PharmaSUG 2024 - Paper RW-125

Reconstruction of Individual Patient Data (IPD) from Published Kaplan-Meier Curves Using Guyot's Algorithm: Step-by-Step Programming in R

Ajay Gupta, Daiichi Sankyo
Natalie Dennis, Daiichi Sankyo

ABSTRACT

Secondary analysis may require the use of reconstructed patient-level data from published Kaplan-Meier (KM) curves to support a number of different objectives, including indirect treatment comparisons within the context of economic evaluations. Guyot (2012) developed an algorithm that reconstructs individual patient data (IPD) for time-to-event endpoints using published KM curves. This presentation will provide step-by-step instructions and a use case for executing the Guyot (2012) algorithm to reconstruct IPD from published KM curves in R.

INTRODUCTION

R is a programming language for statistical computing and data visualization. It has been adopted in the fields of data mining, bioinformatics, and data analysis. The core R language is augmented by many extension packages, containing reusable code, documentation, and sample data. R software is open-source and free software. It is licensed by the GNU Project and available under the GNU General Public License. It is written primarily in C, Fortran, and R itself. Precompiled executables are provided for various operating systems. As an interpreted language, R has a native command line interface. Moreover, multiple third-party graphical user interfaces are available, such as RStudio-an integrated development environment.

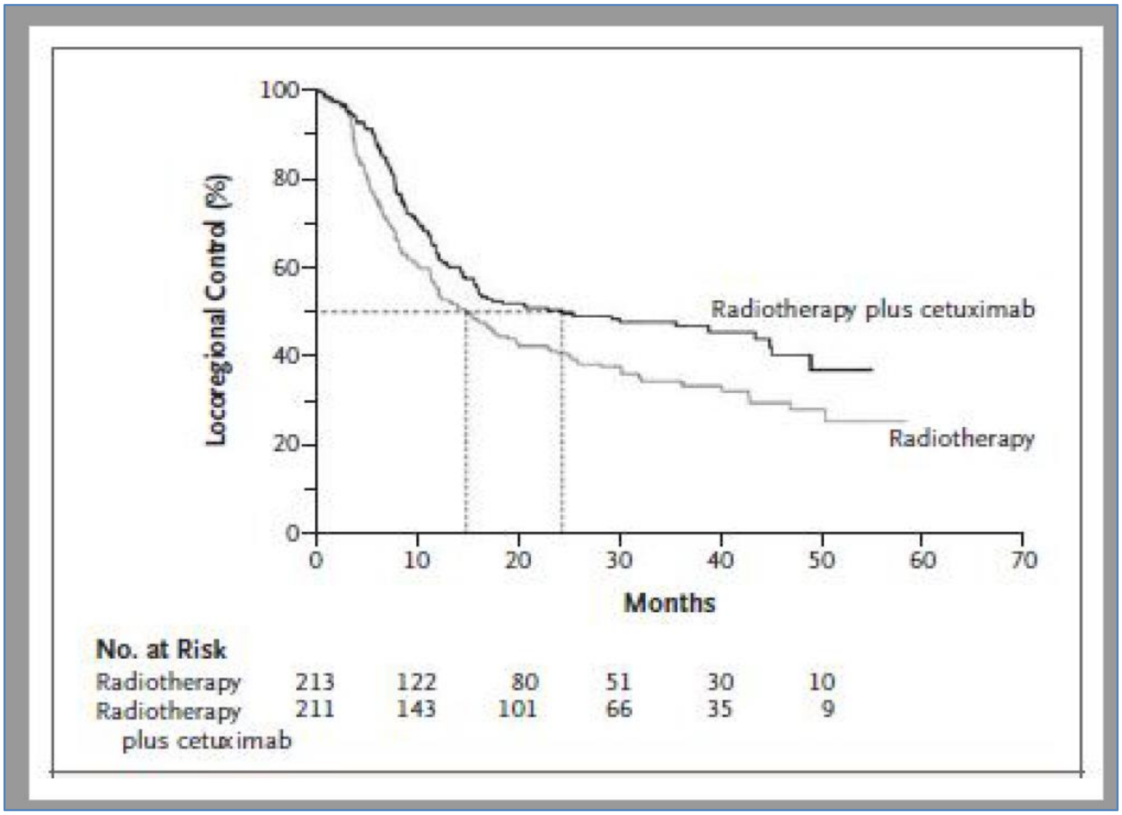
Secondary analysis may require the use of reconstructed patient-level data from published Kaplan-Meier (KM) curves to support a number of different objectives, including indirect treatment comparisons within the context of economic evaluations. Guyot (2012) developed an algorithm that reconstructs individual patient data (IPD) for time-to-event endpoints using published KM curves. This presentation will provide step-by-step instructions and a use case for executing the Guyot (2012) algorithm to reconstruct IPD from published KM curves in R.

STEPS TO RECONSTRUCT INDIVIDUAL PATIENT DATA (IPD) FROM PUBLISHED KM CURVE

1. Digitize Kaplan-Meier curves using published graph (using PlotDigitizer, GetData Graph Digitizer or other application)
2. Save the extracted survival data (digitized x- and y-coordinates) as a CSV/Excel file.
3. Create a file for the number of patients at risk, including the time points and the lower and upper intervals (based on the digitized Kaplan-Meier curves)
4. Identify the total number of events (if published)
5. Run Guyot’s algorithm using R by importing extracted survival data and number of patients at risk files.

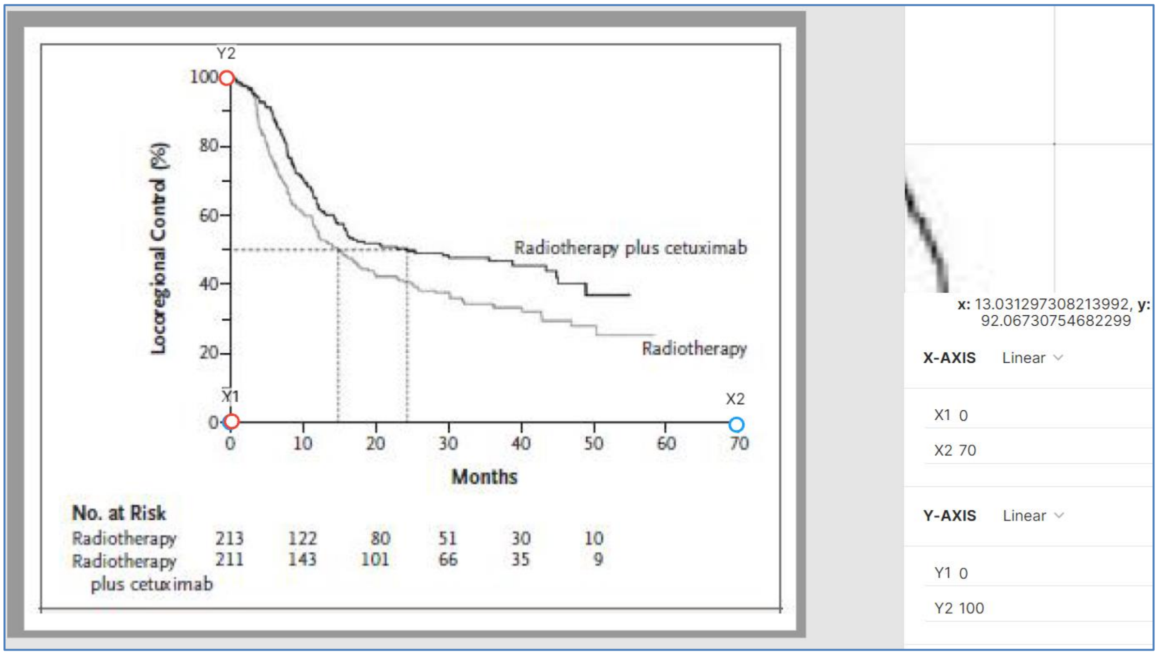
DIGITIZE KAPLAN-MEIER CURVES USING PUBLISHED GRAPH

1. Create image e.g., PNG, GIF from published KM Curve. See, example from published paper (Guyot 2012).



Display 1. Published KM Curve (Guyot 2012)

2. Import the file in plot digitizer user interface and select the range for X and Y axis.


Y2
X1
X2
Locoregional Control (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Months
Radiotherapy plus cetuximab
Radiotherapy
No. at Risk
Radiotherapy 213 122 80 51 30 10
Radiotherapy plus cetuximab 211 143 101 66 35 9

x: 13.031297308213992, **y:** 92.06730754682299

X-AXIS Linear

X1 0

X2 70

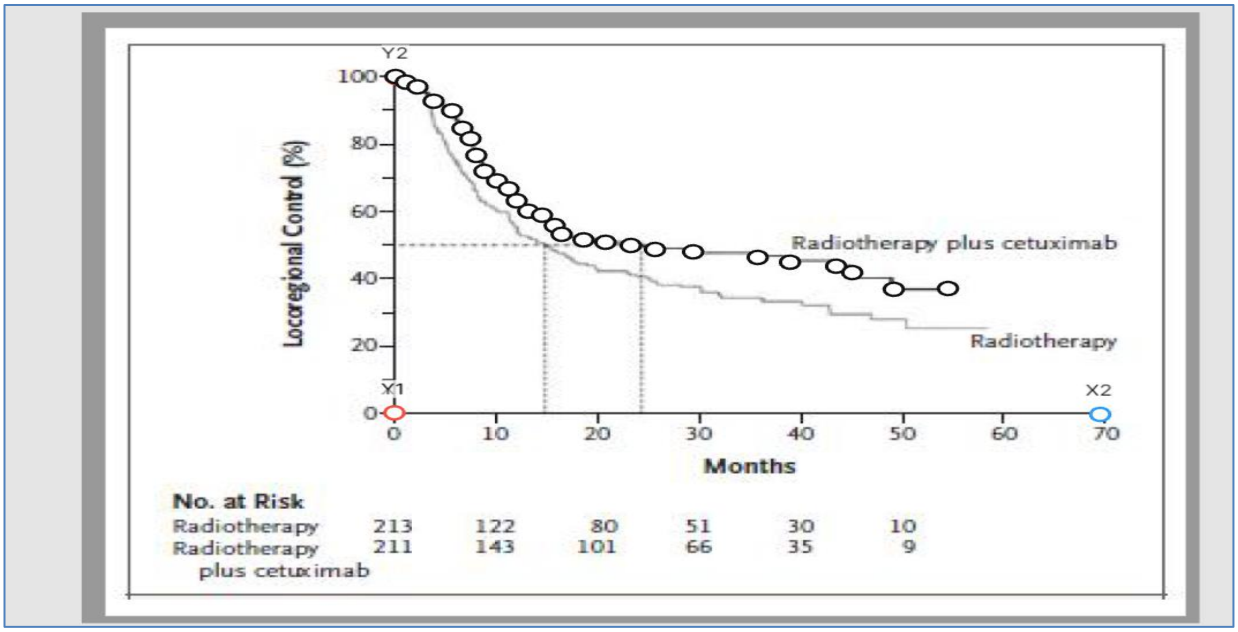
Y-AXIS Linear

Y1 0

Y2 100

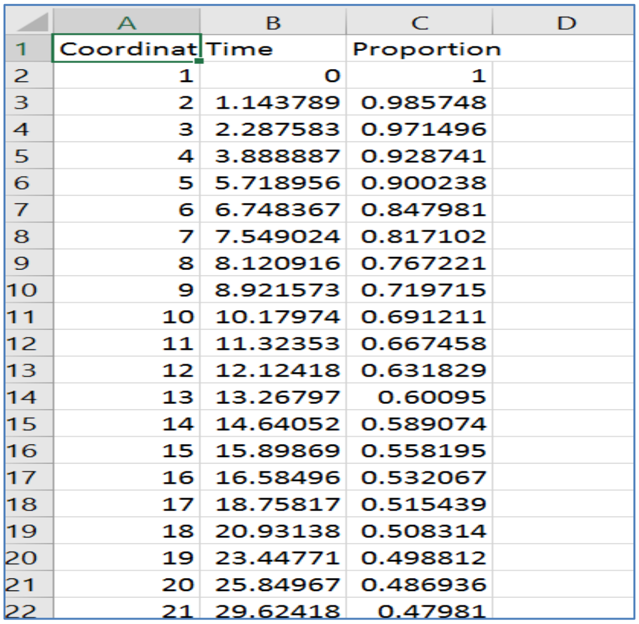
Display 2. Published KM Curve (Guyot 2012) in Plot Digitizer

3. Mark the event times for the Kaplan-Meier curve to be digitized, i.e., at every point in which there is a step down in the Kaplan-Meier curve.



Display 3. Published KM Curve (Guyot 2012) with x and y manual markup.

4. After marking the relevant coordinates of the Kaplan-Meier curve, export the data into .csv file. Make sure to divide value on Y axis from 100 to get the proportion or mark the Y axis goes from 0 to 1 (even it goes to 100).

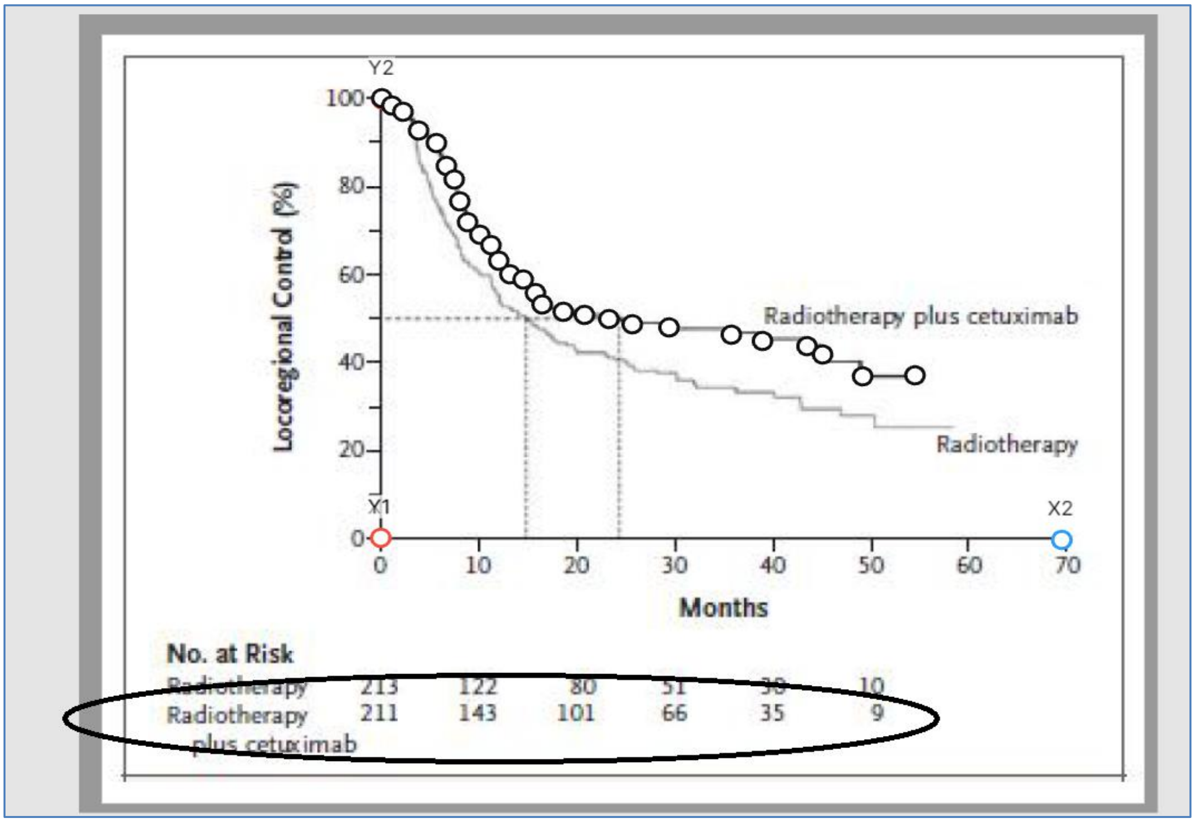


	A	B	C	D
1	Coordinat	Time	Proportion	
2	1	0	1	
3	2	1.143789	0.985748	
4	3	2.287583	0.971496	
5	4	3.888887	0.928741	
6	5	5.718956	0.900238	
7	6	6.748367	0.847981	
8	7	7.549024	0.817102	
9	8	8.120916	0.767221	
10	9	8.921573	0.719715	
11	10	10.17974	0.691211	
12	11	11.32353	0.667458	
13	12	12.12418	0.631829	
14	13	13.26797	0.60095	
15	14	14.64052	0.589074	
16	15	15.89869	0.558195	
17	16	16.58496	0.532067	
18	17	18.75817	0.515439	
19	18	20.93138	0.508314	
20	19	23.44771	0.498812	
21	20	25.84967	0.486936	
22	21	29.62418	0.47981	

Display 4. CSV file from plot digitizer.

CREATE A FILE FOR THE NUMBER OF PATIENTS AT RISK:

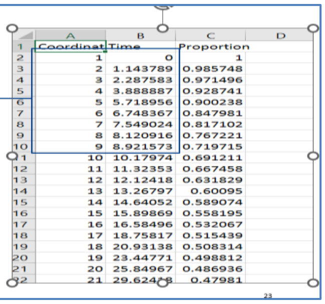
See below graph and follow the instructions below to create the number of patients at risk file.



Display 5. Published KM Curve (Guyot 2012) showing number of patients at risk.

- Create a file for the number of patients at risk, including the time points and the lower and upper intervals.
- Nrisk: value provided in graph above.
- Trisk: time value corresponding to each Nrisk (every 10 months in this example)
- Lower and Upper: the coordinates in the digitize file corresponding to each time window (e.g., coordinates 1-9 fall between 0 and 10 months).

nrisk	trisk	lower	upper
211	0	1	9
143	10	10	17
101	20	18	21
66	30	22	23
35	40	24	26
9	50	27	27

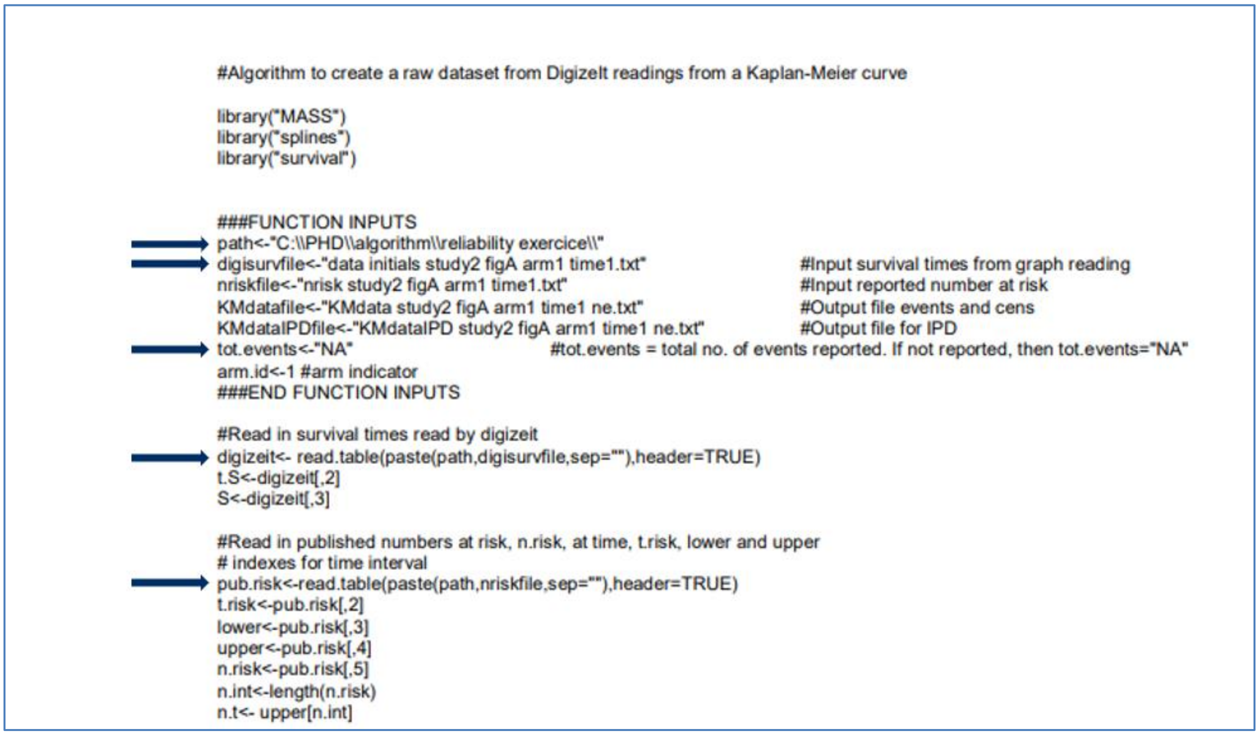


	A	B	C	D
1	Coordinat	Time	Proportion	
2	1	0	1	
3	2	1.143789	0.985748	
4	3	2.287583	0.971496	
5	4	3.888887	0.928741	
6	5	5.718956	0.900238	
7	6	6.748367	0.847981	
8	7	7.549024	0.817102	
9	8	8.120916	0.767221	
10	9	8.921573	0.719715	
11	10	10.17974	0.691211	
12	11	11.32353	0.667458	
13	12	12.12418	0.631829	
14	13	13.26797	0.60095	
15	14	14.64052	0.589074	
16	15	15.89869	0.558195	
17	16	16.58496	0.532067	
18	17	18.75817	0.515439	
19	18	20.93138	0.508314	
20	19	23.44771	0.498812	
21	20	25.84967	0.486936	
22	21	29.624[illegible]8	0.47981	

Display 6. Number of patients at risk file

RUN GUYOT’S ALGORITHM USING R:

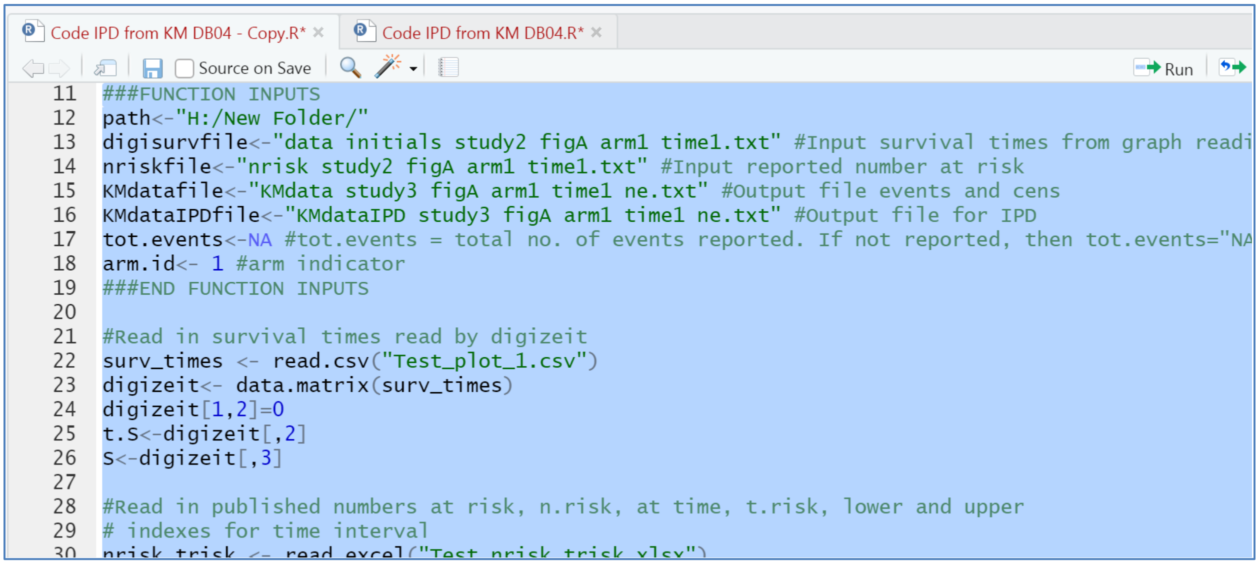
- Download R program containing Guyot’s algorithm from following location. 12874_2011_700_MOESM1_ESM.PDF (springer.com) and update the R programs with respective values (shown in figure below with arrow)



```
#Algorithm to create a raw dataset from DigizeIt readings from a Kaplan-Meier curve

library("MASS")
library("splines")
library("survival")


###FUNCTION INPUTS
path<-"C:\\PHD\\algorithm\\reliability exercice\\"
digisurvfile<-"data initials study2 figA arm1 time1.txt"          #Input survival times from graph reading
nriskfile<-"nrisk study2 figA arm1 time1.txt"                     #Input reported number at risk
KMdatafile<-"KMdata study2 figA arm1 time1 ne.txt"                #Output file events and cens
KMdataIPDfile<-"KMdataIPD study2 figA arm1 time1 ne.txt"          #Output file for IPD
tot.events<-"NA"                     #tot.events = total no. of events reported. If not reported, then tot.events="NA"
arm.id<-1 #arm indicator
###END FUNCTION INPUTS

#Read in survival times read by digizeit
digizeit<- read.table(paste(path,digisurvfile,sep=""),header=TRUE)
t.S<-digizeit[,2]
S<-digizeit[,3]

#Read in published numbers at risk, n.risk, at time, t.risk, lower and upper
# indexes for time interval
pub.risk<-read.table(paste(path,nriskfile,sep=""),header=TRUE)
t.risk<-pub.risk[,2]
lower<-pub.risk[,3]
upper<-pub.risk[,4]
n.risk<-pub.risk[,5]
n.int<-length(n.risk)
n.t<- upper[n.int]
```

Display 7. R program from the Guyot (2012) publication

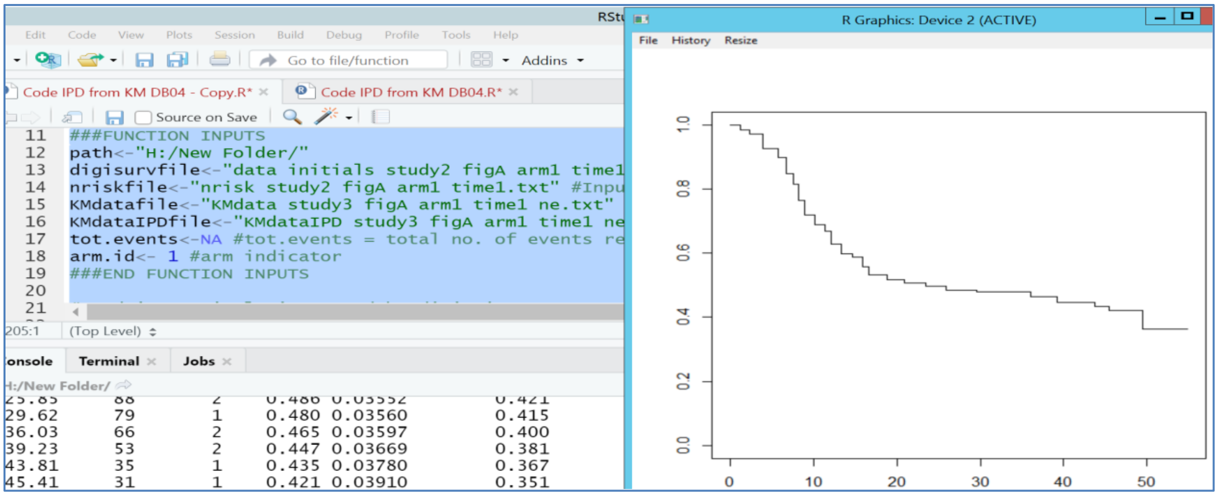
- Execute the R code in R studio:



```
###FUNCTION INPUTS
path<-"H:/New Folder/"
digisurvfile<-"data initials study2 figA arm1 time1.txt" #Input survival times from graph readi
nriskfile<-"nrisk study2 figA arm1 time1.txt" #Input reported number at risk
KMdatafile<-"KMdata study3 figA arm1 time1 ne.txt" #Output file events and cens
KMdataIPDfile<-"KMdataIPD study3 figA arm1 time1 ne.txt" #Output file for IPD
tot.events<-NA #tot.events = total no. of events reported. If not reported, then tot.events="NA
arm.id<- 1 #arm indicator
###END FUNCTION INPUTS

#Read in survival times read by digizeit
surv_times <- read.csv("Test_plot_1.csv")
digizeit<- data.matrix(surv_times)
digizeit[1,2]=0
t.S<-digizeit[,2]
S<-digizeit[,3]

#Read in published numbers at risk, n.risk, at time, t.risk, lower and upper
# indexes for time interval
nrisk_trisk <- read_excel("Test_nrisk_trisk.xlsx")
```

Display 8. R program in R studio

- After execution of the programs, to ensure similarity and that the function accurately created pseudo-IPD based on the digitized Kaplan-Meier curve it is possible to recreate the KM curve to compare to the published one.

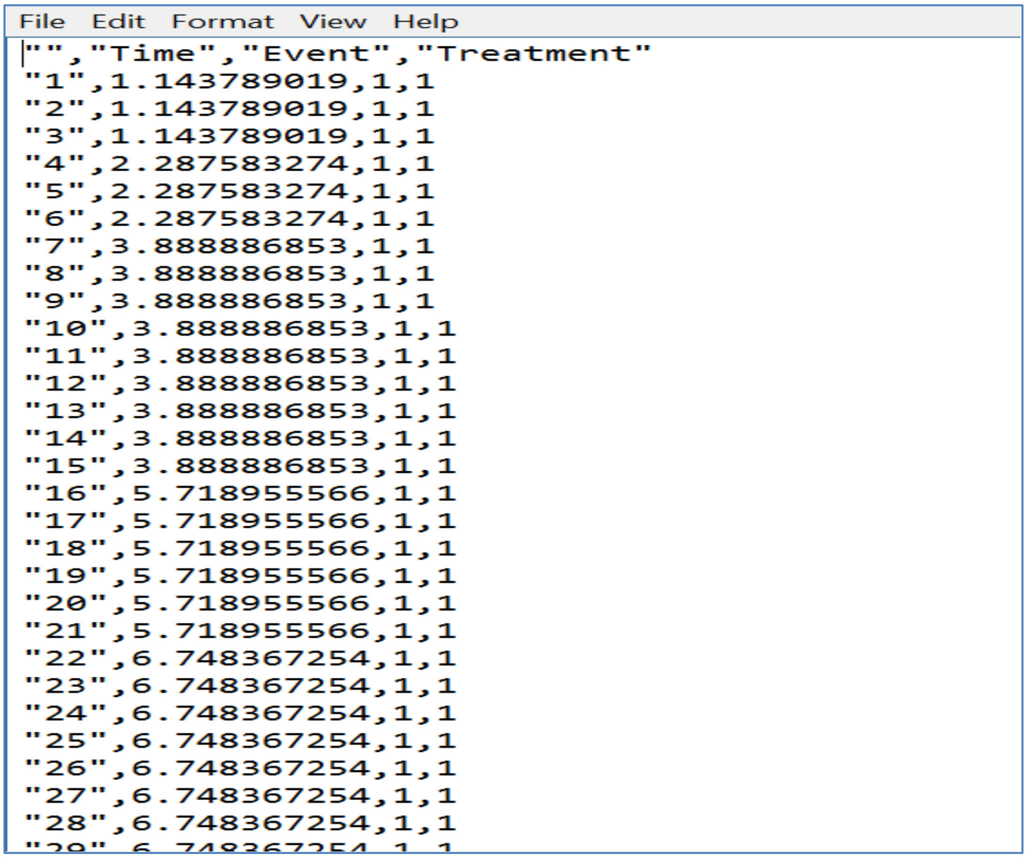


```
###FUNCTION INPUTS
path<-"H:/New Folder/"
digisurvfile<-"data initials study2 figA arm1 time1
nriskfile<-"nrisk study2 figA arm1 time1.txt" #Inpu
KMdatafile<-"KMdata study3 figA arm1 time1 ne.txt"
KMdataIPDfile<-"KMdataIPD study3 figA arm1 time1 ne
tot.events<-NA #tot.events = total no. of events re
arm.id<- 1 #arm indicator
###END FUNCTION INPUTS
```

```
25.85     88      2    0.486 0.03552         0.421
29.62     79      1    0.480 0.03560         0.415
36.03     66      2    0.465 0.03597         0.400
39.23     53      2    0.447 0.03669         0.381
43.81     35      1    0.435 0.03780         0.367
45.41     31      1    0.421 0.03910         0.351
```

Display 9. Recreating the KM curve in R studio

- The programs also create a .txt file with pseudo-patient level data that can be use in secondary analysis.

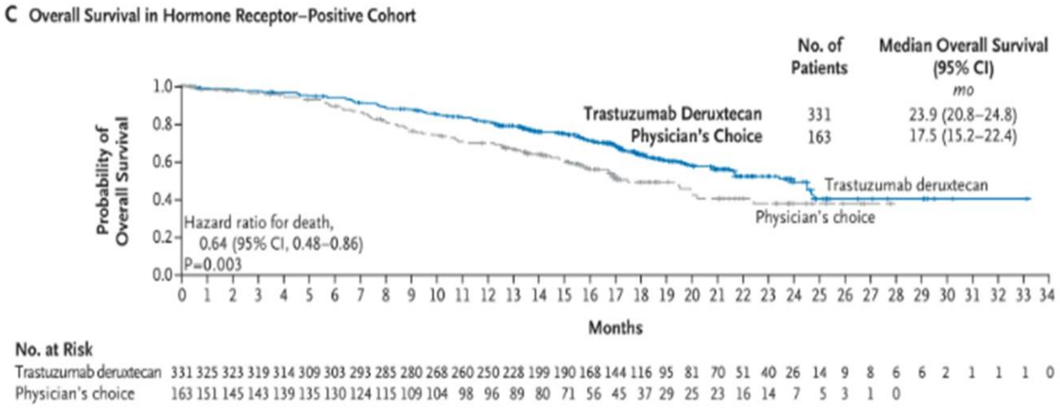


```
"","Time","Event","Treatment"
"1",1.143789019,1,1
"2",1.143789019,1,1
"3",1.143789019,1,1
"4",2.287583274,1,1
"5",2.287583274,1,1
"6",2.287583274,1,1
"7",3.888886853,1,1
"8",3.888886853,1,1
"9",3.888886853,1,1
"10",3.888886853,1,1
"11",3.888886853,1,1
"12",3.888886853,1,1
"13",3.888886853,1,1
"14",3.888886853,1,1
"15",3.888886853,1,1
"16",5.718955566,1,1
"17",5.718955566,1,1
"18",5.718955566,1,1
"19",5.718955566,1,1
"20",5.718955566,1,1
"21",5.718955566,1,1
"22",6.748367254,1,1
"23",6.748367254,1,1
"24",6.748367254,1,1
"25",6.748367254,1,1
"26",6.748367254,1,1
"27",6.748367254,1,1
"28",6.748367254,1,1
"29",6.748367254,1,1
```

Display 10. Individual Patient level data

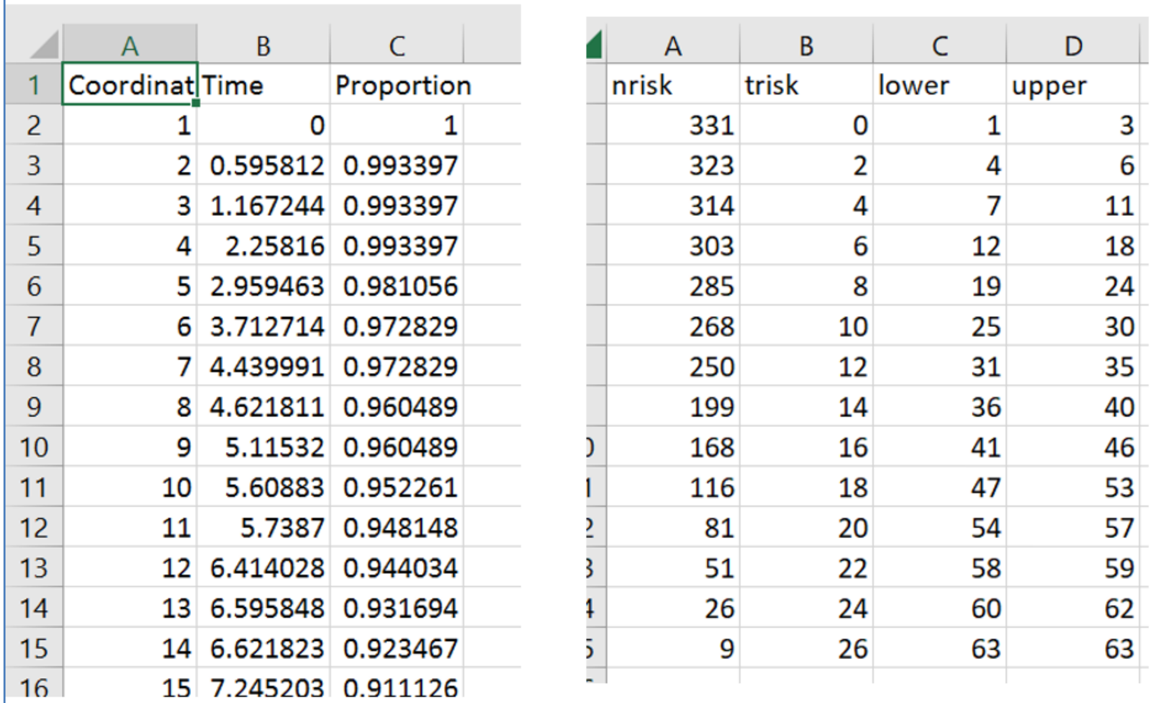
MORE EXAMPLES

See below few more examples. Follow the steps above to recreate the IPD from KM curve.



Reference: Modi, Shanu, et al. "Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer." New England Journal of Medicine 387.1 (2022): 9-20.

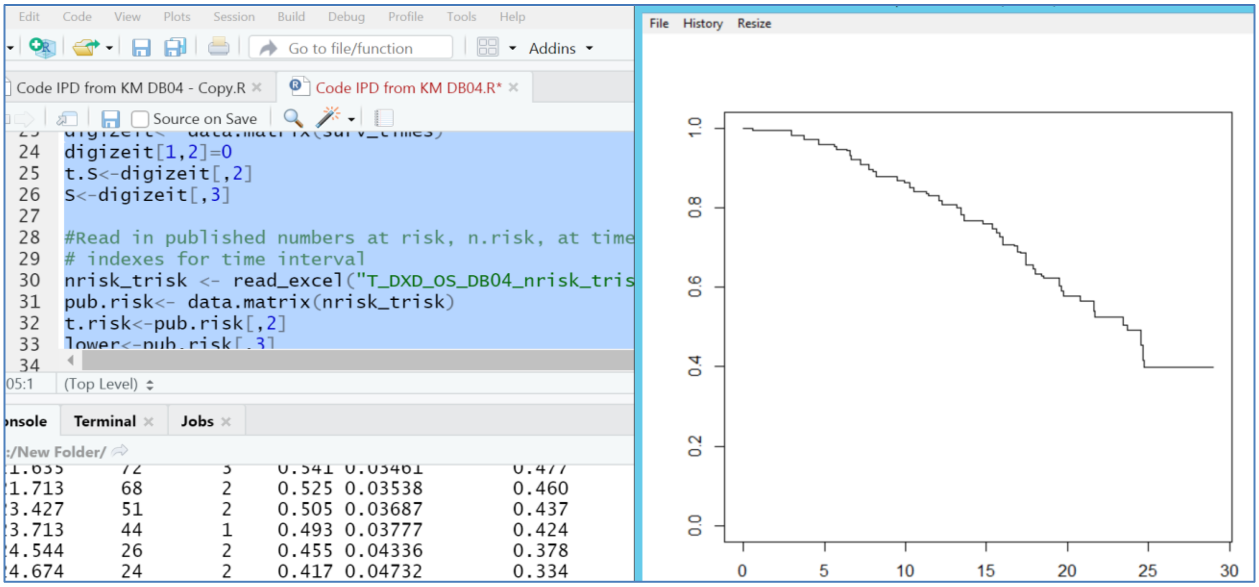
Display 11. Published KM curve.



	A	B	C
1	Coordinat	Time	Proportion
2	1	0	1
3	2	0.595812	0.993397
4	3	1.167244	0.993397
5	4	2.25816	0.993397
6	5	2.959463	0.981056
7	6	3.712714	0.972829
8	7	4.439991	0.972829
9	8	4.621811	0.960489
10	9	5.11532	0.960489
11	10	5.60883	0.952261
12	11	5.7387	0.948148
13	12	6.414028	0.944034
14	13	6.595848	0.931694
15	14	6.621823	0.923467
16	15	7.245203	0.911126

A	B	C	D
nrisk	trisk	lower	upper
331	0	1	3
323	2	4	6
314	4	7	11
303	6	12	18
285	8	19	24
268	10	25	30
250	12	31	35
199	14	36	40
168	16	41	46
116	18	47	53
81	20	54	57
51	22	58	59
26	24	60	62
9	26	63	63

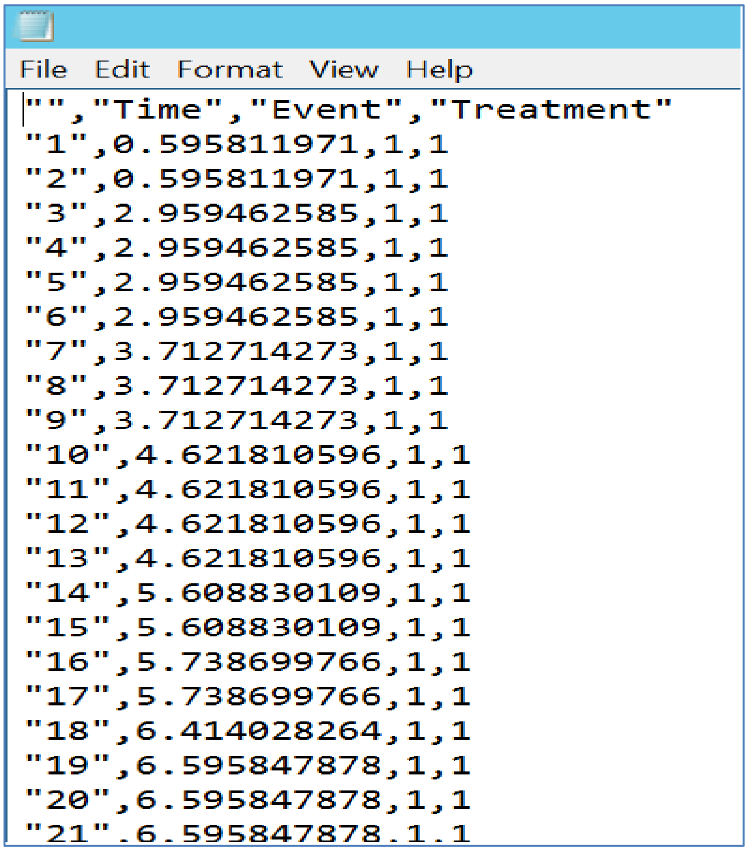
Display 12. Markup file and Risk file for above KM curve



```
digizeit<- data.matrix(surv_times)
digizeit[1,2]=0
t.S<-digizeit[,2]
S<-digizeit[,3]

#Read in published numbers at risk, n.risk, at time
# indexes for time interval
nrisk_trisk <- read_excel("T_DXD_OS_DB04_nrisk_tris
pub.risk<- data.matrix(nrisk_trisk)
t.risk<-pub.risk[,2]
lower<-pub.risk[,3]
```

```
1.635    72     3    0.541 0.03461      0.477
1.713    68     2    0.525 0.03538      0.460
3.427    51     2    0.505 0.03687      0.437
3.713    44     1    0.493 0.03777      0.424
4.544    26     2    0.455 0.04336      0.378
4.674    24     2    0.417 0.04732      0.334
```

Display 13. Recreate KM curve using Guyot (2012) algorithm.



```
"","Time","Event","Treatment"
"1",0.595811971,1,1
"2",0.595811971,1,1
"3",2.959462585,1,1
"4",2.959462585,1,1
"5",2.959462585,1,1
"6",2.959462585,1,1
"7",3.712714273,1,1
"8",3.712714273,1,1
"9",3.712714273,1,1
"10",4.621810596,1,1
"11",4.621810596,1,1
"12",4.621810596,1,1
"13",4.621810596,1,1
"14",5.608830109,1,1
"15",5.608830109,1,1
"16",5.738699766,1,1
"17",5.738699766,1,1
"18",6.414028264,1,1
"19",6.595847878,1,1
"20",6.595847878,1,1
"21",6.595847878,1,1
```

Display 14. Individual Patient level data

ALTERNATIVE APPROACH

- In 2021, Na Liu, Yanhong Zhou & J. Jack Lee proposed a modified, more flexible version of Guyot's algorithm to reconstruct IPD from published K-M curves and developed a R package and Shiny application. See below publication link for more detail. More details will be provided in future presentation.
  - IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves | BMC Medical Research Methodology | Full Text (biomedcentral.com)
- The alternative approach does not require a manual approach to digitizing curves using external software and relaxes some of the requirements for data input. It is therefore an all-in-one approach that reconstructs IPD directly from the KM curve.

CONCLUSION

Using the digitization software and Guyot (2012) algorithm we can efficiently reconstructs individual patient data (IPD) for time-to-event endpoints using published KM curves. This data can be especially useful in secondary analysis to support a number of different objectives, including indirect treatment comparisons within the context of economic evaluations.

REFERENCES

R: The R Project for Statistical Computing (r-project.org)

Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves | BMC Medical Research Methodology | Full Text (biomedcentral.com)

PNS210 A Comparison of Graph Digitization Software for the Reconstruction of Published Kaplan Meier Curves - Value in Health (valueinhealthjournal.com)

PlotDigitizer Online App

IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves | BMC Medical Research Methodology | Full Text (biomedcentral.com)

R (programming language) - Wikipedia

ACKNOWLEDGMENTS

Thanks to Antonio Lovatin, Véronique ROBERT and Giacomo Mordenti for their reviews and comments. Thanks to our family for their support.

CONTACT INFORMATION

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

Ajay Gupta, M.S.
Daiichi Sankyo, Inc
211 Mt Airy Rd
Basking Ridge, NJ 07920
Phone : (269) -873-1145
E-mail : Ajgupta@dsi.com,
ajaykailasgupta@gmail.com

Natalie Dennis
Daiichi Sankyo, Inc
France
Phone : +33 6 77 50 02 77
E-mail : natalie.dennis@daiichi-sankyo.eu

DISCLAIMER

The content of this paper are the works of the authors and do not necessarily represent the opinions, recommendations, or practices of Daiichi Sankyo.

Any brand and product names are trademarks of their respective companies.