

Survival data extraction: the comparison of methods, risk proportionality and informative censoring

I-080

M&S

MODELING & SIMULATION DECISIONS

Kseniia Danichkina¹, Kirill Zhudenzov^{2,3}, Kirill Peskov^{2,3,4}, Ekaterina Khrameeva¹

1 - Skolkovo Institute of Science and Technology, Moscow, Russia; 2 - M&S Decisions FZ-LLC, Dubai, UAE;
3 - I.M. Sechenov First Moscow State Medical University, Moscow, Russia; 4 - Marchuk Institute of Numerical Mathematics RAS, Moscow, Russia

Objective

The objectives of the presented work included: a development of a seamless workflow for IPD data digitization and analysis: identification and comparison of efficient algorithms for the published KM curves in clinical trials; application the diagnostic tests to check the proportionality of hazards to verify the consistency of rank tests and hazard ratio (HR) assessment; performing imbalanced censoring tests and sensitivity analysis to identify possible bias in survival comparison outcomes.

Introduction

- Survival analysis is a gold standard of treatment efficacy assessment in oncology clinical trials. Individual patient data (IPD) are integrated to construct Kaplan-Meier (KM) curves representing published data for patient survival [1]. The basic assumptions used in survival analysis are the proportionality of hazards (PH) (change in probability of event due to change in covariate does not depend on time) and non-informativeness of censoring (dropouts occur completely at random or at least with similar intensity across arms compared) [2]. However, this is not always the case in various clinical trials [3-5]. Here, we investigate and compare available methods of published KM curves digitization and IPD retrieval [6 - 9].

Methods

Survival data representation

- The survival and progression data were obtained from a set of representative published clinical trials in advanced cancer [10-12,16-21]. Pairs of KM curves from different patient subgroups, representing comparison of investigated drugs and control with various sample sizes, were digitized using the PlotDigitizer software [13].

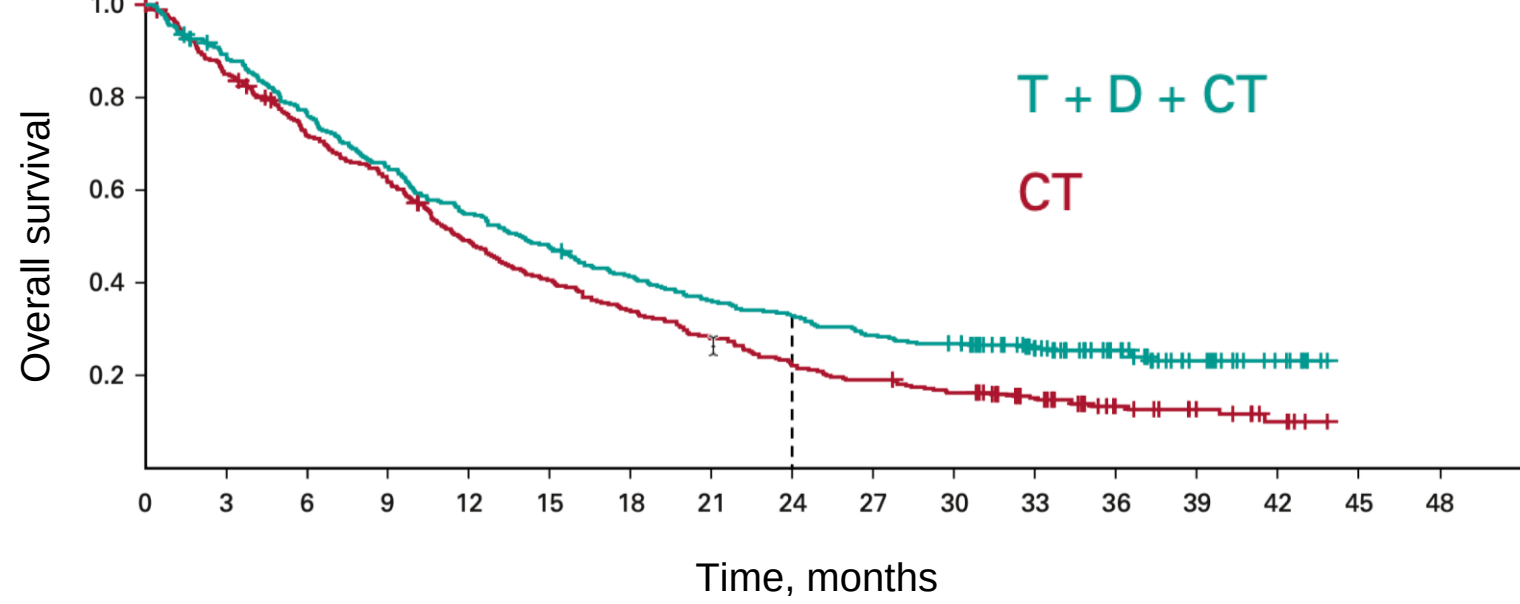


Figure 1. Exemplar published Kaplan-Meier estimator of survival function for study NCT03164616 [11]. T – tremelimumab, D – durvalumab, CT – chemotherapy.

IPD retrieval and hazard proportionality

- The Liu IPD reconstruction algorithm requires most extensive curve shape tracing as well as total number of events and subjects. Risk table introduction is optional [7].
- The Rogula IPD method requires the tracing of curve steps only while requires total number of subjects and exhaustive tracing of censoring times [6].
- Reconstruction adequacy implied the overlay of original and reconstructed KM curves. Median survival times, published HRs and their uncertainty were also compared.
- The PH assumption (as well as the reliability of the log-rank test and HR estimates) was verified using the Schoenfeld residuals.

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

T – time when event occur
 t – current moment

Proportional hazards assumption

$$\lambda(t|x) = \lambda_0(t)e^{x'\beta}$$
$$\frac{\lambda(t|x=1)}{\lambda(t|x=0)} = \exp(\beta) = \text{const.} = \text{HR}$$

$$S(t) = \exp\left\{-\int_0^t \lambda(u)du\right\}$$
$$r_{Sk} = \sum_i \delta_{ik}(x_i - x_{avrg}(\beta, tk))$$

for all i (patients at risk at tk)

Schoenfeld residuals

Figure 2. The concept of proportional hazards (PH) as well as Schoenfeld residuals assessment for PH violation.

Censoring analysis

- Imbalanced censoring tests for reconstructed IPD and sensitivity analysis were performed using the reverse KM method described in Gilboa et al. [5]. All analyses were held in R [14].

Figure 3. Reconstructed IPD (left) and imbalanced censoring adjusted KM (right). See Figure1 for original KM data. Log-rank test p-values are provided.

Results

Data exploration

- The basic analysis set included the OS data from advanced lung cancer clinical studies [10 - 12] of immune therapy: (atezolizumab, durvalumab, tremelimumab and relevant combinations) compared to chemotherapy (docetaxel, cisplatin, pemetrexed and relevant combinations)
- To extend the analysis of balanced censoring and focus on earlier assessment times, the PFS outcomes of other studies on advanced cancer were also incorporated [16 - 21].

Table 1. The sample sizes of published survival data used in the analysis [10-12, 16-21] and reversed HR comparison for 2 IPD reconstruction methods

Study code [ref], endpoint	Population/Subgroup	Total N	Investigated treatment	Control	Reversed HR		Result of sensitivity analysis (new HR)	
					Liu method	Rogula method	Liu method	Rogula method
NCT02657434 [10], OS	ITT	578	292	286	0.80 (0.38 - 1.68)	0.60 (0.25 - 1.47)		
	PD-L1 high	45	25	20	0.70 (0.18 - 2.82)	0.52 (0.12 - 2.34)		
	PD-L1 low	136	63	73	0.85 (0.20 - 3.56)	1.21 (0.32 - 4.53)		
	PD-L1 negative	163	88	75	0.88 (0.18 - 4.34)	0.90 (0.13 - 6.37)		
NCT03164616 [11], OS	T+D+CT & CT	675	338	337	0.26 (0.10 - 0.72)	0.69 (0.24 - 2.00)	0.87 (0.73 - 1.03)	
	D+CT & CT	675	338	337	0.66 (0.32 - 1.39)	0.59 (0.19 - 1.79)		
NCT02008227 [12], OS	ITT	850	425	425	2.39 (1.73-3.32)*	0.68 (0.38 - 1.24)	1.30 (1.07 - 1.58)	
	PD-L1 ≥ 1%	463	241	222	0.41 (0.22 - 0.78)	0.45 (0.22 - 0.95)	0.94 (0.73 - 1.15)	0.89 (0.71 - 1.11)
	PD-L1 ≥ 5%	265	129	136	0.26 (0.11 - 0.64)	0.36 (0.13 - 1.01)	1.03 (0.75 - 1.40)	
	PD-L1 ≥ 50% TC or ≥10% IC	137	72	65	0.50 (0.14 - 1.78)	0.58 (0.16 - 2.18)		
	PD-L1 < 1%	379	180	199	0.61 (0.31 - 1.22)	1.16 (0.51 - 2.63)		
	Non-squamous	628	313	315	0.40 (0.23 - 0.69)	0.65 (0.34 - 1.23)	0.91 (0.74 - 1.11)	
	Squamous	222	112	110	0.90 (0.34 - 2.41)	1.04 (0.38 - 2.87)		
	ITT	154	92	62	0.59 (0.30 - 1.13)	0.49 (0.23 - 1.06)		
NCT02184195 [16], PFS	ITT	1248	623	625	0.74 (0.38 - 1.42)	0.51 (0.22 - 1.18)		
NCT00262847 [17], PFS	ITT	1971	1137	834	1.26 (1.09 - 1.46)*	0.52 (0.43 - 0.62)*	1.07 (0.92 - 1.24)	0.63 (0.56 - 0.71)
ISRCTN49407852 [18], PFS	ITT	733	487	246	0.99 (0.76 - 1.29)	0.55 (0.39 - 0.79)		
NCT02655016 [19], PFS	ITT	392	261	131	—	0.62 (0.28 - 1.38)	—	
NCT01321554 [20], PFS	ITT	302	205	97	—	0.72 (0.44 - 1.18)	—	
NCT01524783 [21], PFS	ITT							

ITT – Intention to treat population; TC – Tumor cell; IC – Tumor-infiltrating immune cell; PD-L1 – Programmed Death-Ligand 1. IPower132 (NCT02657434)[10]; POSEIDON (NCT03164616)[11]; OAK (NCT2008227)[12]; POLO (NCT02184195) [16]; GOG-0218 (NCT00262847) [17]; Myeloma XI (ISRCTN49407852) [18]; GOG-3012 (NCT02655016) [19]; SELECT (NCT01321554) [20]; RADIANT-4 (NCT01524783) [21]

IPD extraction efficiency

- Both tested digitization methods expressed similar and adequate KM representation
- The outcomes of rank comparison tests and estimated HRs showed fine coincidence with the published outcomes:

Table 2. Selected HRs for studies [10-12]: published, Liu and Rogula method results within 95%CI

Study code [ref]	Population	Published	Liu method	Rogula method
NCT02657434 [10]	ITT	0.86 (0.71-1.06)	0.87 (0.71-1.06)	0.88 (0.72-1.07)
NCT03164616 [11]	T+D+CT & CT	0.77 (0.65-0.92)	0.78 (0.66-0.93)	0.77 (0.65-0.92)
	D+CT & CT	0.86 (0.72-1.02)	0.84 (0.71-0.995)	0.84 (0.71-0.99)
NCT02008227 [12]	ITT	0.73 (0.62-0.87)	0.73 (0.61-0.87)	0.67 (0.57-0.78)

Strikethrough – PH is violated; HR is not reliable (see Table 3)

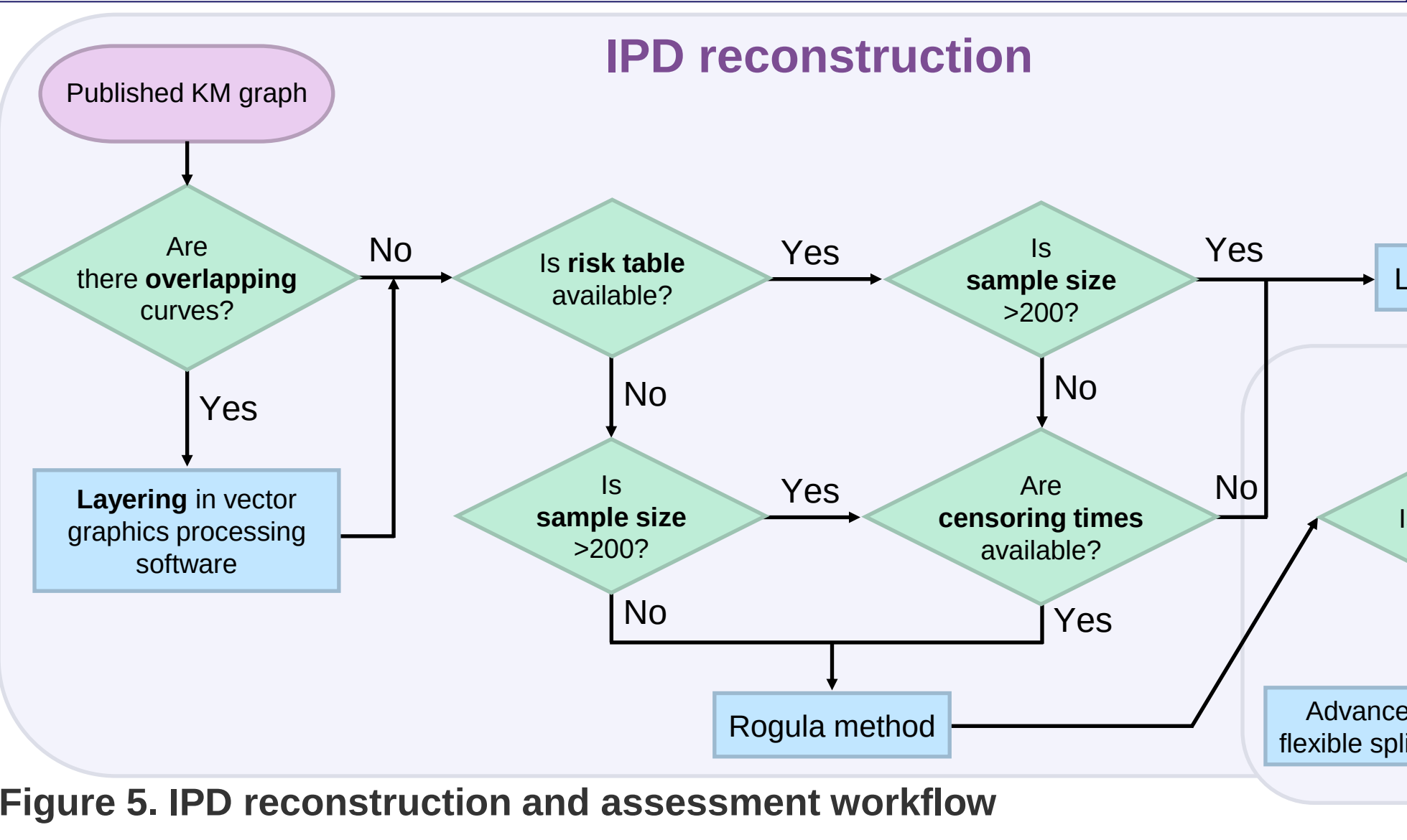


Figure 5. IPD reconstruction and assessment workflow

Hazard proportionality assessment

- However, hazard proportionality was violated in 7 pairs of all tested KM curves [10 - 12, 16 - 21], resulting in a possible bias in the rank comparison test results as well as HR (both published and estimated from reconstructed IPD).

Table 3. Schoenfeld p-values indicating PH violation

Study code [ref], population	Schoenf. p, Liu	Schoenf. p, Rogula
NCT02657434 [10], PD-L1 high	0.03	0.03
NCT03164616 [11], T+D+CT&CT	0.04	0.03
NCT02008227 [12], ITT	—	0.08
NCT00262847 [17], ITT	2.8e-07	1.4e-07
ISRCTN49407852 [18], ITT	—	0.04
NCT02655016 [19], ITT	—	0.03
NCT01524783 [21], ITT	—	5.2e-04

Figure 4. Detected violation of PH assumption

Table 4. Appropriate rank tests to compare KM curves while PH is not valid

Test name	NCT02657434 [10], High PD-L1; p > 0.05		NCT03164616 [11], T+D+CT & CT; p = 3E-3		NCT02008227 [12], ITT; p = 3E-4	
	Liu p	Rogula p	Liu p	Rogula p	Liu p	Rogula p
Log-rank test	0.4752	0.4766	0.0043	0.0031	0.0004	0E+00
Peto-Peto's modified	0.2757	0.2778	0.0540	0.0478	0.0023	2E-04

Regular log-rank test imposes equal weights through the timeline and strictly requires the validity of PH assumption. In case of PH violation the following tests are applicable: Gehan-Breslow (generalized Wilcoxon) — very heavy in the beginning, PH should be either <1 or >1 all the time; Tarone-Ware — heavy beginning, equal population variance assumption is needed; Peto-Peto's modified and modified Peto-Peto (by Andersen) — slightly heavy beginning, no requirements addressing PH, more resistant than Gehan-Breslow. In case of NCT03164616 [11] (T+D+CT & CT) Peto-Peto tests show the loss of significance

Imbalanced censoring outcomes

- Censoring diagnostics by means of reverse KM method [5] revealed 2 cases of imbalanced censoring among 19 curve pairs for both methods and 5 additional signals for the IPD reconstructed using the less robust Liu digitizing method, not sensitive to exact censoring times (Table 1).
- Sensitivity analysis led to lost of substantial clinical result in 1/2 case for both methods and in 4/5 cases specific for Liu.

Conclusions

- Here we presented a workflow for the IPD reconstruction along with necessary inspections for the data prior to the qualification of survival models (Figure 6). Both tested methods for IPD reconstruction showed high performance at a qualitative level for various sample sizes.
- The Rogula method was more robust, though it required more information on censoring of subjects. In certain cases, the data showed a violation of hazard proportionality. Thus, in further analysis, advanced survival models should be developed [15].
- Finally, the analysis of censoring revealed additional differences in treatment effects that would require further application of competing risk models and other advanced approaches.

References

1. Fojo T and Simon R (2021). Lancet Oncol, 22(10), 1358 – 1360;
2. Emmerson J and Brown J (2021). Clinical Oncology, 33(1), 12 - 14;
3. Rosen K et al. (2020). Eur J Cancer, 141, 152 – 161.
4. Templeton A et al. (2020). Nat Rev Clin Oncol, 17(6), 327 - 328;
5. Gilboa S et al. (2021). Eur J Cancer, 153, 190-202;
6. Rogula B et al. (2022). MDM Policy Pract, 7(1), 23814683221077643.
7. Liu N et al. (2021). BMC Med Res Methodol, 21(1), 111;
8. Guyot P et al. (2012). BMC Med Res Methodol, 12, 9;
9. Wei Y and Royston P (2017). Stata J., 17(4), 786 – 802.
10. Nishio M et al. (2021). J Thorac Oncol, 16(4), 653–664;
11. Johnson M et al. (2023). J Clin Oncol, 41(6), 1213–1227;
12. Rittmeyer A et al. (2017). Lancet, 389(10066), 255–265.
13. PlotDigitizer website: <https://plotdigitizer.sourceforge.net/>;
14. R Core Team (2024). <https://www.R-project.org/>;
15. Bardo M et al. (2024). Statistical Methods in Medical Research, 33(6), 1069-1092.
16. Ilson D et al. (2019) NEJM Journal Watch, NA49254;
17. Burger R et al.(2011). NEJM, 365(26), 2473-2483;
18. Jackson G et al.(2019). The Lancet Oncology, 20(1), 57-73.
19. González-Martin A et al. (2019). NEJM, 381(25), 2391-2402;
20. Schlumberger M et al. (2015). NEJM, 372(7), 621-630;
21. Yao J et al. (2016). The Lancet, 387(10022), 968-977.

