A New Scheme to Estimate Median Progression-free Survival Time in Oncology Clinical Trials

Peng Li, Xing Sun, Cong Chen and Anderson Keaven *

Apr, 2013

Abstract

Progression-free survival (PFS) has been accepted as a valid primary endpoint in many previous anti-tumor therapeutic product marketing approvals. Yet, it has been long known that current standard statistical methods to handle PFS data widely adopted in pharmaceutical industry may lead to serious bias under certain circumstances. Although most statistical issues related to two or multiple treatment group comparison problem for PFS data can be resolved (c.f. [10]), the challenge to provide a practical and better median PFS estimation still remains, which is an even more important question to address in order to get a thorough understanding of the treatment effect in everyday medical practice. In this paper, the author proposes a simple alternative trial design idea together with a new estimation scheme to solve this issue. The performance of the new method based on this new scheme will be compared to the standard method currently used by the pharmaceutical industry via simulations under typical oncology study settings.

Keywords: survival analysis, interval censor, progression-free survival, Kaplan-Meier, NPMLE

1 Introduction

More and more clinical trials in oncology clinical research area involve interval censored time-to-event type of data [2, 13]. Popular data of this kind in oncology clinical researches includes progression-free survival(PFS), time to disease progression(TTP), and disease-free survival(DFS). A feature in common with this type of data is that the event of interest is either right censored or the exact time of its occurrence is unknown. Only the time interval in which the event occurs is observed, where the left-point of the time interval represents the last time the individual is known to be event-free and the right-point of the interval represents the earliest time that the individual is observed with an event. In statistical literatures, this type of data is often called case II interval-censored data and it is also often assumed that interval-censoring is non-informative (see [9]), i.e., the censoring interval does not provide any additional information nor depends on the distribution of event time T.

As the number of applications of PFS (TTP) as endpoints to demonstrate anti-tumor effect in clinical trials increases[2, 13], researches have been re-drawn to investigate the statistical methods used to analyze this type of data in real world practice. Despite that the first non-parametric maximum likelihood (NPMLE) based method to estimate the survival function [12] has been available for almost four decades, the most prevalent approach in pharmaceutical industry is still imputation based, i.e., the true unknown event time is replaced by the recorded event time and then the standard survival analysis approach for right-censored data is applied. An important reason that the imputation based approach is widely used today is the computational inconvenience of these

*Peng Li and Xing Sun are with Sanofi & Co., Inc., 5F, No. 112 Jian Guo Lu, Beijing, China, 100022. Email: peng.li, xing.sun@sanofi.com. Cong Chen and Anderson Keaven are with Merck & Co., Inc., 351 N Sumneytown Pike, North Wales, PA, 19454, US. Email: cong_chen, anderson_keaven@merck.com.

NPMLE methods where interval-censored data has to be analyzed as it is without any imputation. Research [10] has shown that the naive imputation methods are far from optimal and often lead to erroneous results in multiple group comparison problems under realistic clinical trial settings. In contrast, the NPMLE methods (see [1, 3, 5, 9, 11, 14] and the references therein) can perform much better and be more robust. Recent guidance issued by EMEA on PFS endpoints [2] also encourages applying these NPMLE approaches in two (multiple) treatment group comparison analysis as a supplemental analysis.

Besides investigations mentioned above on multiple group comparison problems, not much has been done as a practically orientated research for survival curve estimation especially the median event time estimation. On the other hand, the median time estimation is crucial in evaluation for anti-tumor therapy approvals. As observed in many recent clinical studies(e.g. [4, 7]), current imputation based approach may artificially dampen the median PFS (TTP) differences between treatment groups, and lead to results that contradict to what had been implied by the associated HRs or other measures. It is also known that current approach (i.e. using right-point imputation) will artificially prolong the median PFS (TTP) time for each group [6]. In this paper, we will compare the performance of several available median estimation methods and also propose a new but simple study design framework under which a better median estimate can be achieved.

The rest of this paper is organized into 5 sections. Section 2 is a review of the statistical methods related to single curve estimation for interval-censored time-to-event data. Section 3 introduces our new study design framework for better median PFS(TTP) estimate. In Section 4, we will compare the performance of these methods through simulations. A discussion will follow in Section 5. All detailed simulation results can be found in Appendix.

2 Review of Available Methods

In PFS or TTP analyses, given n interval-censored event time, one can represent them by $(L_i, R_i]$ where L_i is the last known event-free observation and R_i is the first known time with an event. If the exact time for this event is known then $L_i = R_i$. If this event is right censored then L_i is the last assessment time and $R_i = \infty$. The standard imputation based methods used in pharmaceutical industry is to impute an unknown exact event time with R_i if the event is truly interval-censored. If an event is right-censored, L_i will be used as the right-censoring time. Standard Kaplan-Meier estimates can then be computed to estimate the median. It is well known that this type of estimate prolong the median PFS or TTP[6]. The amount of time artificially extended can be as large as the length of an entire assessment interval. Also when comparing median PFS time between two groups, the difference between two groups however may equal to zero if the true median PFS time of both happen to be different but within the same assessment interval.

Different than the standard imputation based approach, NPMLE does not use any fixed point to impute the data. Instead it treats the data as it is. To illustrate its idea, one can reformulate the case II interval censored data as $\{(s_j, s_{j+1}],\ 0 = s_1 < ... < s_m = \infty\}$, where $\{s_j, j = 1, ..., m\} \in \{\{L_i\} \cup \{R_i\}, i = 1, ..., n\}$. The logarithm of nonparametric full likelihood proposed by Turnbull [12] is therefore

$$L = \sum_{i=1}^{n} \log\left(S(L_i) - S(R_i)\right) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} (S(s_{j-1}) - S(s_j))\right). \tag{1}$$

Maximizing L gives NPMLE $\{S(s_j),\ j = 1, ..., m-1\}$. Turnbull [12] first proposed an EM algorithm to obtain the NPMLE based on this logarithm of nonparametric full likelihood function. Since then many researches have been done on related topics (see [11, 14] for details). It is also noted that the NPMLE may not be uniquely determined on certain intervals. This may happen for time intervals when there is no observed $(L_i, R_i]'s$ overlapping with this interval or when only right-censored observations (L_i, ∞) overlapping with this interval. Another difference from the standard survival

analysis methods for right-censored data is that the survival distribution is given for each time intervals $(s_{j-1}, s_j],\ j = 1, ..., m$ instead of specific timepoints. Similar to the standard fixed point imputation approach, one can still use the right-point of the resulting time interval to represent the associated quantile of the survival distribution even this is a suboptimal solution. In this paper, we will use this right-point representation in the simulation when comparing all these methods.

3 Design with Shifted Assessments

In two or multiple sample comparison problems NPMLE methods can largely improve the performance of statistical testing and HR estimations. However, NPMLE methods alone can not make too much improvement in single curve estimation as shown in the simulations in Section 4. The reason is that NPMLE can not make any finer estimate within an interval $(s_{j-1}, s_j]$. It is natural to seek for designs with more frequent disease progression(PD) assessments, which have less information loss. In clinical trial practice, this is often infeasible as the evaluation cost increases rapidly and the visit frequency may also be further constrained by patient's convenience or comfort. To accommodate this desire to have finer interval while keeping almost the same number of assessments, we propose a new design framework with shifted assessment. The idea is illustrated as below.

Suppose that n patients are enrolled into one treatment group in a study with a PD assessment interval I. We randomly and equally assign these n patients into four groups with different assessment schedules,

A.1 Assessed at time $\{0.5I, 1.5I, 2.5I, ...\}$,
A.2 Assessed at time $\{0.75I, 1.75I, 2.75I, ...\}$,
A.3 Assessed at time $\{I, 2I, 3I, ...\}$,
A.4 Assessed at time $\{1.25I, 2.25I, 3.25I, ...\}$.

Then NPMLE of the survival curve will be calculated with the observed data. The right-point of the interval with estimated cumulative survival probability closest to 0.5 will be used as the median PFS or TTP time estimate. The lower and the upper CI for this generalized survival curve can also be constructed for each interval $(s_{j-1}, s_j]$ based on the variance estimation method proposed in [8] via a resampling technique.

Note that the allocation into four assessment schedules is just for illustrative purposes. The statistician should decide the number of different schedules and the allocation ratios based on disease background information and clinical practice convenience at the time of design. Nevertheless, under this proposed framework the total number of disease assessments remains almost the same but the estimation intervals $(s_{j-1}, s_j]$ becomes much shorter. As shown in Section 4 by simulations, this new framework can provide better median estimations under certain suitable oncology trial settings.

4 Simulation

Simulations were conducted to compare the performance of the methods introduced in previous sections for median PFS time estimation. We generated simulation data with size analogous to typical oncology Phase III clinical studies with 200 patients in a treatment arm. The assessment schedule was assumed as every 2 months until Month 16 in the simulations when event time T followed exponential distributions. Simulated events after the last assessment were right-censored at the last assessment time. There was also data simulated to be right-censored before the last scheduled assessment time as Month 16. Approximate 15%- 25% of the simulated data was randomly right-censored. The most recent assessment time before each right-censoring was chosen as the right-censoring time per common practice in PFS analysis. A thousand random repeats were generated for each assumed true median PFS time T_{Median}. As the true median T_{Median} is unknown in real clinical trials, it is important to evaluate the overall performance of these estimation methods for many different T_{Median}s. In our simulation, T_{Median} takes values from Month 4 to Month

6. During each random repeat, the same simulated dataset of exact event time was used to derive interval-censored PFS data using both the non-shifted assessment schedule and the shifted assessment schedule. For the shifted assessment schedule, each 200 generated data was randomly and equally assigned into 4 different schedules listed below. Both the standard right-point imputation based method and the NPMLE method were applied to each interval-censored data.

S.1 First assessment at Month 2 then every two months until Month 16;
S.2 First assessment at Month 2.5 then every two months until Month 16.5;
S.3 First assessment at Month 3 then every two months until Month 17;
S.4 First assessment at Month 3.5 then every two months until Month 17.5.

Empirical mean and CI of the median estimates and the MSE are shown in Table 1. Rates that confident intervals (CIs) constructed by each method covering true median time and the empirical mean and variance of the width of the constructed CI are presented in Table 2.

Note that the statistical inference on the median estimation in survival analysis is often based on normal approximation. To assess the normality assumption with median estimates obtained by different methods, a histogram sample is presented in Figure 1 for selected assumed true median time.

We also compared the above methods in simulations when event time T follows Weibull distribution and in simulations when there are random perturbations to the scheduled assessment time to simulate voluntary or involuntary deviations from the assessment schedule in real clinical trial practice. These simulations were set up in a similar way to the simulation described above. Due to the page limit of this paper, their results will only be shown in future publications.

5 Discussion

As demonstrated by the simulation results in Table 1, NPMLE with shifted assessment schedule can lead to better estimation of the median event time for interval-censored data such as PFS or TTP for a wide range of scenarios. Particularly, the NPMLE with shifted assessment schedule outperforms other estimation approaches consistently across the entire interval of the unknown true median time, while other approaches, especially methods with non-shifted schedules overestimate the true median time in a very variable fashion depending on where the true median time locates. NPMLE with shifted schedule also has a smaller MSE. Note that the constructed CIs by NPMLE with shifted schedules are wider than the CIs from KM methods. This is due to the additional resampling variance term in the variance calculation formula for NPMLE [8]. However, the coverage rate in Table 2 shows that the CIs by NPMLE with shifted schedules can provide more robust and relatively accurate coverages than CIs by other methods. This is partly due to the fact that the distribution of the median estimate from NPMLE with shifted schedules is better approximated by a normal distribution as shown in Figure 1. In another word, although the MSE can not go to zero even for NPMLE with shifted schedules since the associated median estimates will always be from a number of discrete timepoints not necessarily containing the true median, the distribution of the median estimate can still be more concentrated and symmetrically distributed around the true median.

One may expect that the standard imputation based approach i.e. KM estimator, together with a shifted assessment schedule could perform better. However, the improvement by this method is very limited. This is due to the fact that the estimation process from this approach completely ignores the information provided by the left-points of the interval. The nature in its estimating process does not reduce the interval length of the data, where the uncertainty associated with the estimate is determined by the interval length and the accuracy of the estimates improves only when the interval length can get smaller. Note that in the simulation presented in this paper where no deviation to the assessment schedule exist, there is no difference in point estimation of median between the standard imputation based method and the NPMLE method as expected by the algorithm of NPMLE. However these two methods perform differently when deviations do exist.

Although this paper shows that the proposed median estimate method based on shifted assessment schedules seems promising to provide more accurate estimation of median event time for interval-censored data such as PFS or TTP, more factors should be considered when applying it in a clinical trial cautiously. It should not be applied to a rapidly progressed disease where disease progression may likely happen before the first assessment. A shifted schedule may further delay the discontinuation of progressed patients from the study and delay their entries into other therapies. Another factor to consider is the operational challenge in assessment visit management. Since different patients may be on different assessment schedules, monitoring their visits can be more intensive than monitoring visits following a same assessment schedule.

As explained earlier results presented in this paper are limited to exponential distribution and no deviation to visit schedule due to the page limit. Simulation results from the Weibull distribution and the case with random deviations are consistently showing that NPMLE with shifted schedule can provide better estimate of median PFS or TTP time.

References

[1] Chen, D., Sun, J. and Peace, K. (2012). Interval-censored Time-to-Event Data: Methods and Applications. *Chapman Hall/CRC*.

[2] EMEA CHMP (2007). Methodological Considerations for Using Progression-free Survival (PFS) as Primary Endpoint in Confirmatory Trials for Registration.

[3] Fay M. (1999). Comparing Several Score Tests for Interval-Censored Data. *Statistics in Medicine*, **18**, 273-285.

[4] Hodi, F.S., O'Day, S.J., McDermott, D.F. and *et. al.*(2010).Improved survival with ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine*, **363(13)**, 1290.

[5] Huang J. and Wellner J.A. (1997). Interval Censored Survival Data: A Review of Recent Progress. *In Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis,*

[6] Panageas K.S., Ben-Porat L., Dickler M.N., Chapman P.B., and Schrag D. (2007). When You Look Matters: the Effect of Assessment Schedule on Progression-free Survival. *Journal of National Cancer Institute*, **99**, 428-432.

[7] Robert, C, Thomas, L., Bondarenko, I. and *et. al.*(2011).Ipilimumab plus Dacarbazine for Previously Untreated Metastatic Melanoma. *New England Journal of Medicine*, **364**, 2517-2526.

[8] Sun J. (2001). Variance Estimation of a Survival Function for Interval-Censored Survival Data. *Statistics in Medicine*, **20**, 1249-1257.

[9] Sun J. (2006). The Statistical Analysis of Interval-censored Failure Time Data. *Statistics for Biology and Health. Springer, New York, NY.*

[10] Sun X. and Chen C. (2010). Comparison of Finkelstein's Method with the Conventional Approach for Interval-Censored Data Analysis. *Statistics in Biopharmaceutical Research*, **2**, **1**, 97-108.

[11] Sun, X., Li, XY, Chen, C. and Song, Y.(*to appear*). A Review of Statistical Issues with Progression-Free Survival as an Interval-Censored Time-To-Event Endpoint . *Journal of Biopharmaceutical Statistics.*

[12] Turnbull B.W. (1976). The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society, Series B*, **38**, 290-295.

[13] US FDA (2007). Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics.

[14] Zhang Z. and Sun J. (2010). Interval Censoring. *Statistical Methods in Medical Research*, **19**, 53-70.

6 Appendix

Table 1: Empirical mean, 95% CI and MSE of median estimate

True median	Mean (95% CI)				MSE			
	KM	NP	KM*	NP*	KM	NP	KM*	NP*
4	4.57 (4.0, 6.0)	4.57 (4.0, 6.0)	4.73 (4.0, 5.5)	4.03 (3.0, 5.0)	1.14	1.14	0.72	0.32
4.2	4.94 (4.0, 6.0)	4.94 (4.0, 6.0)	4.95 (4.0, 6.0)	4.23 (3.0, 5.5)	1.55	1.55	0.77	0.34
4.4	5.29 (4.0, 6.0)	5.29 (4.0, 6.0)	5.13 (4.5, 6.0)	4.42 (3.5, 5.5)	1.71	1.71	0.76	0.37
4.6	5.56 (4.0, 6.0)	5.56 (4.0, 6.0)	5.30 (4.5, 6.5)	4.58 (3.5, 6.0)	1.61	1.61	0.73	0.39
4.8	5.79 (4.0, 6.0)	5.79 (4.0, 6.0)	5.51 (4.5, 6.5)	4.81 (3.5, 6.0)	1.38	1.38	0.78	0.39
5	5.91 (4.0, 6.0)	5.91 (4.0, 6.0)	5.70 (5.0, 6.5)	5.00 (4.0, 6.5)	1.06	1.06	0.77	0.43
5.2	6.00 (4.0, 8.0)	6.00 (4.0, 8.0)	5.90 (5.0, 7.0)	5.17 (4.0, 6.5)	0.86	0.86	0.81	0.47
5.4	6.13 (6.0, 8.0)	6.13 (6.0, 8.0)	6.08 (5.0, 7.0)	5.36 (4.0, 6.5)	0.85	0.85	0.80	0.48
5.6	6.23 (6.0, 8.0)	6.23 (6.0, 8.0)	6.29 (5.0, 7.5)	5.58 (4.5, 7.0)	0.84	0.84	0.83	0.50
5.8	6.41 (6.0, 8.0)	6.41 (6.0, 8.0)	6.48 (5.5, 8.0)	5.79 (4.5, 7.5)	1.04	1.04	0.88	0.58
6	6.65 (6.0, 8.0)	6.65 (6.0, 8.0)	6.68 (5.5, 8.0)	5.99 (4.5, 7.5)	1.31	1.31	0.85	0.57

*: with shifted assessment schedule

Table 2: Coverage rate, width of constructed 95% CI

True median	95% CI coverage Rate (%)				Mean width of 95% CI				SD width of 95% CI			
	KM	NP	KM*	NP*	KM	NP	KM*	NP*	KM	NP	KM*	NP*
4	98.7	99.5	71.4	95.3	1.9	1.9	1.6	1.7	0.5	0.5	0.4	0.6
4.2	95.4	96.1	54.1	87.9	2.0	2.0	1.7	1.8	0.4	0.5	0.4	0.7
4.4	91.5	93.9	38.4	87.7	2.0	2.0	1.7	1.9	0.6	0.6	0.4	0.7
4.6	85.2	90.8	73.5	88.6	2.0	2.2	1.8	2.0	0.7	0.7	0.4	0.6
4.8	74.8	80.3	58.1	88.8	2.1	2.2	1.9	2.1	0.9	0.9	0.4	0.7
5	57.1	67.2	82.3	95.6	2.0	2.3	2.0	2.2	1.0	1.0	0.5	0.7
5.2	44.4	55.8	71.5	91.6	2.0	2.3	2.1	2.3	1.0	1.0	0.5	0.7
5.4	33.3	43.6	57.5	90.2	2.1	2.4	2.2	2.3	0.9	1.0	0.5	0.8
5.6	22.4	30.2	79.2	91.5	2.2	2.3	2.2	2.5	0.9	0.9	0.5	0.7
5.8	11.9	18.1	69.8	91.6	2.2	2.3	2.3	2.5	0.8	0.9	0.5	0.8
6	98.9	99.4	85.3	95.9	2.3	2.4	2.4	2.6	0.8	0.9	0.5	0.8

*: with shifted assessment schedule

Figure 1: Histogram of median estimate