Furthermore, it helps the programmer to understand the concept of censored data. This concept is shown for two widely used endpoints in oncological trials, overall survival and progression free survival. To comprehend the Kaplan-Meier approach in detail is an advantage when it comes to developing programs in order to apply this method. In this paper, example data is used to show the cumulative structure of the Kaplan-Meier estimator step by step. Finally, Kaplan-Meier plots are shown to enhance the understanding of the method further.

REFERENCES

CDISC Analysis Data Model (ADaM) Team: Analysis Data Model (ADaM). Version 2.1. 2009 (http://www.cdisc.org/adam)

CDISC Analysis Data Model (ADaM) Team: The ADaM Basic Data Structure for Time-to-Event Analysis. Version 1.0. 2012 (http://www.cdisc.org/adam)

CDISC Analysis Data Model (ADaM) Team: Analysis Data Model (ADaM) Implementation Guide. Version 1.1. 2016 (http://www.cdisc.org/adam)

Cheson BD, Pfistner B, Juweid ME, Gascoyne RD, Specht L, et al. Revised Response Criteria for Malignant Lymphoma. Journal of Clinical Oncology. Vol 25 Number 5. 2007.

Eisenhauer E.A., et al. New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1.1). European Journal of Cancer. Vol 45. 2009. (http://www.eortc.be/recist/)

EvaluatePharma®. Word Preview 2017, Outlook to 2022. 10th Edition June 2017.

Ferlay J, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11, Lyon, France: International Agency for Research on Cancer. 2013

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481. doi:10.1080/01621459.1958.10501452)

CONTACT INFORMATION

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

Frieder Wolff
Chrestos Concept GmbH & Co.KG
Girardetstr. 1-5
45131 Essen
Work Phone: +49 175 / 2139780
Email: frieder.wolff@chrestos.de
Web: http://www.chrestos.de

Vitali Gering
Chrestos Concept GmbH & Co.KG
Girardetstr. 1-5
45131 Essen
Work Phone: +49 160 / 97761157
Email: vitali.gering@chrestos.de
Web: http://www.chrestos.de