Extracting Data from Publications to inform Clinical Trials

Vincent Wolowski
June 2023

Data Extraction from Publications

Rindopepimut with temozolomide for patients with newly diagnosed, EGFRvIII-expressing glioblastoma (ACT IV): a randomised, double-blind, international phase 3 trial

*Michael Weller, Nicholas Butowski, David D Tran, Lawrence D Recht, Michael Lim, Hal Hirte, Lynn Ashby, Laszlo Mechtler, Samuel A Goldlust, Fabio Iwamoto, Jan Drappatz, Donald M O'Rourke, Mark Wong, Mark G Hamilton, Gaetano Finocchiaro, James Perry, Wolfgang Wick, Jennifer Green, Yi He, Christopher D Turner, Michael J Yellin, Tibor Keler, Thomas A Davis, Roger Stupp, and John H Sampson, for the ACT IV trial investigators**

Summary

Background Rindopepimut (also known as CDX-110), a vaccine targeting the *EGFR* deletion mutation EGFRvIII, consists of an EGFRvIII-specific peptide conjugated to keyhole limpet haemocyanin. In the ACT IV study, we aimed to assess whether or not the addition of rindopepimut to standard chemotherapy is able to improve survival in patients with EGFRvIII-positive glioblastoma.

Methods In this randomised, double-blind, phase 3 trial, we recruited patients aged 18 years and older with glioblastoma from 165 hospitals in 22 countries. Eligible patients had newly diagnosed glioblastoma confirmed to express EGFRvIII by central analysis, and had undergone maximal surgical resection and completion of standard chemoradiation without progression. Patients were stratified by European Organisation for Research and Treatment of Cancer recursive partitioning analysis class, *MGMT* promoter methylation, and geographical region, and randomly assigned (1:1) with a prespecified randomisation sequence (block size of four) to receive rindopepimut (500 μg admixed with 150 μg GM-CSF) or control (100 μg keyhole limpet haemocyanin) via monthly intradermal injection until progression or intolerance, concurrent with standard oral temozolomide (150–200 mg/m² for 5 of 28 days) for 6–12 cycles or longer. Patients, investigators, and the trial funder were masked to treatment allocation. The primary endpoint was overall survival in patients with minimal residual disease (MRD; enhancing tumour <2 cm² post-chemoradiation by central review), analysed by modified intention to treat. This trial is registered with ClinicalTrials.gov, number NCT01480479.

Findings Between April 12, 2012, and Dec 15, 2014, 745 patients were enrolled (405 with MRD, 338 with significant residual disease [SRD], and two unevaluable) and randomly assigned to rindopepimut and temozolomide (n=371) or control and temozolomide (n=374). The study was terminated for futility after a preplanned interim analysis. At final analysis, there was no significant difference in overall survival for patients with MRD: median overall survival was 20·1 months (95% CI 18·5–22·1) in the rindopepimut group versus 20·0 months (18·1–21·9) in the control group (HR 1·01, 95% CI 0·79–1·30; p=0·93). The most common grade 3–4 adverse events for all 369 treated patients in the rindopepimut group versus 372 treated patients in the control group were: thrombocytopenia (32 [9%] *vs* 23 [6%]), fatigue (six [2%] *vs* 19 [5%]), brain oedema (eight [2%] *vs* 11 [3%]), seizure (nine [2%] *vs* eight [2%]), and headache (six [2%] *vs* ten [3%]). Serious adverse events included seizure (18 [5%] *vs* 22 [6%]) and brain oedema (seven [2%] *vs* 12 [3%]). 16 deaths in the study were caused by adverse events (nine [4%] in the rindopepimut group and seven [3%] in the control group), of which one—a pulmonary embolism in a 64-year-old male patient after 11 months of treatment—was assessed as potentially related to rindopepimut.

Interpretation Rindopepimut did not increase survival in patients with newly diagnosed glioblastoma. Combination approaches potentially including rindopepimut might be required to show efficacy of immunotherapy in glioblastoma.

Funding Celldex Therapeutics, Inc.

Lancet Oncol 2017; 18: 1373–85

Published Online August 22, 2017 http://dx.doi.org/10.1016/S1470-2045(17)30517-X

See Comment page 1294

*Investigators who participated in this trial are listed in the appendix

Department of Neurology (Prof M Weller MD) and Department of Oncology (Prof R Stupp MD), University Hospital and University of Zurich, Zurich, Switzerland; Department of Neurological Surgery, University of California, San Francisco, CA, USA (N Butowski MD); Washington University, St Louis, MO, USA (D D Tran MD); Stanford University Medical Center, Palo Alto, CA, USA (L D Recht MD); The Johns Hopkins Hospital, Baltimore, MD, USA (M Lim MD); Ivanwieski Cancer Centre, Hamilton, ON, Canada (H Hirte MD); Barrow Neurological Institute, Phoenix, AZ, USA (L Ashby MD); DENT Neurologic Institute, Buffalo, NY, USA (L Mechtler, MD); John Theurer Cancer Center, Hackensack, NJ, USA (S A Goldlust MD); Columbia University Medical Center, New York, NY, USA (F Iwamoto MD); University of Pittsburgh Medical Center, Pittsburgh, PA, USA (J Drappatz MD); Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA (D M O'Rourke MD); Westmead Hospital, Westmead, NSW, Australia (M Wong MD); University of Calgary, Department of Clinical Neurosciences, Division of Neurosurgery, Foothills Hospital, Calgary, AB, Canada (Prof M G Hamilton MD); Fondazione IRCCS Istituto

Introduction

Glioblastoma is the most common malignant primary brain tumour in adults. Its annual incidence is more than three per 100 000 people worldwide without major regional variation, and men are affected more frequently than women.[1] The standard of care—maximum feasible surgical resection followed by radiotherapy with concomitant and maintenance temozolomide chemotherapy—generally leads to a median overall survival of about 15 months.[2,3]

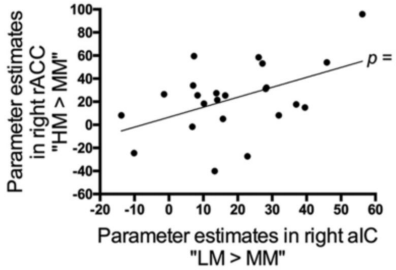
The tumour-treating fields device, recently reported to extend survival to 20·5 months, represents an additional treatment option for glioblastoma.[4] Treatment at recurrence, which might include second surgery, re-irradiation, alkylating chemotherapy using nitrosoureas such as lomustine or temozolomide rechallenge, or antiangiogenic therapy using bevacizumab, is less well standardised and has not shown a significant improvement in survival in a randomised trial. Poor prognostic factors include poor performance status, older age, incomplete

Tabular data

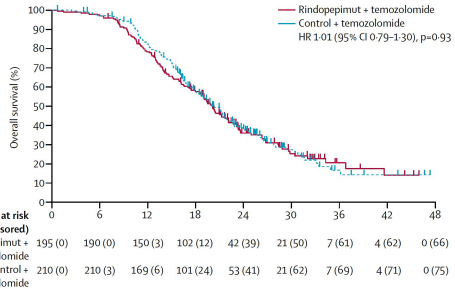
Table 1. Main Characteristics of the Study Population.*

Characteristic	Patients with Spontaneous Thrombosis (N=153)	Patients with Secondary Thrombosis (N=146)	Control Subjects (N=150)
Age — yr	67.0±16.7	65.8±17.4	65.4±15.7
Male sex — no. (%)	71 (46.4)	65 (44.5)	68 (45.3)
Smoker — no. (%)	40 (26.1)	49 (33.6)	45 (30.0)
Hypertension — no. (%)	46 (30.1)	37 (25.3)	46 (30.7)
Hyperlipidemia — no. (%)	25 (16.3)	17 (11.6)	25 (16.7)
Obesity — no. (%)	11 (7.2)	12 (8.2)	16 (10.7)
Diabetes — no. (%)	16 (10.5)	12 (8.2)	18 (12.0)
Screened for thrombophilia — no. (%)	68 (44.4)	64 (43.8)	—
Thrombophilia — no.	25†	15‡	—

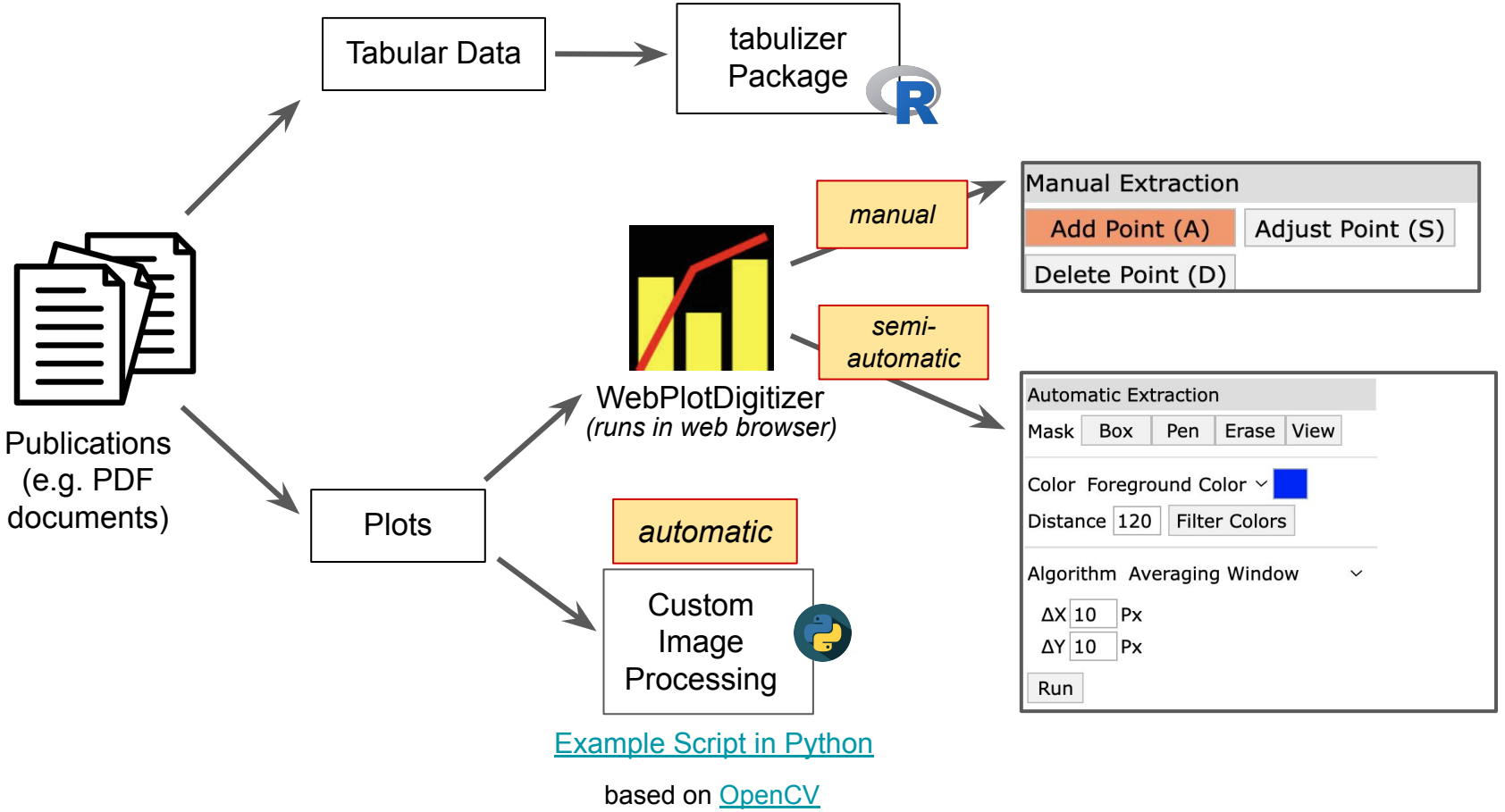
Plots



Survival curves

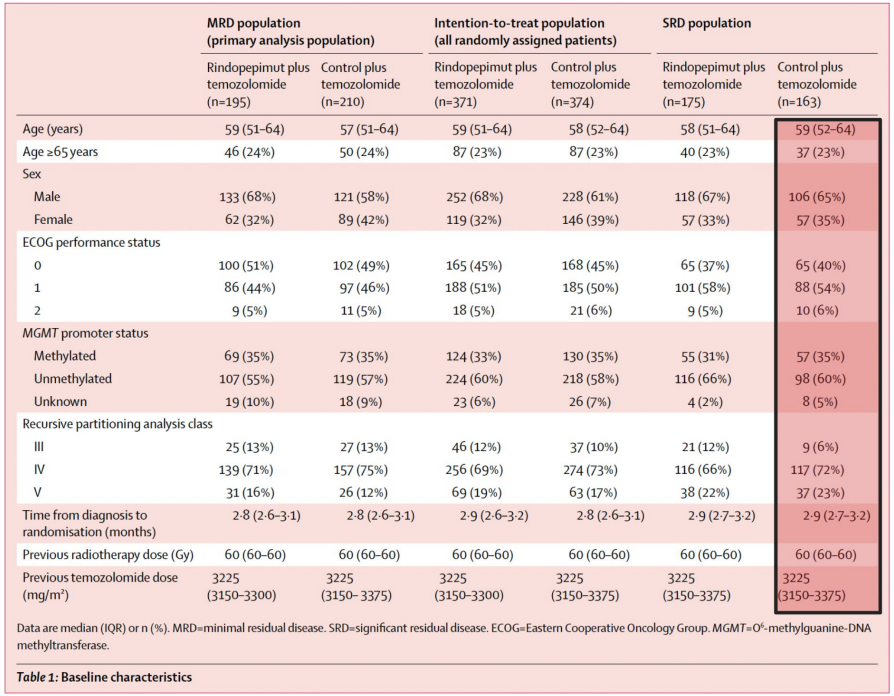


Overview: Data Extraction Modes



Data Extraction of published Tables

tabulizer R package: github, tutorial



	MRD population (primary analysis population)		Intention-to-treat population (all randomly assigned patients)		SRD population	
	Rindopepimut plus temozolomide (n=195)	Control plus temozolomide (n=210)	Rindopepimut plus temozolomide (n=371)	Control plus temozolomide (n=374)	Rindopepimut plus temozolomide (n=175)	Control plus temozolomide (n=163)
Age (years)	59 (51–64)	57 (51–64)	59 (51–64)	58 (52–64)	58 (51–64)	59 (52–64)
Age ≥65 years	46 (24%)	50 (24%)	87 (23%)	87 (23%)	40 (23%)	37 (23%)
Sex						
Male	133 (68%)	121 (58%)	252 (68%)	228 (61%)	118 (67%)	106 (65%)
Female	62 (32%)	89 (42%)	119 (32%)	146 (39%)	57 (33%)	57 (35%)
ECOG performance status						
0	100 (51%)	102 (49%)	165 (45%)	168 (45%)	65 (37%)	65 (40%)
1	86 (44%)	97 (46%)	188 (51%)	185 (50%)	101 (58%)	88 (54%)
2	9 (5%)	11 (5%)	18 (5%)	21 (6%)	9 (5%)	10 (6%)
MGMT promoter status						
Methylated	69 (35%)	73 (35%)	124 (33%)	130 (35%)	55 (31%)	57 (35%)
Unmethylated	107 (55%)	119 (57%)	224 (60%)	218 (58%)	116 (66%)	98 (60%)
Unknown	19 (10%)	18 (9%)	23 (6%)	26 (7%)	4 (2%)	8 (5%)
Recursive partitioning analysis class						
III	25 (13%)	27 (13%)	46 (12%)	37 (10%)	21 (12%)	9 (6%)
IV	139 (71%)	157 (75%)	256 (69%)	274 (73%)	116 (66%)	117 (72%)
V	31 (16%)	26 (12%)	69 (19%)	63 (17%)	38 (22%)	37 (23%)
Time from diagnosis to randomisation (months)	2·8 (2·6–3·1)	2·8 (2·6–3·1)	2·9 (2·6–3·2)	2·8 (2·6–3·1)	2·9 (2·7–3·2)	2·9 (2·7–3·2)
Previous radiotherapy dose (Gy)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)
Previous temozolomide dose (mg/m²)	3225 (3150–3300)	3225 (3150–3375)	3225 (3150–3300)	3225 (3150–3375)	3225 (3150–3375)	3225 (3150–3375)

Data are median (IQR) or n (%). MRD=minimal residual disease. SRD=significant residual disease. ECOG=Eastern Cooperative Oncology Group. *MGMT*=O⁶-methylguanine-DNA methyltransferase.

Table 1: **Baseline characteristics**

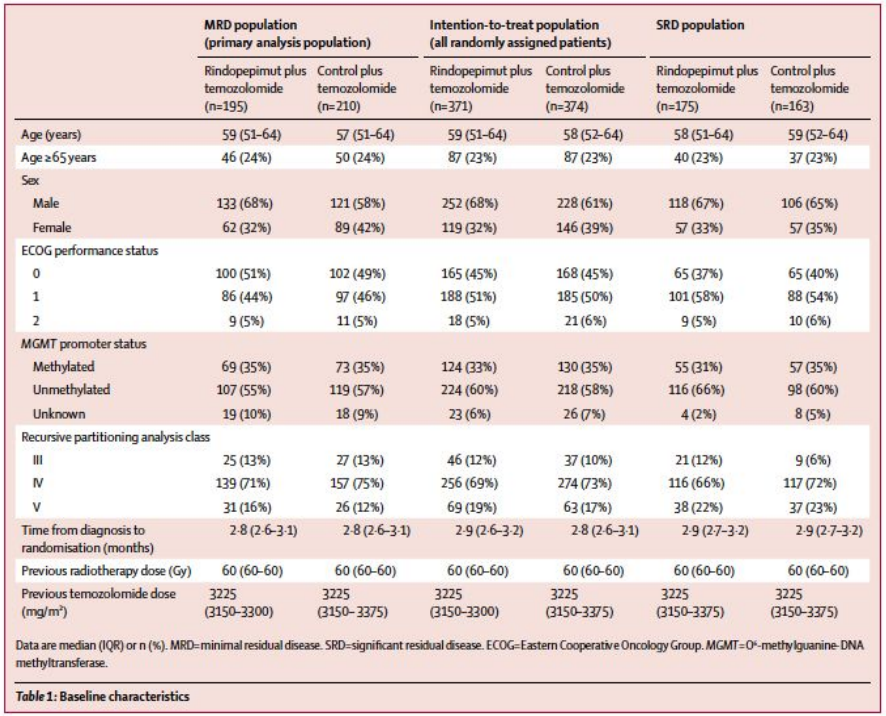
interactive

```
> out
[[1]]
      [,1]
 [1,] "59 (52–64)"
 [2,] "37 (23%)"
 [3,] "106 (65%)"
 [4,] "57 (35%)"
 [5,] "65 (40%)"
 [6,] "88 (54%)"
 [7,] "10 (6%)"
 [8,] "57 (35%)"
 [9,] "98 (60%)"
[10,] "8 (5%)"
[11,] "9 (6%)"
[12,] "117 (72%)"
[13,] "37 (23%)"
[14,] "2·9 (2·7–3·2)"
[15,] "60 (60–60)"
[16,] "3225"
[17,] "(3150–3375)"
```

```
out <- tabulizer::extract_areas('paper.pdf', pages = 6)
```

Data Extraction of published Tables

→ some further manual processing required



	MRD population (primary analysis population)		Intention-to-treat population (all randomly assigned patients)		SRD population	
	Rindopepimut plus temozolomide (n=195)	Control plus temozolomide (n=210)	Rindopepimut plus temozolomide (n=371)	Control plus temozolomide (n=374)	Rindopepimut plus temozolomide (n=175)	Control plus temozolomide (n=163)
Age (years)	59 (51–64)	57 (51–64)	59 (51–64)	58 (52–64)	58 (51–64)	59 (52–64)
Age ≥65 years	46 (24%)	50 (24%)	87 (23%)	87 (23%)	40 (23%)	37 (23%)
Sex						
Male	133 (68%)	121 (58%)	252 (68%)	228 (61%)	118 (67%)	106 (65%)
Female	62 (32%)	89 (42%)	119 (32%)	146 (39%)	57 (33%)	57 (35%)
ECOG performance status						
0	100 (51%)	102 (49%)	165 (45%)	168 (45%)	65 (37%)	65 (40%)
1	86 (44%)	97 (46%)	188 (51%)	185 (50%)	101 (58%)	88 (54%)
2	9 (5%)	11 (5%)	18 (5%)	21 (6%)	9 (5%)	10 (6%)
MGMT promoter status						
Methylated	69 (35%)	73 (35%)	124 (33%)	130 (35%)	55 (31%)	57 (35%)
Unmethylated	107 (55%)	119 (57%)	224 (60%)	218 (58%)	116 (66%)	98 (60%)
Unknown	19 (10%)	18 (9%)	23 (6%)	26 (7%)	4 (2%)	8 (5%)
Recursive partitioning analysis class						
III	25 (13%)	27 (13%)	46 (12%)	37 (10%)	21 (12%)	9 (6%)
IV	139 (71%)	157 (75%)	256 (69%)	274 (73%)	116 (66%)	117 (72%)
V	31 (16%)	26 (12%)	69 (19%)	63 (17%)	38 (22%)	37 (23%)
Time from diagnosis to randomisation (months)	2·8 (2·6–3·1)	2·8 (2·6–3·1)	2·9 (2·6–3·2)	2·8 (2·6–3·1)	2·9 (2·7–3·2)	2·9 (2·7–3·2)
Previous radiotherapy dose (Gy)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)
Previous temozolomide dose (mg/m²)	3225 (3150–3300)	3225 (3150–3375)	3225 (3150–3300)	3225 (3150–3375)	3225 (3150–3375)	3225 (3150–3375)

Data are median (IQR) or n (%). MRD=minimal residual disease. SRD=significant residual disease. ECOG=Eastern Cooperative Oncology Group. MGMT=O⁶-methylguanine-DNA methyltransferase.

Table 1: Baseline characteristics

	V2	V3	V4	V5	V6	V7	V8
1		MRD population		Intention-to-treat p	opulation	SRD population	
2		(primary analysis po	pulation)	(all randomly assigne	d patients)		
3		Rindopepimut plus	Control plus	Rindopepimut plus	Control plus	Rindopepimut plus	Control plus
4		temozolomide	temozolomide	temozolomide	temozolomide	temozolomide	temozolomide
5		(n=195)	(n=210)	(n=371)	(n=374)	(n=175)	(n=163)
6	Age (years)	59 (51-64)	57 (51-64)	59 (51-64)	58 (52-64)	58 (51-64)	59 (52-64)
7	Age ≥65 years	46 (24%)	50 (24%)	87 (23%)	87 (23%)	40 (23%)	37 (23%)
8	Sex						
9	Male	133 (68%)	121 (58%)	252 (68%)	228 (61%)	118 (67%)	106 (65%)
10	Female	62 (32%)	89 (42%)	119 (32%)	146 (39%)	57 (33%)	57 (35%)
11	ECOG performance status						
12	0	100 (51%)	102 (49%)	165 (45%)	168 (45%)	65 (37%)	65 (40%)
13	1	86 (44%)	97 (46%)	188 (51%)	185 (50%)	101 (58%)	88 (54%)
14	2	9 (5%)	11 (5%)	18 (5%)	21 (6%)	9 (5%)	10 (6%)
15	MGMT promoter status						
16	Methylated	69 (35%)	73 (35%)	124 (33%)	130 (35%)	55 (31%)	57 (35%)
17	Unmethylated	107 (55%)	119 (57%)	224 (60%)	218 (58%)	116 (66%)	98 (60%)
18	Unknown	19 (10%)	18 (9%)	23 (6%)	26 (7%)	4 (2%)	8 (5%)
19	Recursive partitioning analysis clas	s					
20	III	25 (13%)	27 (13%)	46 (12%)	37 (10%)	21 (12%)	9 (6%)
21	IV	139 (71%)	157 (75%)	256 (69%)	274 (73%)	116 (66%)	117 (72%)
22	V	31 (16%)	26 (12%)	69 (19%)	63 (17%)	38 (22%)	37 (23%)
23	Time from diagnosis to	2·8 (2·6–3·1)	2·8 (2·6–3·1	) 2·9 (2·6–3·2)	2·8 (2·6–3·1)	2·9 (2·7–3·2)	2·9 (2·7–3·2)
24	randomisation (months)						
25	Previous radiotherapy dose (Gy)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)	60 (60–60)
26	Previous temozolomide dose	3225	3225	3225	3225	3225	3225
27	(mg/m2)	(3150-3300)	(3150- 3375)	(3150-3300)	(3150-3375)	(3150-3375)	(3150-3375)
28	Data are median (IQR) or n (%). MRD=m	inimal residual diseas	e. SRD=significant r	esidual disease. ECOG=Easte	rn Cooperative Onco	logy Group. MGMT=O6-m	ethylguanine-DNA
29	methyltransferase.						
30	Table 1: Baseline characteristics						

```
tabulizer::extract_tables('paper.pdf', pages=6, guess=F,
                          columns=list(c(121.5,218,273.4,320.5,389.1,440.9,501.7,552.9)),
                          area=list(c(405.3,119.1,749.7,558.3)))
```

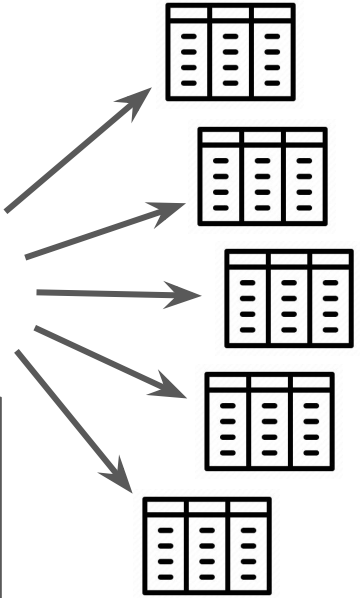
Get coordinates for columns and area via Shiny:

```
tabulizer::locate_areas('paper.pdf', pages = 6)
```

Goal for automatic Data Extraction of published Tables

PDF

```
tabulizer::extract_tables('paper.pdf', guess = TRUE)
```



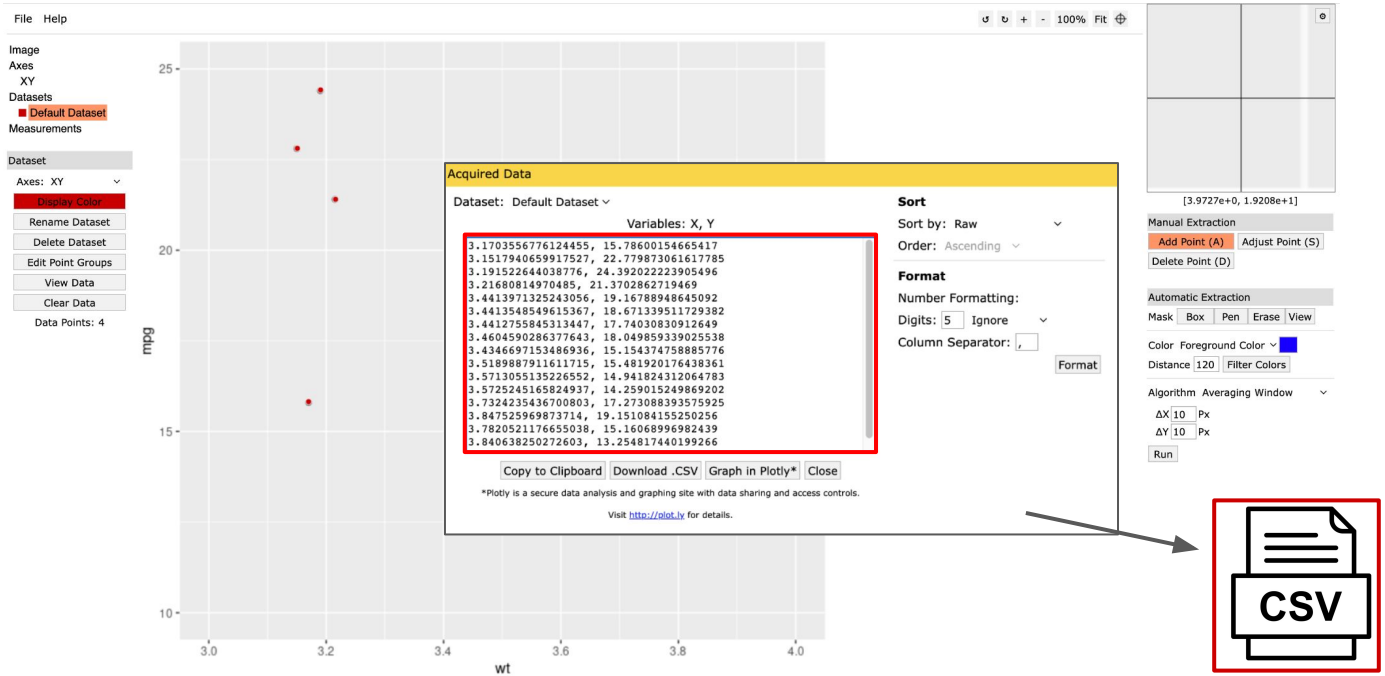
V1	V2	V3	V4	V5
	MRD population	Intention-to-treat population	SRD population	
	(primary analysis population)	(all randomly assigned patients)		
	Rindopepimut plus Control plus	Rindopepimut plus Control plus	Rindopepimut plus	Control plus
	temozolomide temozolomide	temozolomide temozolomide	temozolomide	temozolomide
	(n=195) (n=210)	(n=371) (n=374)	(n=175)	(n=163)
Age (years)	59 (51-64) 57 (51-64)	59 (51-64) 58 (52-64)	58 (51-64)	59 (52-64)
Age ≥65 years	46 (24%) 50 (24%)	87 (23%) 87 (23%)	40 (23%)	37 (23%)
Sex				
Male	133 (68%) 121 (58%)	252 (68%) 228 (61%)	118 (67%)	106 (65%)
Female	62 (32%) 89 (42%)	119 (32%) 146 (39%)	57 (33%)	57 (35%)
ECOG performance status				
0	100 (51%) 102 (49%)	165 (45%) 168 (45%)	65 (37%)	65 (40%)
1	86 (44%) 97 (46%)	188 (51%) 185 (50%)	101 (58%)	88 (54%)
2	9 (5%) 11 (5%)	18 (5%) 21 (6%)	9 (5%)	10 (6%)
MGMT promoter status				
Methylated	69 (35%) 73 (35%)	124 (33%) 130 (35%)	55 (31%)	57 (35%)
Unmethylated	107 (55%) 119 (57%)	224 (60%) 218 (58%)	116 (66%)	98 (60%)
Unknown	19 (10%) 18 (9%)	23 (6%) 26 (7%)	4 (2%)	8 (5%)
Recursive partitioning analysis class				
III	25 (13%) 27 (13%)	46 (12%) 37 (10%)	21 (12%)	9 (6%)
IV	139 (71%) 157 (75%)	256 (69%) 274 (73%)	116 (66%)	117 (72%)
V	31 (16%) 26 (12%)	69 (19%) 63 (17%)	38 (22%)	37 (23%)

Result for *Table 1*

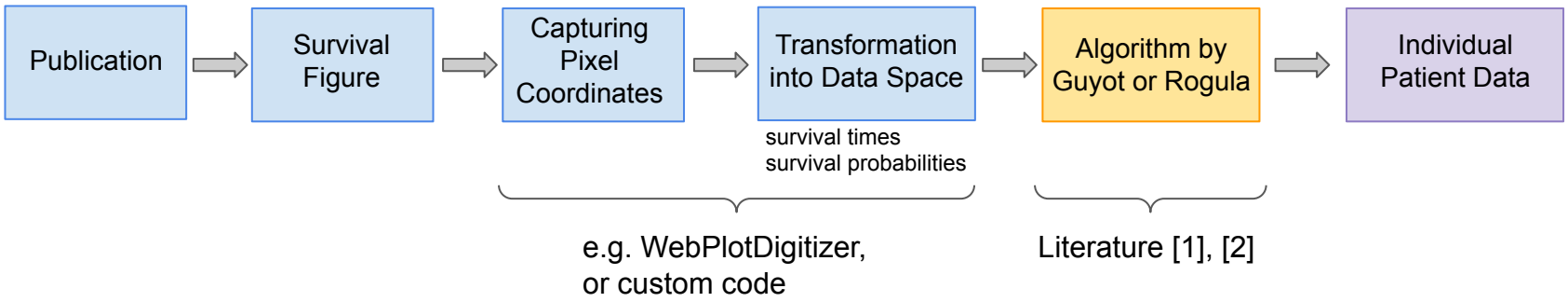
Plots: Getting from Pixel Screen Coordinates to Data Space Coordinates

Digitization tool: *WebPlotDigitizer:* https://apps.automeris.io/wpd/
User Manual: https://automeris.io/WebPlotDigitizer/userManual.pdf
Tutorial: https://automeris.io/WebPlotDigitizer/tutorial.html
Github: https://github.com/ankitrohatgi/WebPlotDigitizer

Use case: load image ⟶ set reference points on axes ⟶ manual or semi-automatic extraction ⟶ get underlying data

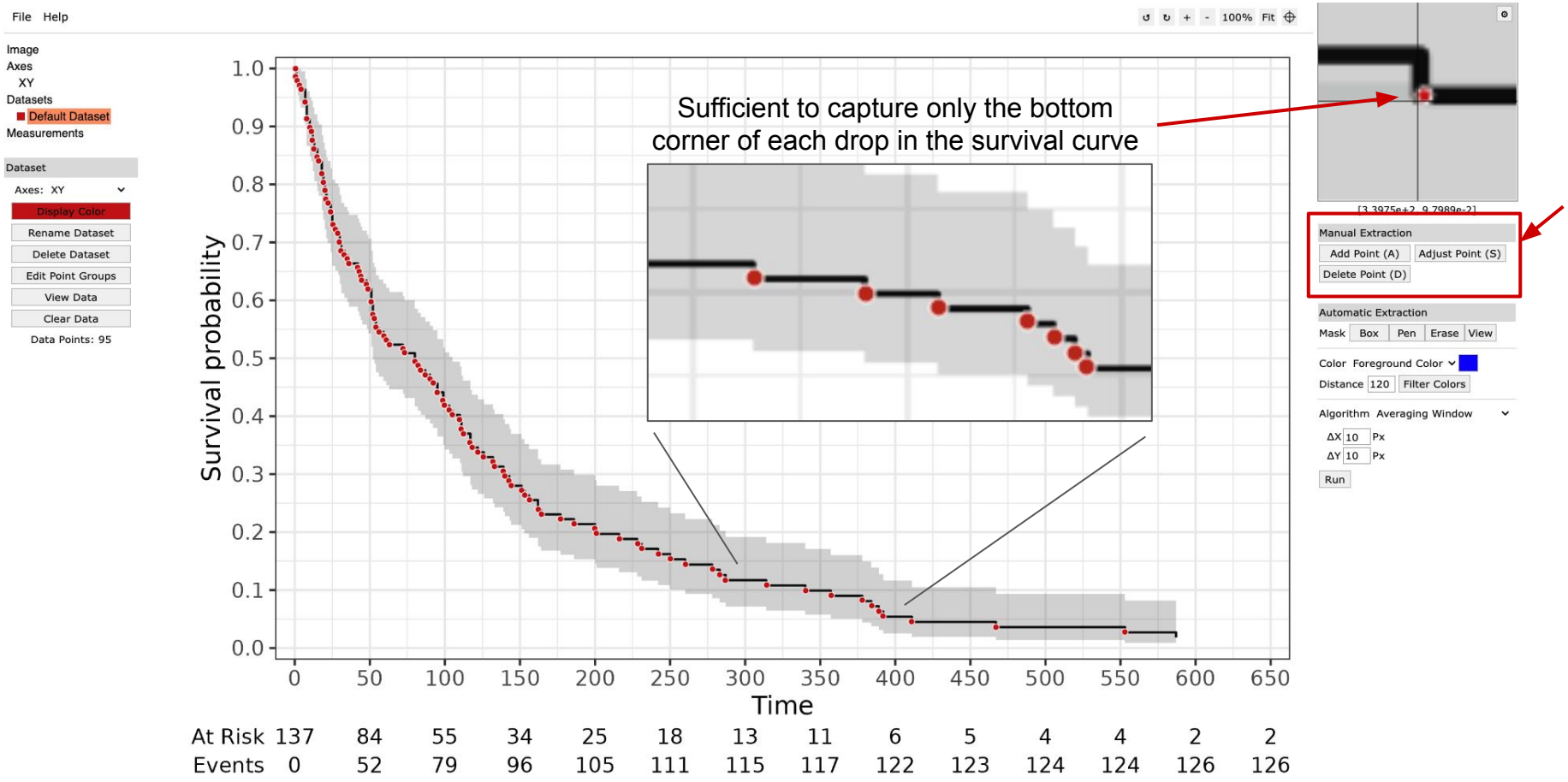


Data Extraction Workflow for Kaplan Meier Survival Curves



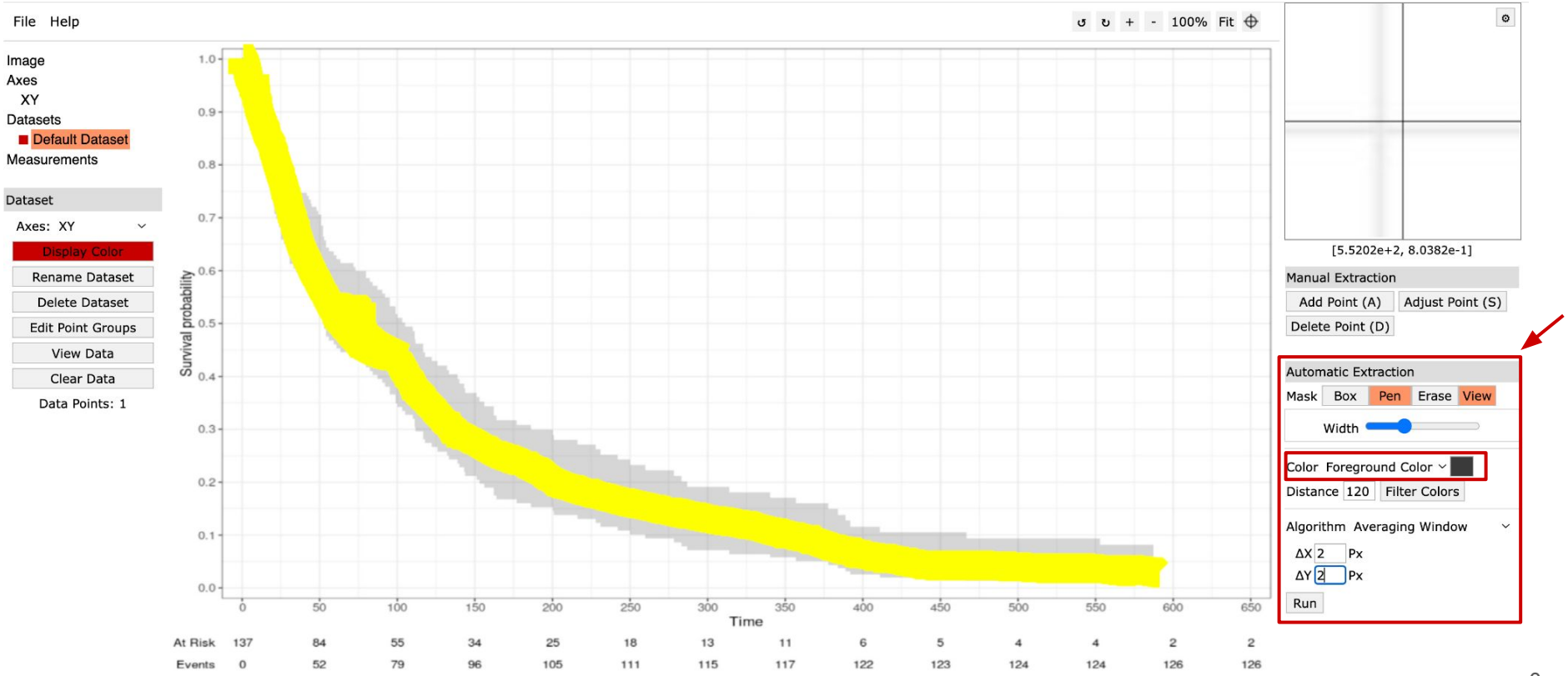
Literature:
[1] P. Guyot et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves, 2012 (link)
[2] B. Rogula et al. A Method for Reconstructing Individual Patient Data From Kaplan-Meier Survival Curves That Incorporate Marked Censoring Times, 2022 (link)

WebPlotDigitizer: Manually marking of Data Points



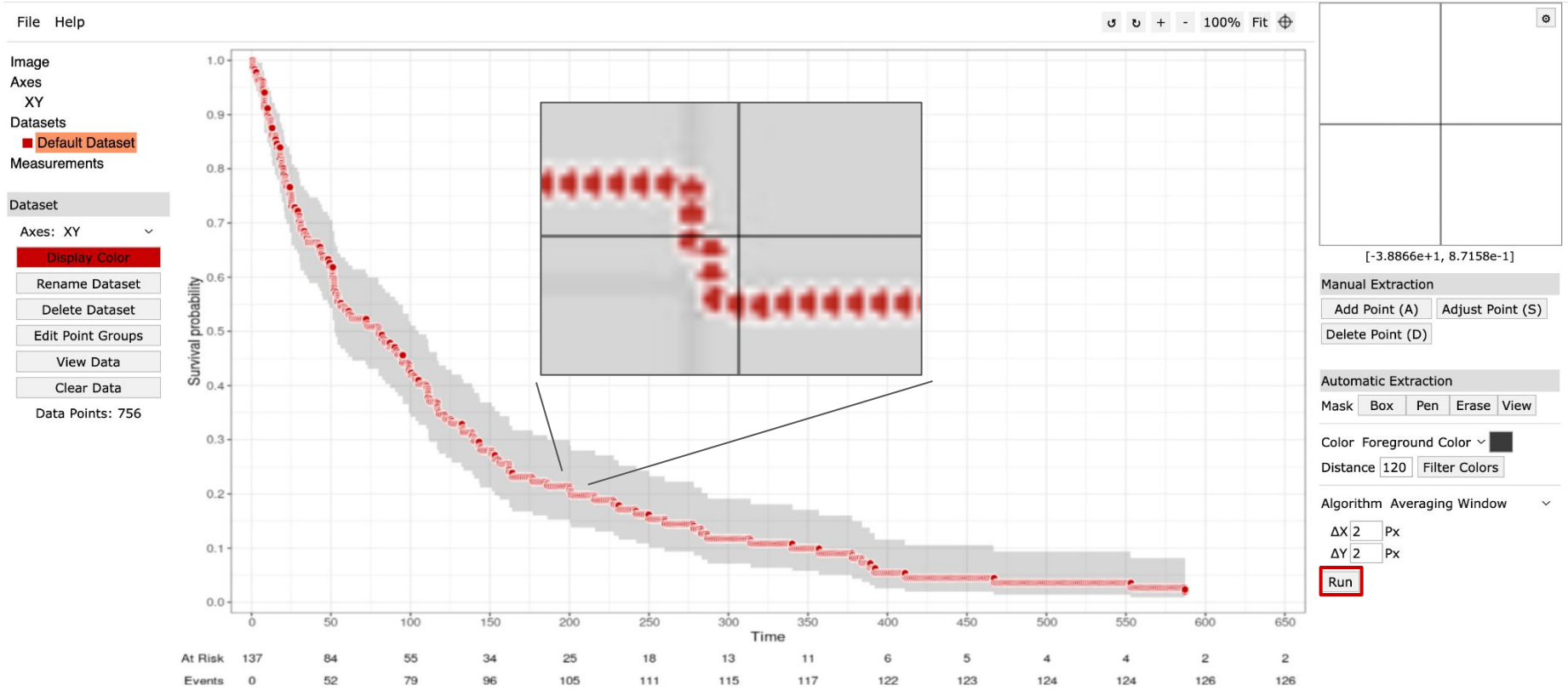
WebPlotDigitizer: Semi-Automatic Extraction of Data Points

Selecting color of KM survival curve (i.e. Color Foreground) and marking it by *Pen*



WebPlotDigitizer: Semi-Automatic Extraction

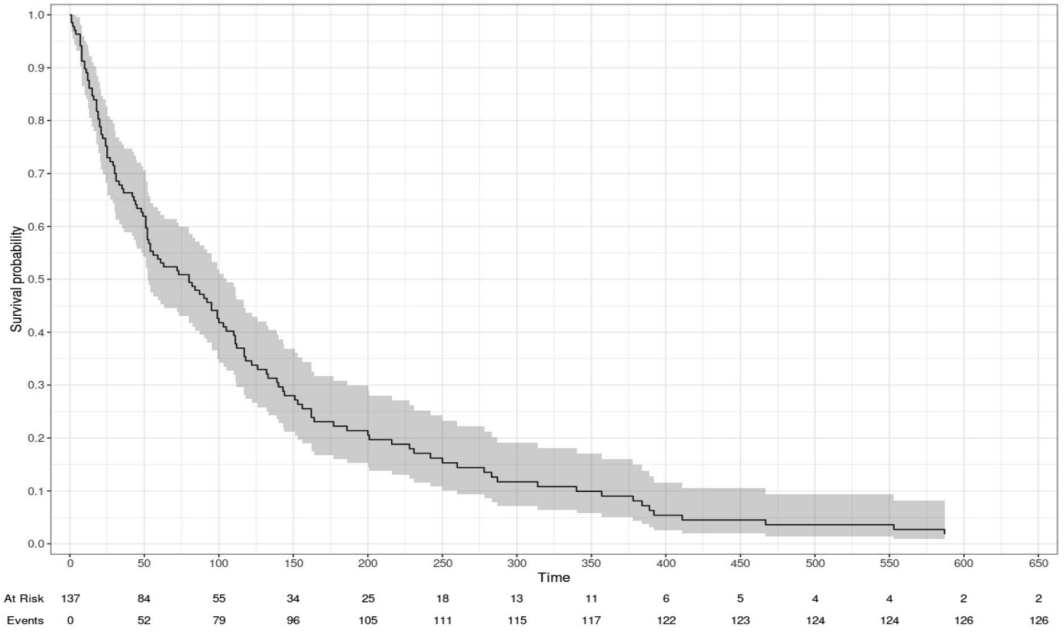
Running *Averaging Window Algorithm* to mark the curve



R Implementation based on Guyot et al to get Patient Level Data

R package ipdrecon (wrapper for Guyot and Rogula algorithm):
https://github.com/vinwol/ipdrecon

Example of reconstructed Survival Curve:



Data: Veteran Dataset from the Survival Package (randomized, two-treatment regime trial for lung cancer)

Input Parameters

- time point at which survival probabilities are observed } from digitization of KM curve
- survival probabilities at each time point
- time points at which number of patients at risk is calculated → from KM plot
- number of patients at risk at each time point } from published risk table
- total number of events (optional)

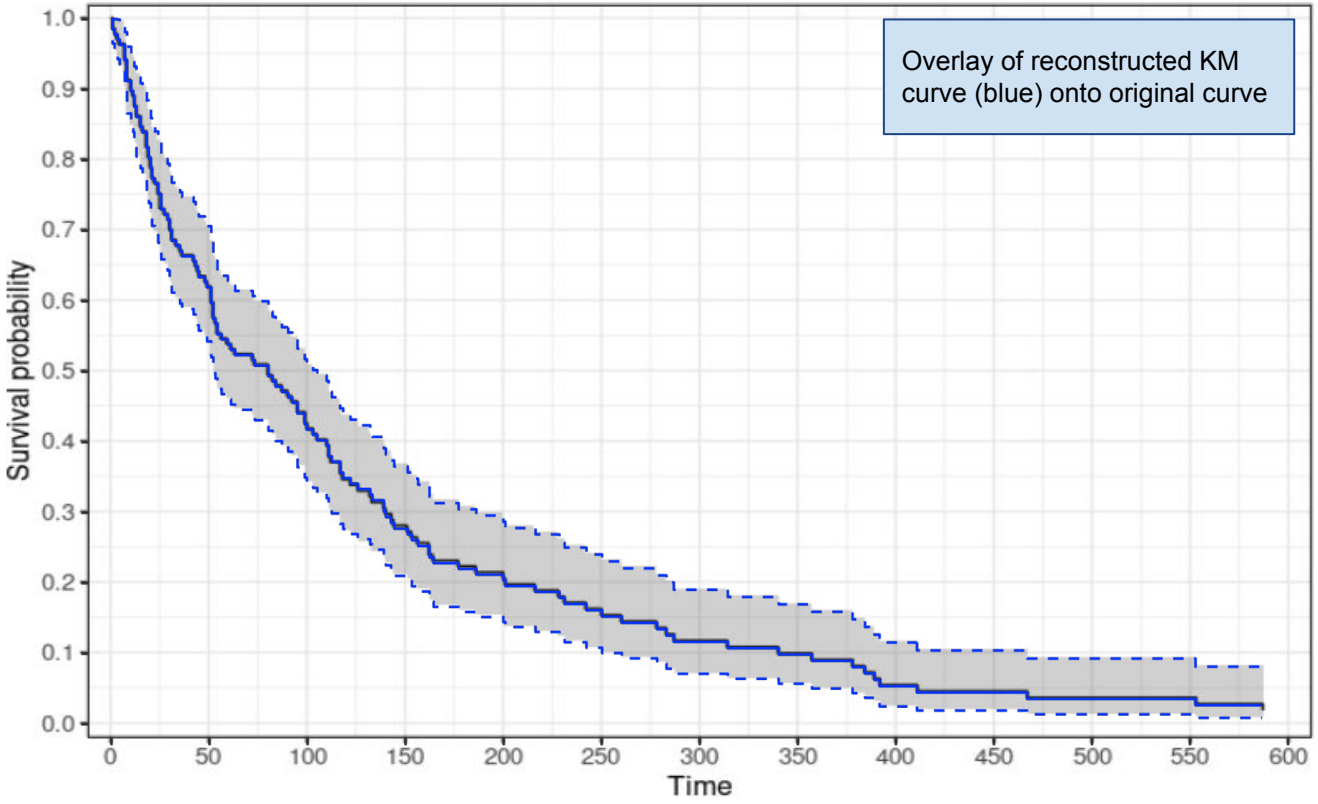
e.g. from WebPlotDigitizer

```
library(ipdrecon)

digitized_data <- read.csv('./data/digitized_data.csv', header = FALSE)
time <- digitized_data[,1]
prob <- digitized_data[,2]
trisk <- seq(from=0, to=550, by=50)
nrisk <- c(137,84,55,34,25,18,13,11,6,5,4,4)
tot_events <- 126
lower <- ipdrecon::get_lower_indices(time, trisk)
upper <- ipdrecon::get_upper_indices(time, trisk)
res <- ipdrecon::get_ipd_guyot(time,
                               prob,
                               trisk,
                               nrisk,
                               lower,
                               upper,
                               tot_events)
ipd <- res[[1]]

fit <- ggsurvfit::survfit2(survival::Surv(time, event) ~ 1, data = ipd)
ggsurvfit::ggsurvfit(fit) +
   ggplot2::labs(x = "Time", y = "Survival probability") +
   ggplot2::scale_x_continuous(limits = c(0,650),
                               breaks = seq(0,650,by=50),
                               expand = c(0.02,0)) +
   ggplot2::scale_y_continuous(limits = c(0,1),
                               breaks = seq(0,1,by=0.1),
                               expand = c(0.02,0)) +
   ggsurvfit::add_confidence_interval() +
   ggsurvfit::add_risktable()
```

Comparison of original KM Curve and reconstructed Version by Guyot et al



Comparison of survival quantiles:

Quantile	Original	Reconstructed
25%	25 (19,36)	25.4 (19.1,36)
50%	80 (52,105)	80 (51.9,104.9)
75%	162 (133,242)	162.2 (133,242.2)

At Risk	137	84	55	34	25	18	13	11	6	5	4	4	2
Events	0	52	79	96	105	111	115	117	122	123	124	124	126

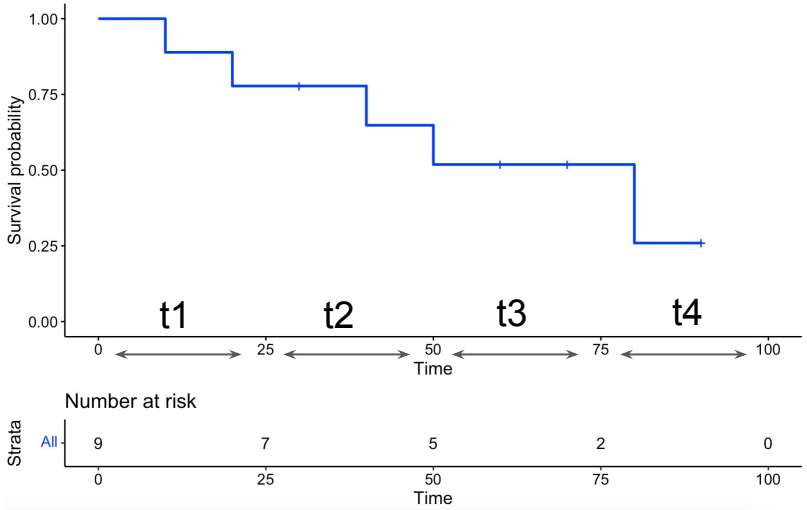
Original risk table

At Risk	137	84	55	34	25	18	13	11	6	5	4	4	0
Events	0	52	79	97	106	111	116	118	123	124	125	125	126

Reconstructed risk table

Prerequisite for Guyot et al Algorithm:
Lower and Upper Bounds of the Time Intervals

→ **Lower** and **upper** denote the ***row indices*** into the CSV file to map digitized time points to the right time intervals



Digitized using e.g. WebPlotDigitizer

	time	prob
lower: 1 upper: 3 t1	0.07745933	1.0001801
	10.06971340	0.8911200
	19.98450813	0.7803478
lower: 4 upper: 4 t2	40.04647560	0.6511192
lower: 5 upper: 5 t3	50.19364833	0.5215516
lower: 6 upper: 6 t4	80.17041053	0.2627474

lower: c(1,4,5,6)

upper: c(3,4,5,6)

Description of Guyot et al Algorithm

Main idea

Estimation of the number of censored patients, number of patients at risk and number of events for each time interval to find numerical solutions to the inverted KM equations

R function call to get reconstructed IPD:

```
ipd <- ipdrecon::get_ipd_guyot(time, prob, trisk, nrisk, lower, upper, tot_events)[[1]]
```

Input

Number of total events ⟶ **tot_events** (optional)

Digitized data from survival curve

Time	Prob
0.0775	1.0002
10.0697	0.8911
19.9845	0.7803
40.0465	0.6511
50.1936	0.5216
80.1704	0.2627

⟶ **time**: time points t1 to t6 here

⟶ **prob**: survival probabilities

Risk table

9,7,5,2
0,25,50,75

⟶ **nrisk**: number of patients at risk

⟶ **trisk**: related time points

for loop over each time interval **t[i-1]** to **t[i]**:

- calculate the estimated number of censored patients and the estimated number of patients at risk.
- distribute the censored patients evenly over the time interval.
- calculate the estimated number of events and the estimated survival probability at each extracted time point.
- adjust the number of patients at risk to account for the events and censored patients.

```
d[k] <- round(n_hat[k] * (1 - (prob[k] / KM_hat[last])))
```
$\longleftrightarrow d_m = n_m \times (1 - \frac{\hat{S}(t_m)}{\hat{S}(t_{m-1})})$

```
KM_hat[k] <- KM_hat[last] * (1 - (d[k] / n_hat[k]))
```
$\longleftrightarrow \hat{S}(t_m) = \hat{S}(t_{m-1}) \times (1 - \frac{d_m}{n_m})$

n_m: number of patients at risk at time m

d_m: number of events at time m

$S_{hat}(t_m)$: survival probability at time m

Description of Rogula et al Algorithm

Main idea

Comparison of survival probabilities at consecutive extracted time points and adding events to the reconstructed IPD as it progresses through all the different extracted time points.

R function call:

```
ipd <- ipdrecon::get_ipd_rogula(n, time, prob, cens_t)[[1]]
```

Input

Patients at risk at t=0 ⟶ **n**

Digitized data from survival curve

Time	Prob
0.0775	1.0002
10.0697	0.8911
19.9845	0.7803
40.0465	0.6511
50.1936	0.5216
80.1704	0.2627

⟶ **time**: time points t1 to t6 here

⟶ **prob**: survival probabilities

Digitized censor data

Time	Prob
29.9768	0.7824
60.0310	0.5219
69.9458	0.5239
90.0852	0.2597

⟶ **cens_t**: specific time points at which patients were censored (optional)

Additional variables:
p: number of patients
IPD: dataframe which is iteratively filled

for loop over all extracted time points (stored in *time*):

- If censoring information is provided, the number of patients who are censored between the previous and current extracted time points are determined and **p** is updated.
- If there are still patients at risk (**p < n**), it finds the number of events at the current time point that results in a survival probability at that time point closest to **prob[i]**. It does this within a **while loop** by repeatedly adding one event to the reconstructed IPD and then calculating the survival probability at that time point.

```
est_prob <- summary(survfit(Surv(time,event) ~ 1, data=IPD), time=time[i])$surv
diff <- est_prob - prob[i]
```

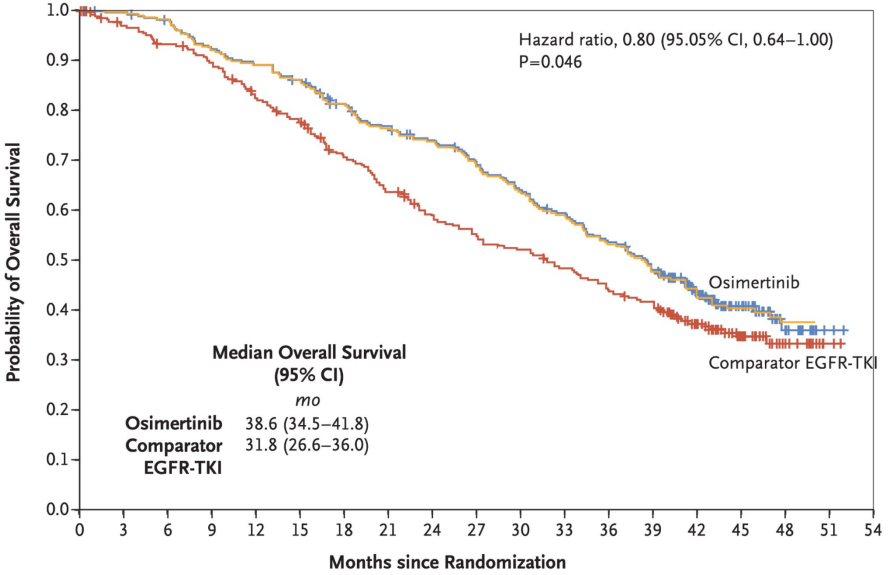
- It updates the number of patients at risk and repeats the steps above until all patients have been accounted for.

At the end, it checks for any censoring information that needs to be added at the end of the follow-up time.

Comparison using an Example from the Publication [1]

Figure 1 and outcome of Guyot's and Rogula's Algorithm for experimental arm

Guyot's Algorithm:
Overlaid reconstructed survival curve (orange) on Osimertinib

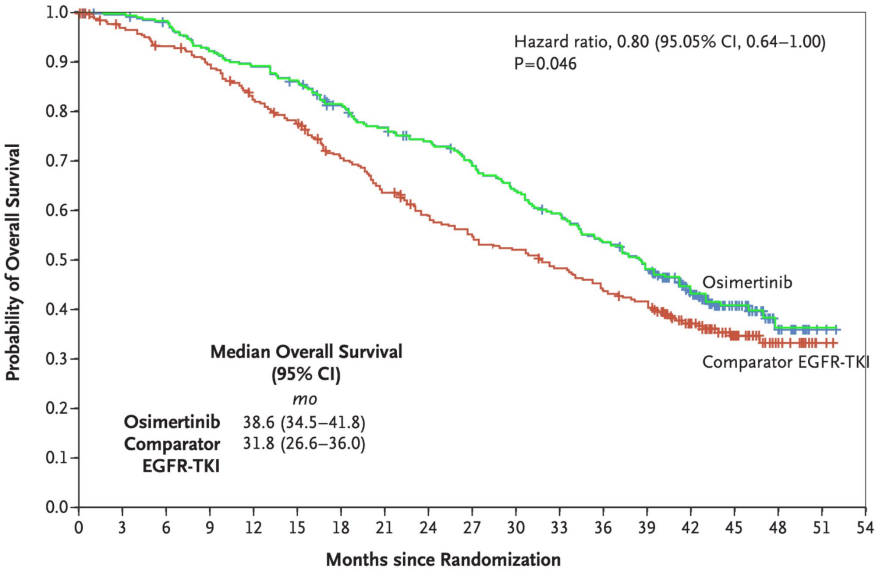


Osim.	279	276	270	254	245	236	217	204	193	180	166	153	138	123	86	50	17	2	0
Comp.	277	263	252	239	219	205	182	165	148	138	131	121	110	101	72	40	17	2	0

Risk table from Guyot's reconstructed IPD:

At Risk	279	274	268	251	243	234	215	202	191	178	164	151	136	121	91	48	37	0	0
Events	0	1	5	22	30	38	51	64	71	84	97	109	124	138	150	154	157	157	157

Rogula's Algorithm:
Overlaid reconstructed survival curve (green) on Osimertinib



Osim.	279	276	270	254	245	236	217	204	193	180	166	153	138	123	86	50	17	2	0
Comp.	277	263	252	239	219	205	182	165	148	138	131	121	110	101	72	40	17	2	0

Risk table from Rogula's reconstructed IPD:

At Risk	279	276	270	253	245	236	218	204	194	180	167	154	139	124	86	51	17	2	0
Events	0	1	5	22	30	38	51	64	71	84	97	109	124	138	149	153	156	156	156

[1] S.S. Ramalingam et al: Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC, The New England Journal of Medicine, 2020

Summary

Example R code showing a complete extraction use case (tables and survival curves)

- For extraction of tabular data, the ***R package tabulizer*** is very handy
- For plot data, ***WebPlotDigitizer*** allows **manual** or **semi-automatic extraction**
- Custom image processing (e.g. **ImageAI** + **OpenCV**) allows to extract the underlying data in an **automatic fashion,** potentially **processing** the images **fully automated** without any further input
- **Algorithm by Guyot et al** can reconstruct the underlying IPD from published KM survival curves pretty well
- **Algorithm by Rogula et al** can incorporate **explicit censoring information** if available
- It would be great if **both algorithms** could be combined into **one algorithm**
- **These extraction techniques** offer data-driven, evidence based enhancement to purely statistical approaches

Material

Publications

- M. W. Hoyle. W. Henley: Improved curve fits to summary survival data: application to economic evaluation of health technologies, 2011 (link)
- P. Guyot et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves, 2012 (link)
- X. Wan,L. Peng,Y. Li: A Review and Comparison of Methods for Recreating Individual Patient Data from Published Kaplan-Meier Survival Curves for Economic Evaluations: A Simulation Study, 2015 (link)
- Y. Wei, P. Royston: Reconstructing time-to-event data from published Kaplan–Meier curves, 2017 (link)
- N. Liu et al. IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves, 2021 (link)
- B. Rogula et al. A Method for Reconstructing Individual Patient Data From Kaplan-Meier Survival Curves That Incorporate Marked Censoring Times, 2022 (link)

Digitization Tool

- WebPlotDigitizer (link)

R Packages

- R package IPDfromKM: Map Digitized Survival Curves Back to Individual Patient Data (link, github)
- R package by B. Rogula et al. KMtoIPD (link)
- R package: KMSubtraction: Reconstruction of unreported subgroup survival data utilizing published Kaplan-Meier survival curves (link)
- R package: reconstructKM: Reconstruct Individual-Level Data from Published KM Plots (link)

Shiny Applications

- Y. Zhou, N. Liu, J. Lee: Shiny App (link)