Poster #: MSR124

Improving the efficiency of using the Guyot algorithm to construct pseudo individual patient data from Kaplan-Meier survival curves: Standardized guidance and data preparation

Emily Aiello,[1] Ioana Gulas,[1] Emma Mackay,[1] Victor Laliman-Khara,[1]
[1]Cytel, Inc., Toronto, ON, Canada

The Guyot Algorithm drives high variability in user implementation due to the nuances in user tracing of published curves and manual edits of digitizer application output. The efforts of this study improved accuracy and standardization when using Guyot to generate pseudo IPD.

Background

- The Guyot method[1] is a widely used algorithm to generate pseudo individual patient data (IPD) from published Kaplan-Meier (KM) curves for post-hoc analyses to compare the relative efficacy of drugs, perform survival assessments and make predictions of projected survival probabilities for the cost-effectiveness model.
- The algorithm imposes restrictions on the input data, for example it assumes that the survival probabilities are monotonically decreasing or that the number at risk table reports the number of subjects at time zero.
- Mixed published guidance on curve tracing and implementation of the algorithm leads to frequent errors and inaccuracies in the final IPD.

Objective

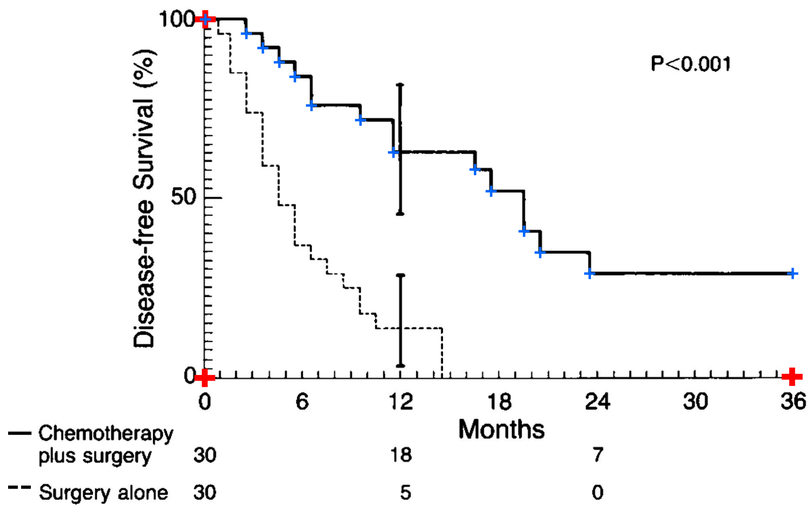
- This study aimed to provide standardized guidance and a data preparation function to use alongside the Guyot algorithm, and to test its performance for reducing errors and improving accuracy.

Methods

Part I: Trace and export survival probabilities

- Identify the KM graph and associated numbers at risk table from the publication
    - Import your file into a digitization software (e.g., Engauge Digitizer Software)
- Trace and export the survival probabilities of each treatment arm individually:
    - When setting axes boundaries, carefully identify the point (0,0); this is not always where the two axis lines intersect;
    - The first point should be chosen to be at time 0 where the survival probability is 100%;
    - Click one time at the ‘bottom’ of each consecutive step in the curve to identify the time at which an event has happened, and the new survival probability at each specific time;
    - To identify the last time a survival probability is recorded:
        - For a horizontal plateau (e.g. chemotherapy plus surgery arm), click at the start and end of the plateau;
        - For a vertical drop (e.g. surgery alone arm), click at the top and bottom of the vertical line.

Figure 1. An Example - Kaplan-Meier Plot of Disease-free Survival in Patients with Stage IIIA Lung Cancer[2]



- Extract the number of patients at risk from the publication into a CSV, as shown in Table 1, which will be used as input for the data preparation.

Table 1. nriskfile.csv

time	number_at_risk
0	30
12	18
24	7

Part II: Data Preparation and Guyot Algorithm

- Prepare the input for the data preparation function, with the following arguments:

  CurveTracingMod(nriskfile, digisurvfile, horizontalend, nsize)

    - *nriskfile* = nriskfile.csv (if not provided, default = NULL)
        - Data frame column A: time (the time number at risk is provided)
        - Data frame column B: number_at_risk (number of patients at risk)
    - *digisurvfile* = digisurvfile.csv
        - Data frame column A: t (time of extracted survival probability)
        - Data frame column B: survival (extracted survival probability from digitizer software)
    - *horizontalend* = TRUE/FALSE
        - Boolean TRUE (if the KM curve is ends in a horizontal plateau)
        - Boolean FALSE (if the KM curve ends in a vertical drop)
    - *nsize* = numeric value;
        - If reported, nsize is the total population size
- Copy and paste the digitizer output into a CSV file, with the column titles “t” and “survival”, as in Table 2, representative of the chemo + surgery arm from Figure 1.

Table 2. digisurvfile.csv

t	survival
-0.5907	1.00302
1.9444	0.95723
2.9405	0.91739
---	---
20.1578	0.32649
23.3264	0.26718
36.0574	0.26638

Part III: Validate data preparation

i. If the number of patients at risk is available in the publication:
    a) The first entry in the number at risk table should be (0, N), where N is the number of patients allocated to the trial arm (if provided).
    b) The last entry of the number at risk table should be (0, 0); at the end of the survival curve there should be 0 patients left, at this time all patients have had either an event or censoring event.

ii. If the number of patients at risk is **not** available:
    a) The first entry in the number at risk table should be (0, *nsize*)
    b) No other entries should exist in this case

iii. The survival probability should start at (0, 1); if the first time entry of the digisurvfile rounds to 0, then the recorded survival probability is not changed.

iv. If the curve ends in a **horizontal** line, the last two time points in the digisurvfile should have identical survival probabilities (survival). If the curve ends in a **vertical** line, the last two survival probabilities should have the same recorded time (t).

v. At each time point the number of patients at risk is provided, there should be information about the survival probability
    a) For example, if the number of patients at risk is recorded at time 4 and time 6, there should be a survival probability at time 4 and time 6.

vi. The survival probabilities should be monotonically decreasing; any survival times that violate this rule will be dropped.

Part IV: Validate and compare the pseudo-IPD

- For each curve:
    - Compute the Kaplan-Meier (KM) survival estimate
    - Generate a plot of the KM curve and overlay on top of the publication
- To compare two curves:
    - Compute the Cox proportional hazards regression model to estimate the log-hazard ratios and the p-value
    - Perform a test of independence between the residuals and time in a Cox proportional hazards model by analyzing the scaled Schoenfeld residuals
    - Graph the two treatment arms on the same axes as (a) survival probability versus time or (b) log(-log(survival)) versus log(time).

Results

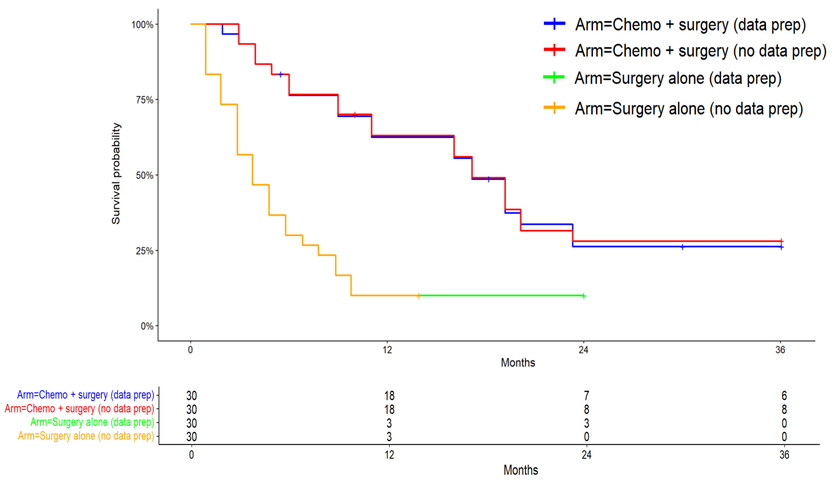
- The *CurveTracingMod()* function prepares the data for the Guyot Algorithm, for example by ensuring there are no negative times extracted from the digitization process. Table 3 shows the results of the data processing step.

Table 3. digisurvfile.csv after data processing

t	survival
0.000	1.0000
1.9444	0.95723
2.9405	0.91739
---	---
23.3264	0.26718
24.0000	0.26718
36.9574	0.26718

- Implementation of the data processing step improves the accuracy of the distribution of the event and censoring times. The number at risk table of the curves with data preparation is slightly closer to the numbers reported in the publication.
- A simple calculation of the hazard ratio (HR) and Schoenfeld residual values highlight a key difference in the results between using the Guyot algorithm with and without the data processing step.
    - The p-value of the HR **without** pre-processing is 0.2279, not significant at the 0.05 level, while the p-value of the HR **with** pre-processing is 0.004, which is significant.
    - The p-value generated by the pre-processing data aligns with the significant p-value presented in the publication.

Figure 2. KM estimates with and without data preparation



Conclusions

- The efforts of this study reduced errors and improved standardization when using Guyot to generate pseudo IPD
- The results of the pseudo-IPD align closely with the results of the publication and indicate a similar level of significance
- Using this function can improve validity of indirect treatment comparisons where pseudo-IPD is required

References

[1] Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012 Feb 1;12:9. doi: 10.1186/1471-2288-12-9. PMID: 22297116; PMCID: PMC3313891.

[2] Rosell R, Gómez-Codina J, Camps C, Maestre J, Padille J, Cantó A, Mate JL, Li S, Roig J, Olazábal A, et al. A randomized trial comparing preoperative chemotherapy plus surgery with surgery alone in patients with non-small-cell lung cancer. N Engl J Med. 1994 Jan 20;330(3):153-8. doi: 10.1056/NEJM199401203300301. PMID: 8043059.

Abbreviations

IPD: individual patient data; PH: proportional hazard; KM: Kaplan-Meier

Presented at ISPOR EU | Copenhagen, Denmark | November 12–15, 2023

Contact information: www.cytel.com