Passion for Innovation.
Compassion for Patients.™

Daiichi-Sankyo

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

Ajay Gupta and Natalie Dennis

INTRODUCTION

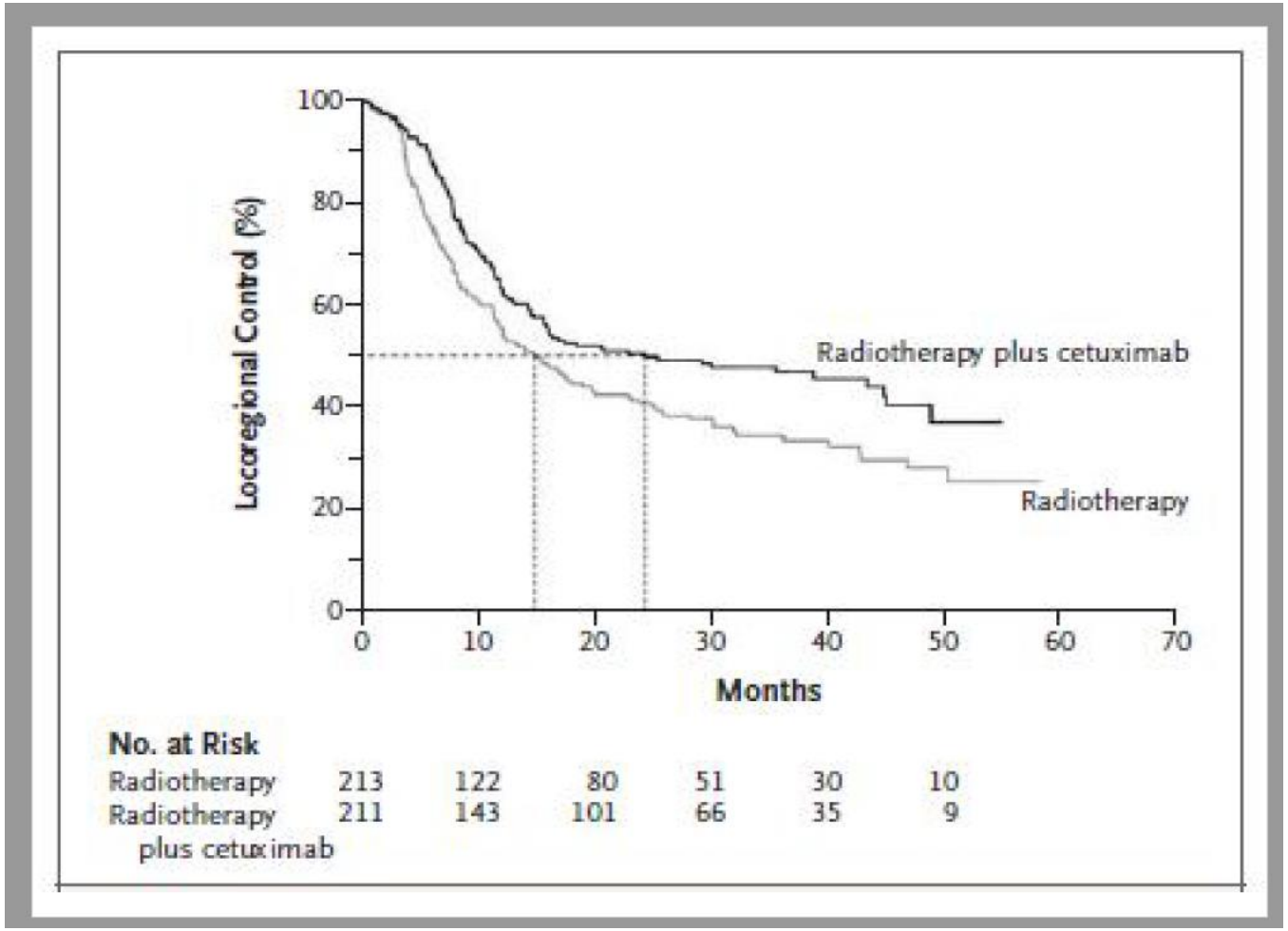
- Secondary analysis may require the use of reconstructed patient-level data from published Kaplan-Meier (KM) curves to support a number of different objectives, including indirect treatment comparisons within the context of economic evaluations.
- Guyot (2012) developed an algorithm that reconstructs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

STEPS TO RECONSTRUCTING IPD

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

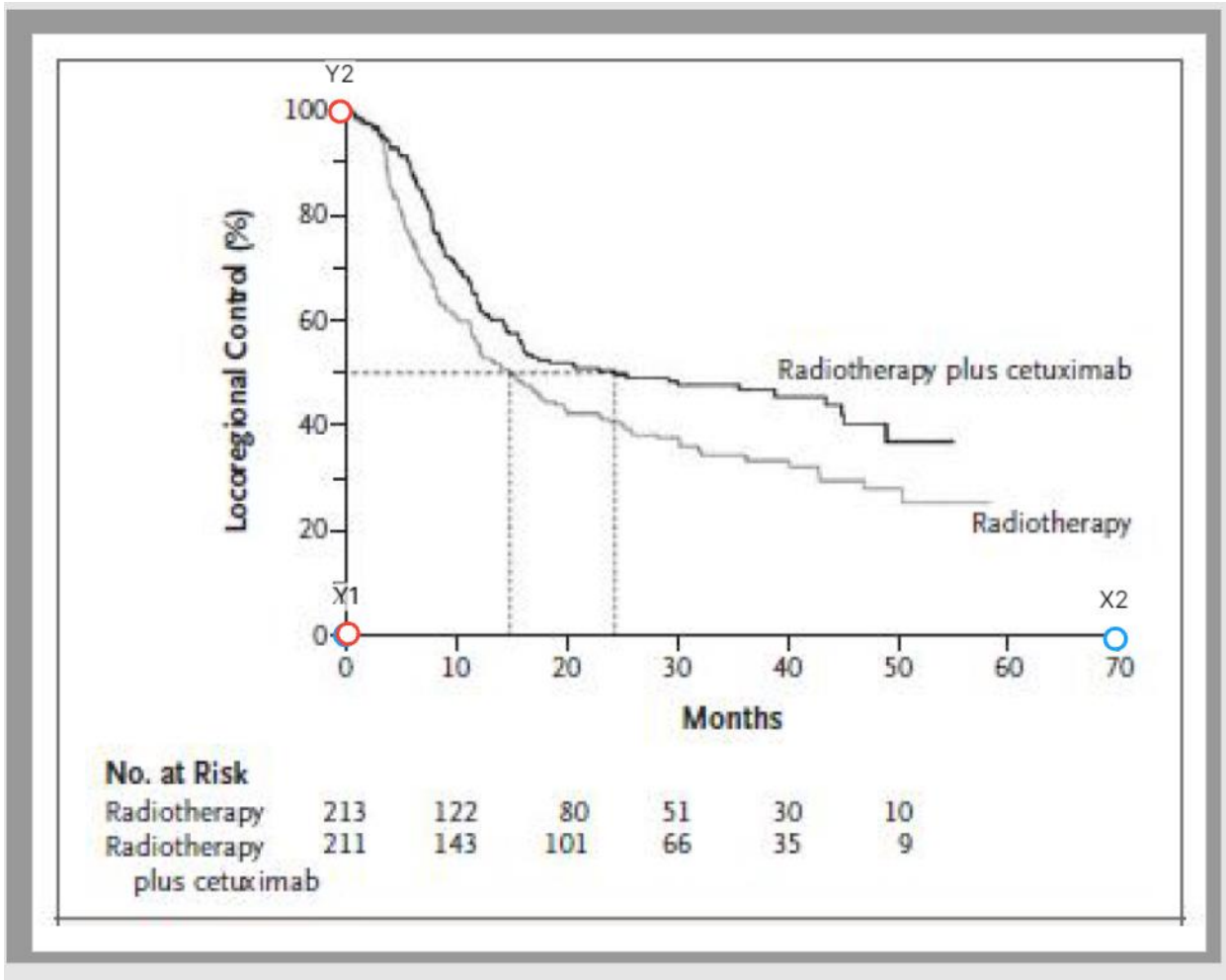
DIGITIZE KAPLAN-MEIER CURVES USING PUBLISHED GRAPH

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



Cont'd.

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



x: 13.031297308213992, y: 92.06730754682299

X-AXIS Linear

X1 0

X2 70

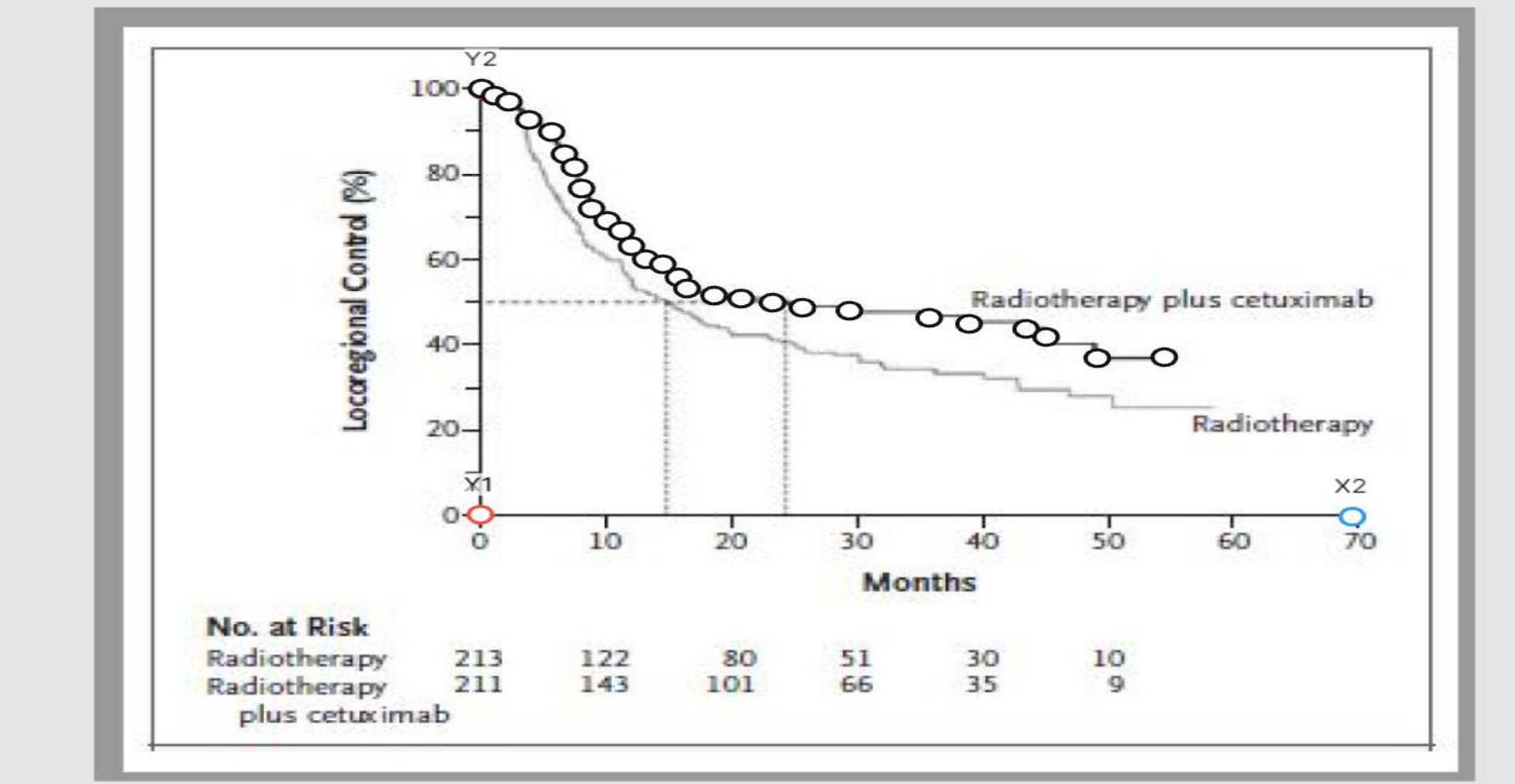
Y-AXIS Linear

Y1 0

Y2 100

Cont'd.

3. Mark the data point manually on the graph when there is change in plot (i.e., at each step down).



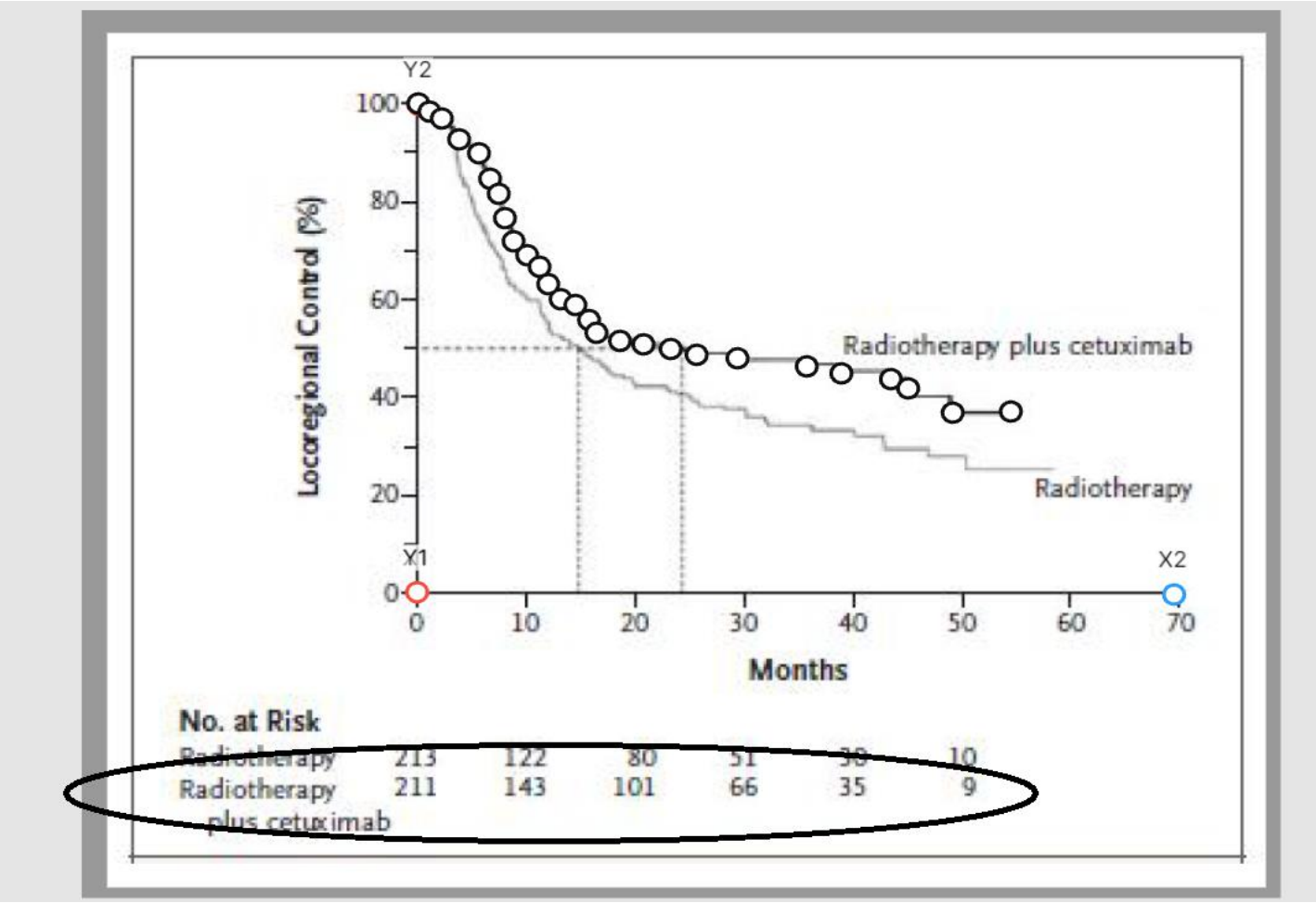
Cont'd.

4. After marking, 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	

CREATE A FILE FOR THE NUMBER OF PATIENTS AT RISK

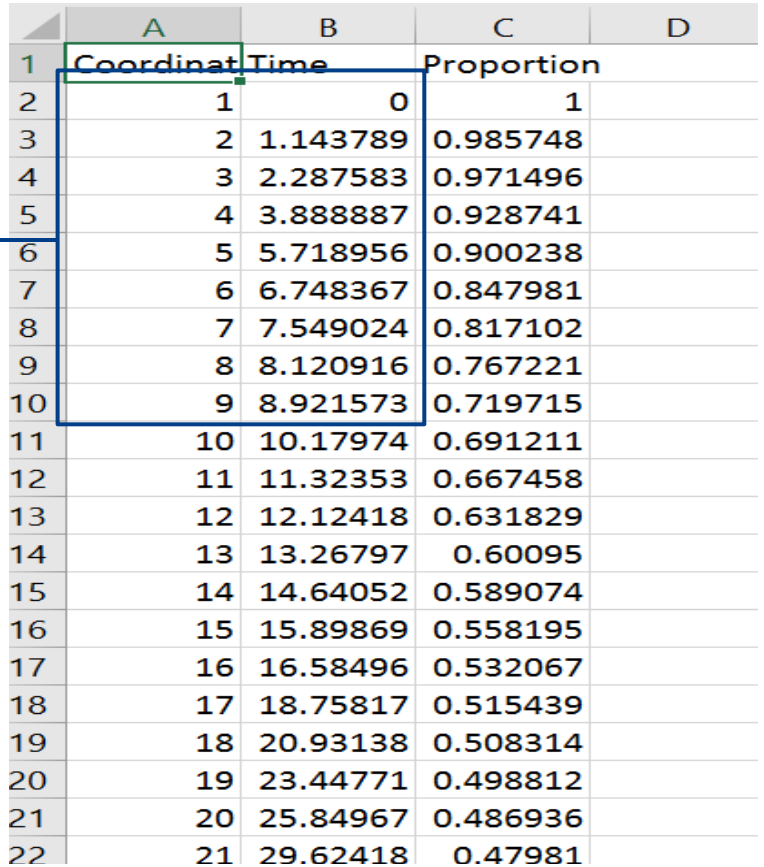
- Create a file for the number of patients at risk, including the time points and the lower and upper intervals.



Cont'd.

- Nrisk: value provided in graph
- 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).

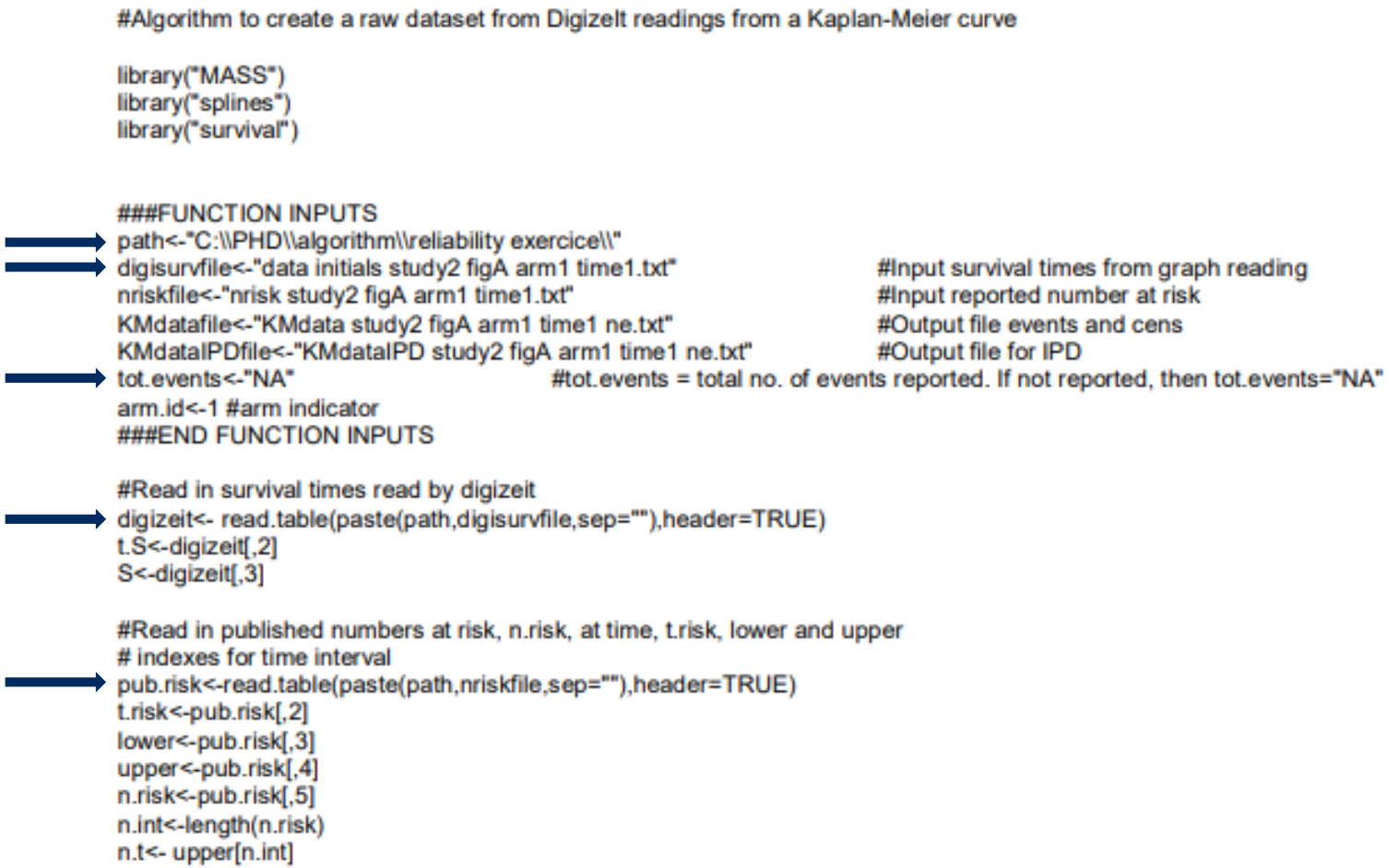
nrisk	trisk	lower	upper
211	0	1	9
143	10	10	17
101	20	18	21
66	30	22	23
35	40	24	26
9	50	27	27



	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	29.62418	0.47981	

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 (show 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]
```

Cont'd.

- Execute the R code in R studio:

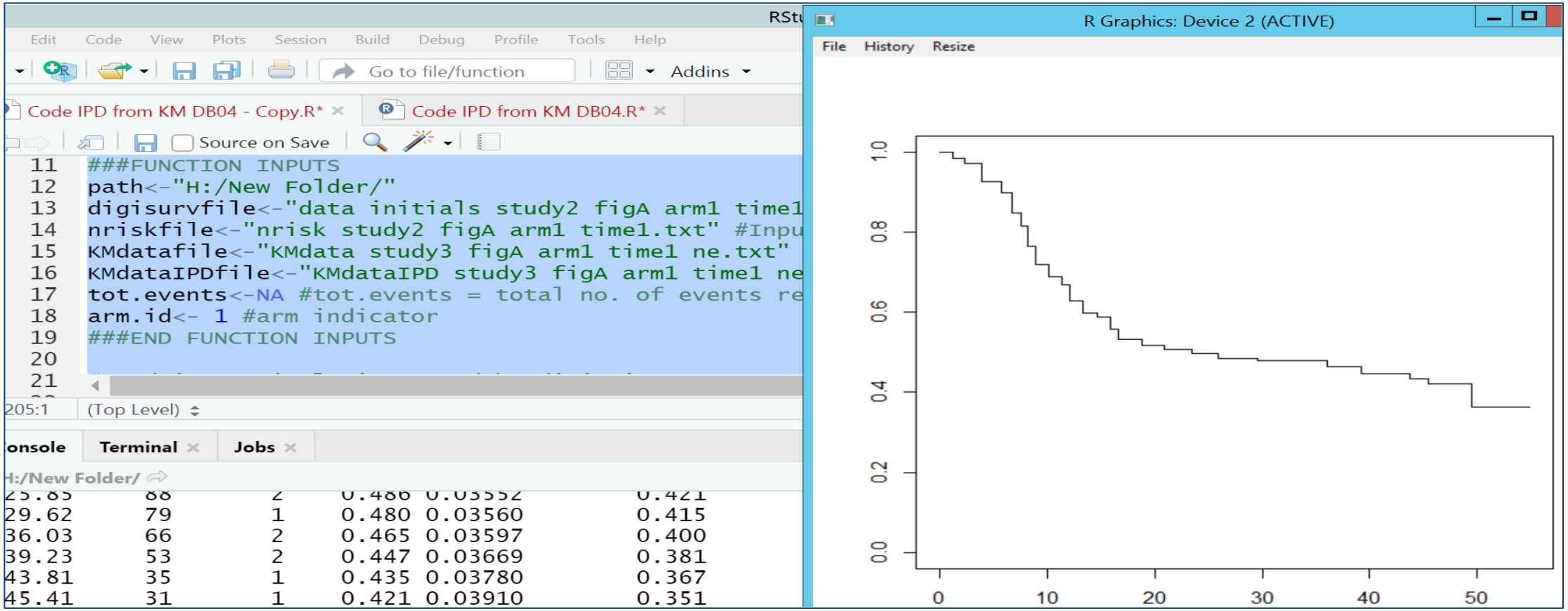
Code IPD from KM DB04 - Copy.R* | Code IPD from KM DB04.R*

Source on Save | Run

```
11  ###FUNCTION INPUTS
12  path<-"H:/New Folder/"
13  digisurvfile<-"data initials study2 figA arm1 time1.txt" #Input survival times from graph readi
14  nriskfile<-"nrisk study2 figA arm1 time1.txt" #Input reported number at risk
15  KMdatafile<-"KMdata study3 figA arm1 time1 ne.txt" #Output file events and cens
16  KMdataIPDfile<-"KMdataIPD study3 figA arm1 time1 ne.txt" #Output file for IPD
17  tot.events<-NA #tot.events = total no. of events reported. If not reported, then tot.events="NA
18  arm.id<- 1 #arm indicator
19  ###END FUNCTION INPUTS
20
21  #Read in survival times read by digizeit
22  surv_times <- read.csv("Test_plot_1.csv")
23  digizeit<- data.matrix(surv_times)
24  digizeit[1,2]=0
25  t.S<-digizeit[,2]
26  S<-digizeit[,3]
27
28  #Read in published numbers at risk, n.risk, at time, t.risk, lower and upper
29  # indexes for time interval
30  nrisk.trisk <- read.excel("Test_nrisk_trisk.xlsx")
```

Cont'd.

- After execution of the programs, it is possible to recreate the KM curve to compare to the published one.



Cont'd.

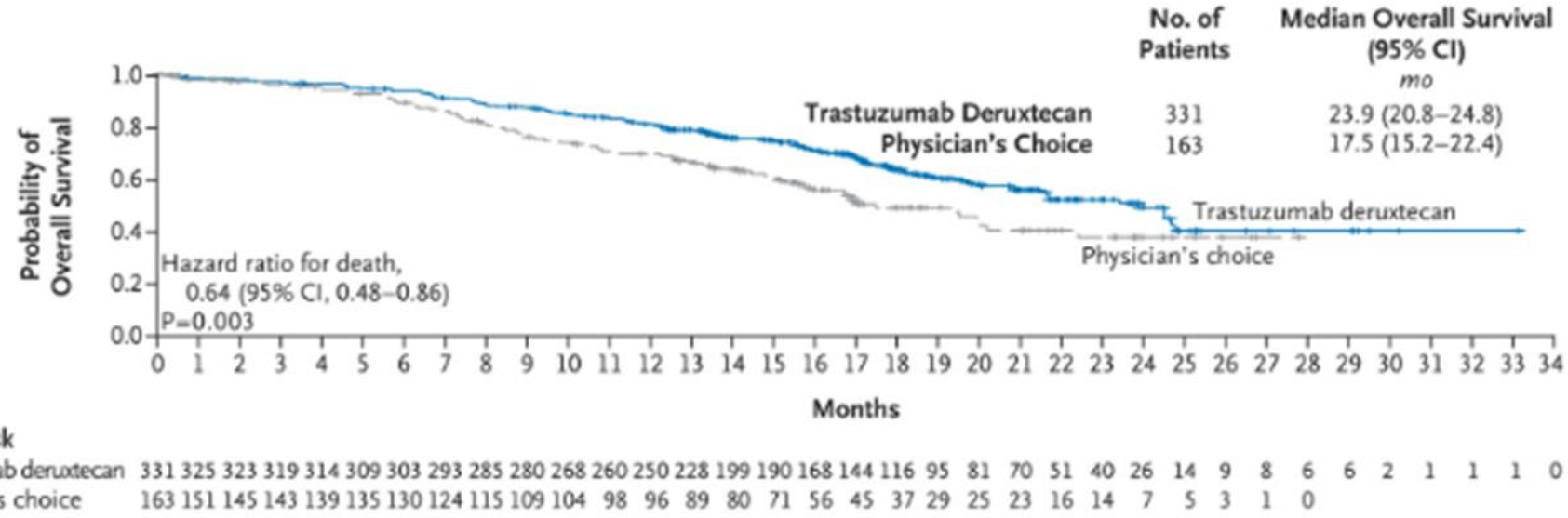
- The programs also create a .txt file with pseudo-patient level data that can be use in secondary analysis.

File Edit Format View Help

```
"","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
```

MORE EXAMPLES

C Overall Survival in Hormone Receptor–Positive Cohort



No. at Risk

Trastuzumab deruxtecan	331	325	323	319	314	309	303	293	285	280	268	260	250	228	199	190	168	144	116	95	81	70	51	40	26	14	9	8	6	6	2	1	1	1	0
Physician's choice	163	151	145	143	139	135	130	124	115	109	104	98	96	89	80	71	56	45	37	29	25	23	16	14	7	5	3	1	0						

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.

Cont'd.

	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
0	168	16	41	46
1	116	18	47	53
2	81	20	54	57
3	51	22	58	59
4	26	24	60	62
5	9	26	63	63

Cont'd.

Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Code IPD from KM DB04 - Copy.R | Code IPD from KM DB04.R*

Source on Save

```
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]
```

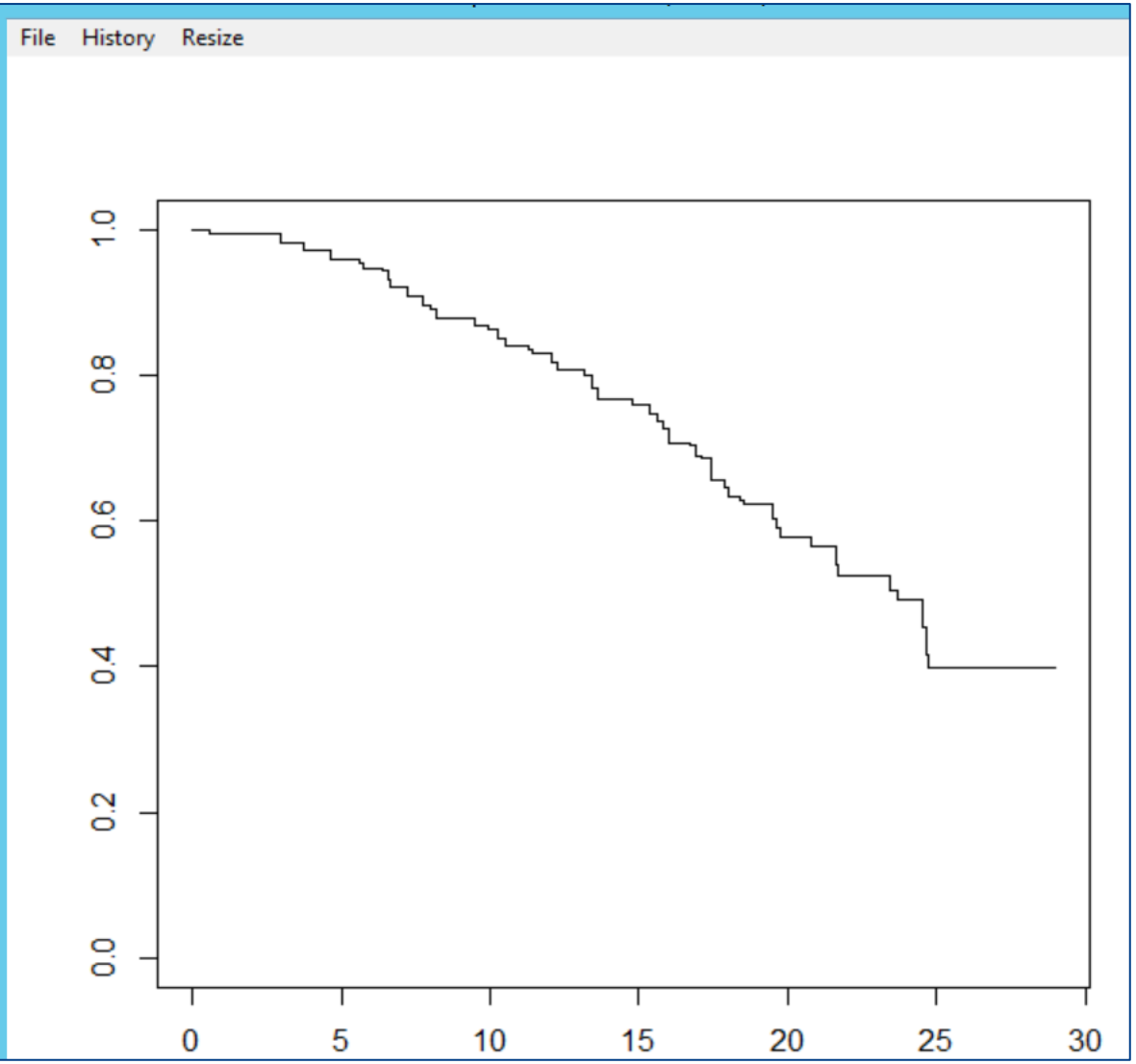
05:1 (Top Level)

Terminal | Jobs

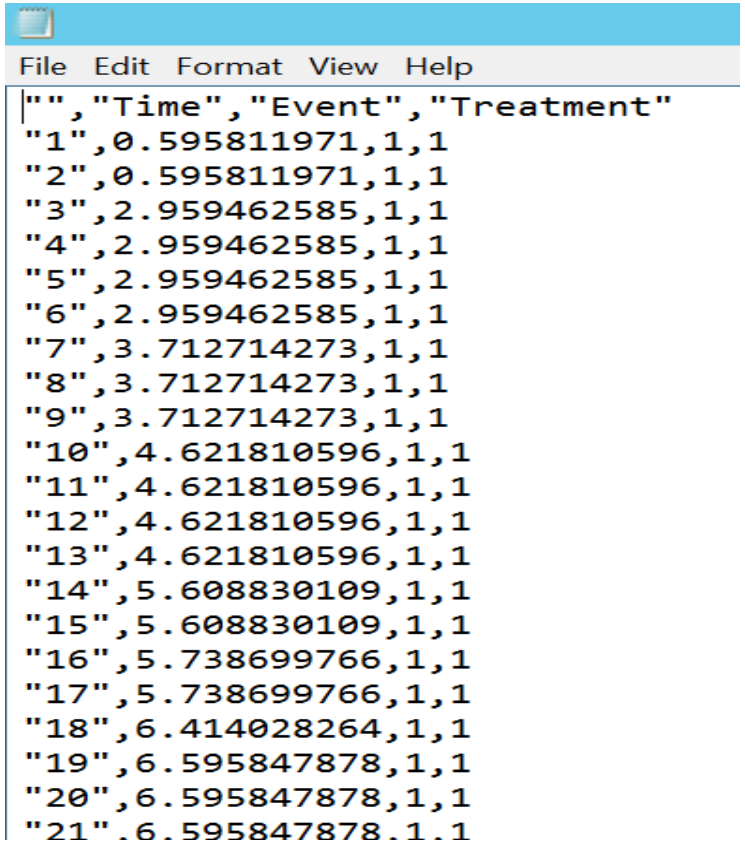
:/New Folder/

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

File History Resize



Cont'd.


File Edit Format View Help
"","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

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 very useful in secondary analysis to support a number of different objectives, including indirect treatment comparisons within the context of economic evaluations

OTHER 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)

USEFUL LINKS

Documentation:

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)

QUESTIONS

Contact Authors:

Ajgupta@dsi.com

Natalie.DENNIS@daiichi-sankyo.eu