Help File for KMshiny Application

Na Liu, J. Jack Lee

July 13, 2018

1 Description

This package is an implementation to reconstruct individual patients records (time, status) from coordinates extracted from an survival curves. Digitized softwares such as DigitizeIt(for MAC or windows) OR ScanIt(for windows) can be used to get the coordinates from published curves. Other available information like [1] reported patient numbers at risk reported at the beginning of time intervals, total number of initial patients, and total number of events, can also be involved during calculations to increase accuracy.

2 Extract Coordinates from published KM curves

2.1 Prepare your image

Obtain an image of your graph from PDF, or from your scanner. It does not matter weather the image is capture straightly. However, better resolutions will present better accuracy in the later steps. I will use the published locoregional control events in head and neck cancer KM curve as example.

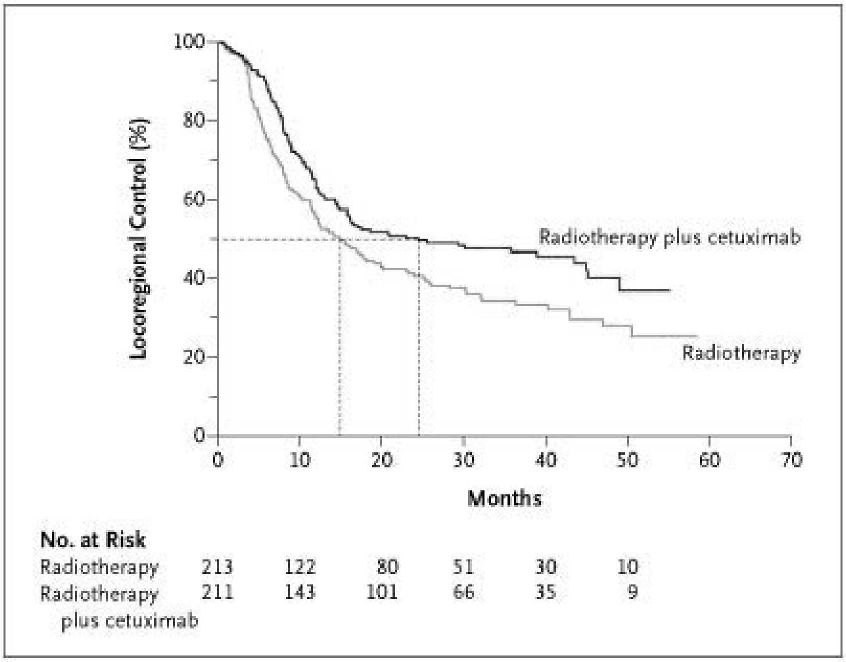
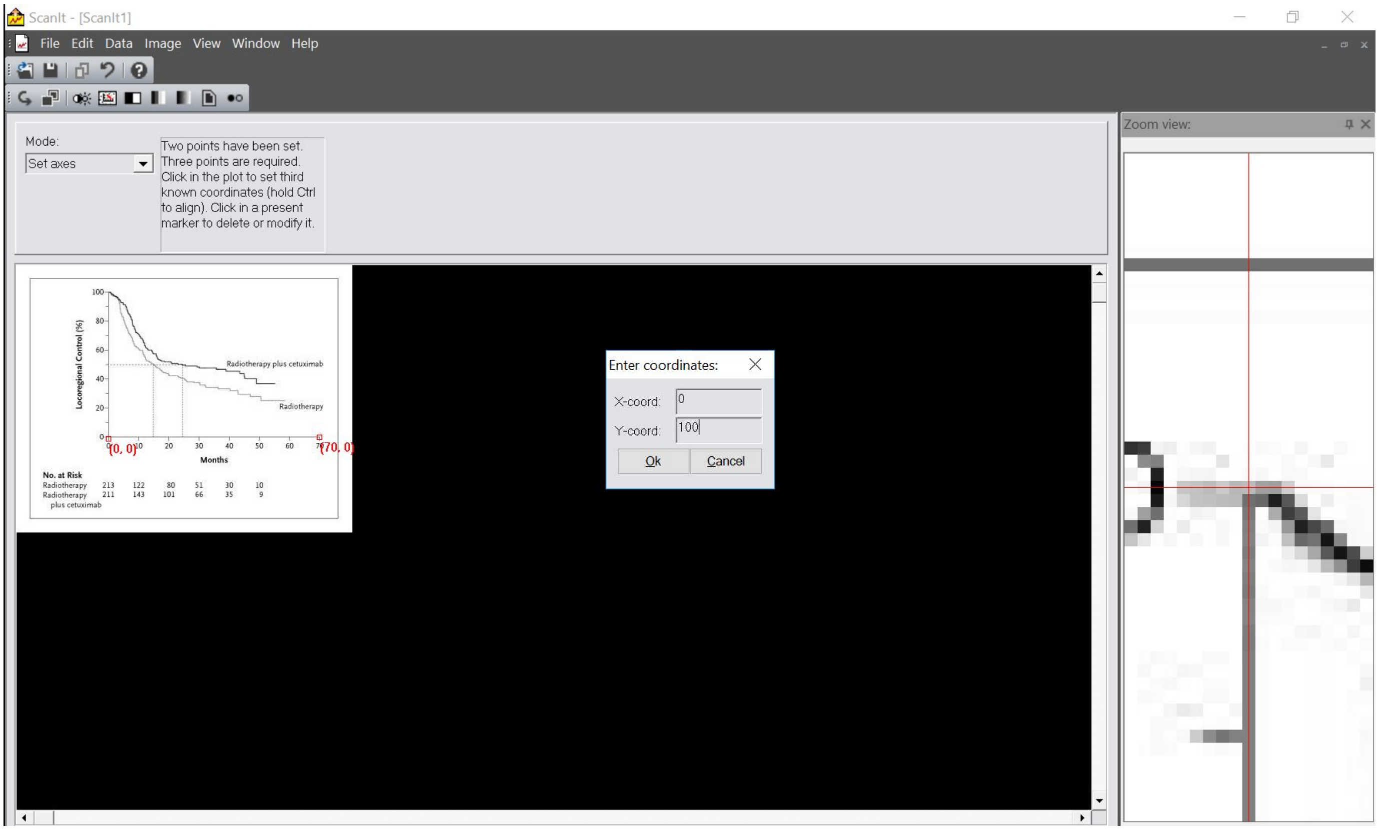


Figure 1: Locoregional control example of published Kaplan-Meier curves. We use the same sample as Guyot's paper to make comparison.

2.2 Use ScanIt software to extracting the coordinates of the survival curve.

ScanIt is a free download software to extracting data from scientific graphs, particularly from line and scatter plots.
Once open your file, first we need to identify the axes. We pick two points on each axis. The zoom view will help to pick accurately. Commonly three points are picked: the intersection of the two axes and another point at each axis. Two separate points on each axis are also supported. ScanIt supports linear and logarithmic axes.



Mode: Set axes

Two points have been set. Three points are required. Click in the plot to set third known coordinates (hold Ctrl to align). Click in a present marker to delete or modify it.

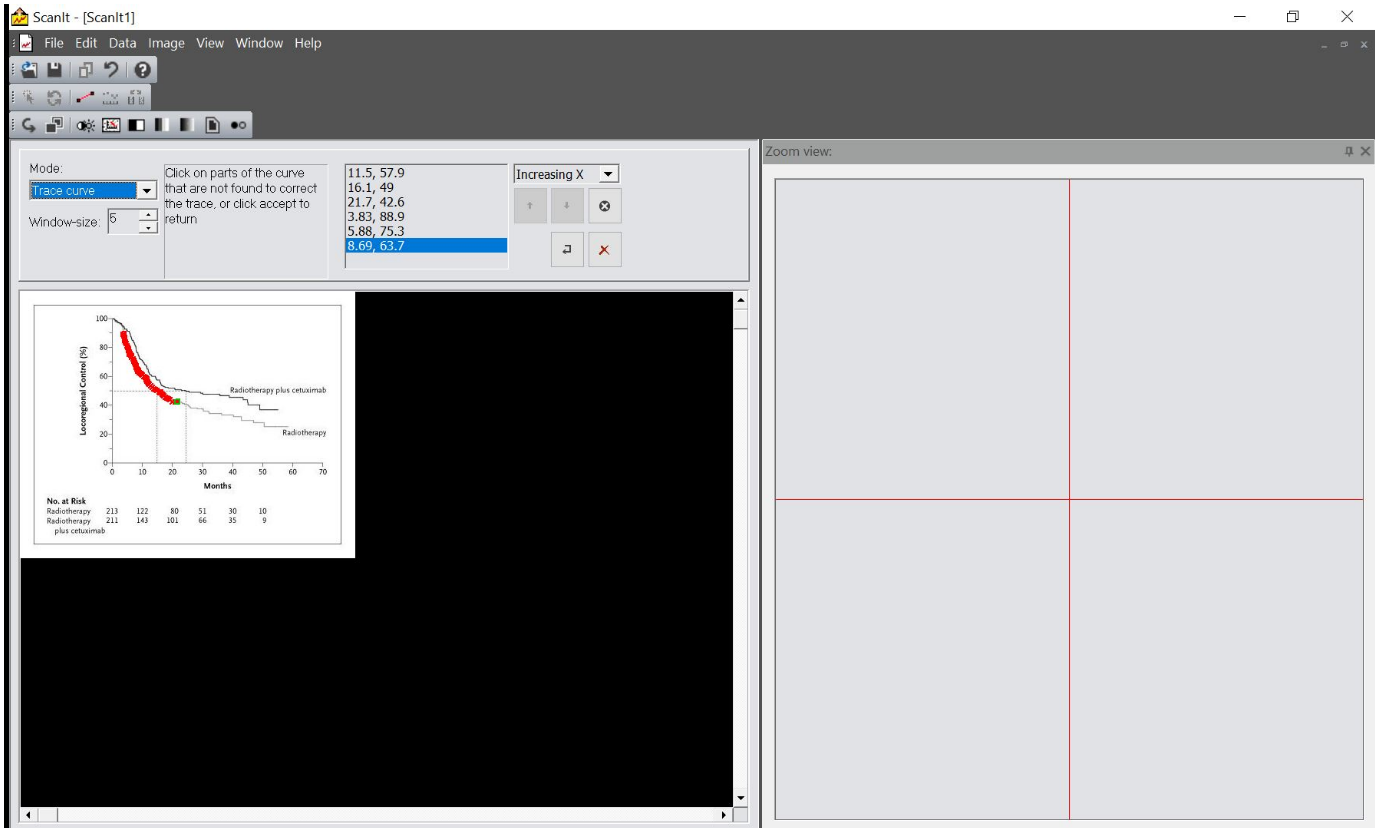
No. at Risk						
Radiotherapy	213	122	80	51	30	10
Radiotherapy plus cetuximab	211	143	101	66	35	9

X-coord: 0

Y-coord: 100

Figure 2: Setting up axis positions
We use zoom view to help pick accuracy.

Then we choose ”Trace curve” mode, where entire curves can be digitized by selecting only a few points (2 or more) on the curve. Since the KM curves. generally have a lot of jumps, I normally pick about 10 to 20 points. Again, the zoom view will help the accuracy.



Mode: Trace curve

Window-size: 5

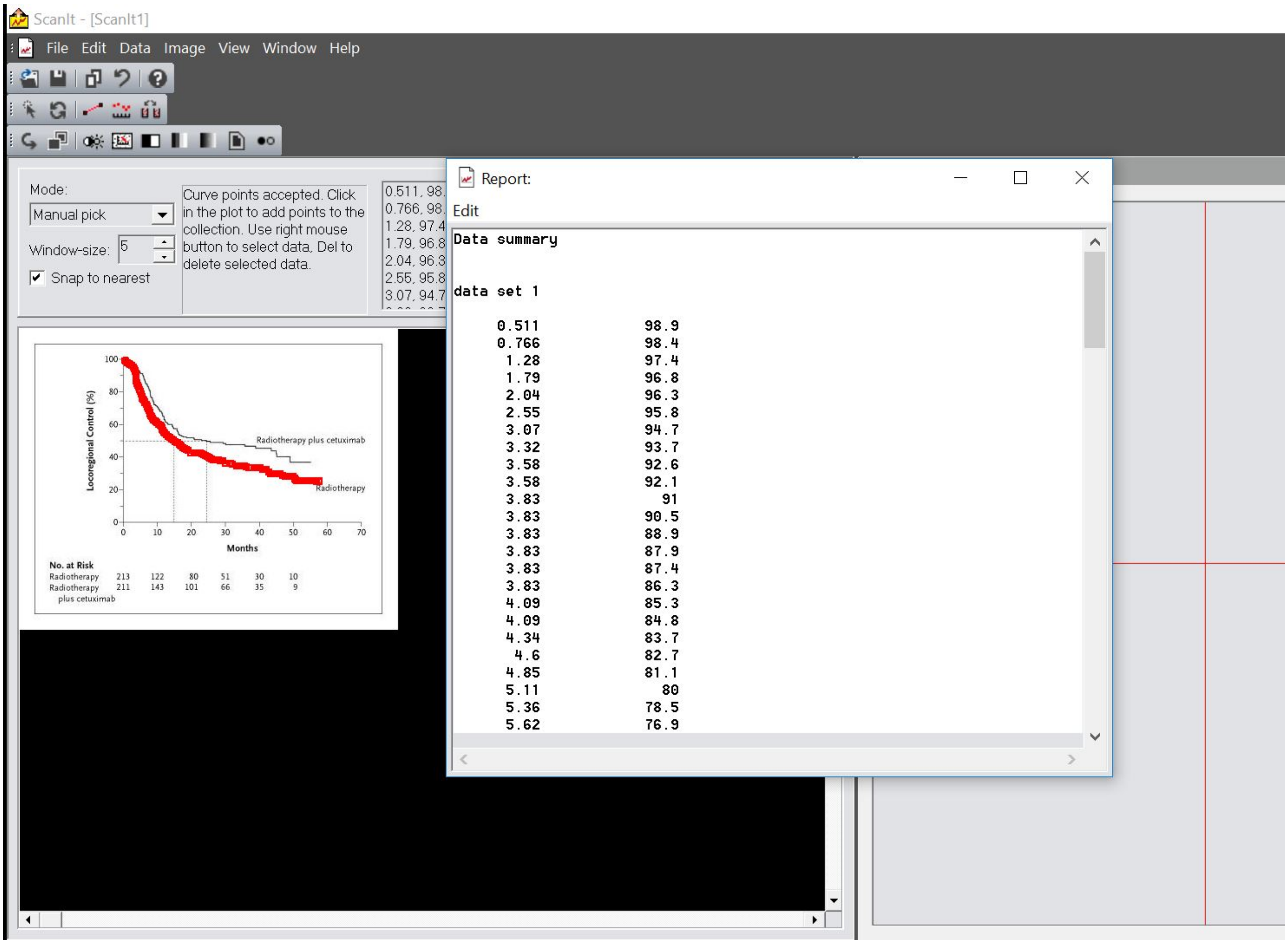
Click on parts of the curve that are not found to correct the trace, or click accept to return

```
11.5, 57.9
16.1, 49
21.7, 42.6
3.83, 88.9
5.88, 75.3
8.69, 63.7
```

Increasing X

Figure 3: Using Trace curve mode to extract the entire curves automatically
We need to pick about 10 points to tell software the jumps positions.

Then we are read to export the data. The data can be exported as .txt file.


ScanIt - [ScanIt1]
File Edit Data Image View Window Help
Report:
Edit
100
80
60
40
20
0
Locoregional Control (%)
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

Mode: Manual pick

Window-size: 5

☑ Snap to nearest

Curve points accepted. Click in the plot to add points to the collection. Use right mouse button to select data, Del to delete selected data.

```
0.511, 98.
0.766, 98.
1.28, 97.4
1.79, 96.8
2.04, 96.3
2.55, 95.8
3.07, 94.7
```

```
Data summary


data set 1

    0.511           98.9
    0.766           98.4
     1.28           97.4
     1.79           96.8
     2.04           96.3
     2.55           95.8
     3.07           94.7
     3.32           93.7
     3.58           92.6
     3.58           92.1
     3.83             91
     3.83           90.5
     3.83           88.9
     3.83           87.9
     3.83           87.4
     3.83           86.3
     4.09           85.3
     4.09           84.8
     4.34           83.7
      4.6           82.7
     4.85           81.1
     5.11             80
     5.36           78.5
     5.62           76.9
```

Figure 4: Data can be exported as .txt file

3 Using the shiny application to find the individual patients records

3.1 When the patient number at risk and time interval known

Most of published KM survival also reported five to ten time points, and the patient number at risk at those time. These can be found under the x-axis of the survival curve.

For example, the radio therapy example we used here, reported at time

$$(0, 10, 20, 30, 40, 50, 60, 70)$$

months, the patient number at risk observed were

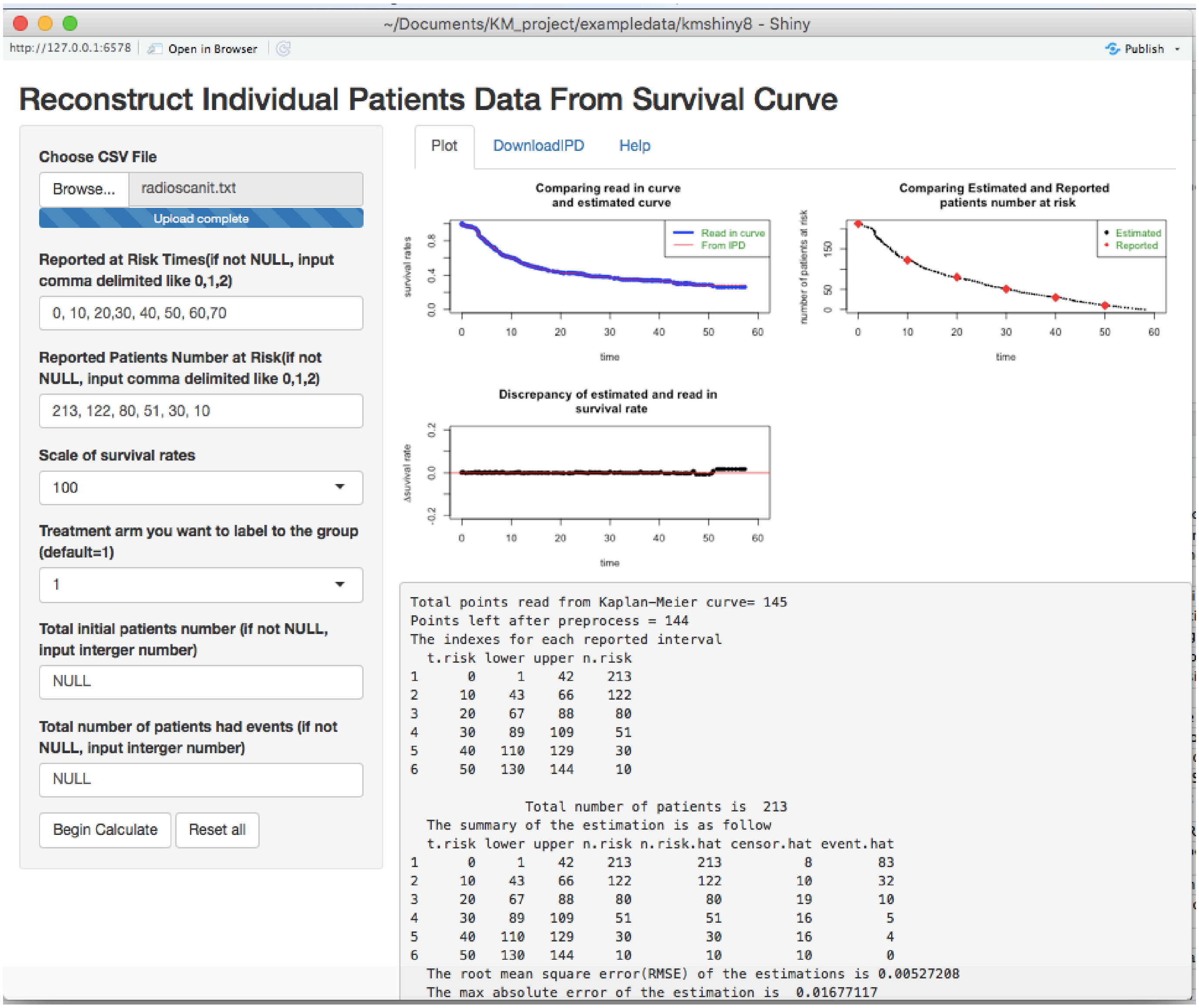
$$(213, 122, 80, 51, 30, 10)$$

.

We upload the coordinate datafile we got from ScanIt software, and then input the time at risk and patient number at risk as shown in Figure 5. Other input includs:

- Scale of survival rates. For the case the survival rates reported in the published graph showed as the percentages, we set the scale to 100. For the case using the decimal, we set the scale to 1.
- Treatment arm is the just label for the convenience of next step analysis. You can just pick the right label you want, the default is 1.
- The initial patients is just equal to the patient number at risk at time zero. You can leave it as 'NULL', since no additional information involved in this case.
- Some paper also reported total number of events. Input the number can increase the accuracy of estimation.

Once finish input data, click " Begin Calculate". Click "Reset all" to reset all the input as default status.


~/Documents/KM_project/exampledata/kmshiny8 - Shiny
http://127.0.0.1:6578
Open in Browser
Publish
Plot
DownloadIPD
Help
Comparing read in curve and estimated curve
survival rates
0.0
0.4
0.8
0
10
20
30
40
50
60
time
Read in curve
From IPD
Comparing Estimated and Reported patients number at risk
number of patients at risk
0
50
150
0
10
20
30
40
50
60
time
Estimated
Reported
Discrepancy of estimated and read in survival rate
Δsurvival rate
-0.2
0.0
0.2
0
10
20
30
40
50
60
time

Reconstruct Individual Patients Data From Survival Curve

Choose CSV File

Browse... radioscanit.txt

Upload complete

Reported at Risk Times(if not NULL, input comma delimited like 0,1,2)

0, 10, 20,30, 40, 50, 60,70

Reported Patients Number at Risk(if not NULL, input comma delimited like 0,1,2)

213, 122, 80, 51, 30, 10

Scale of survival rates

100

Treatment arm you want to label to the group (default=1)

1

Total initial patients number (if not NULL, input interger number)

NULL

Total number of patients had events (if not NULL, input interger number)

NULL

Begin Calculate | Reset all

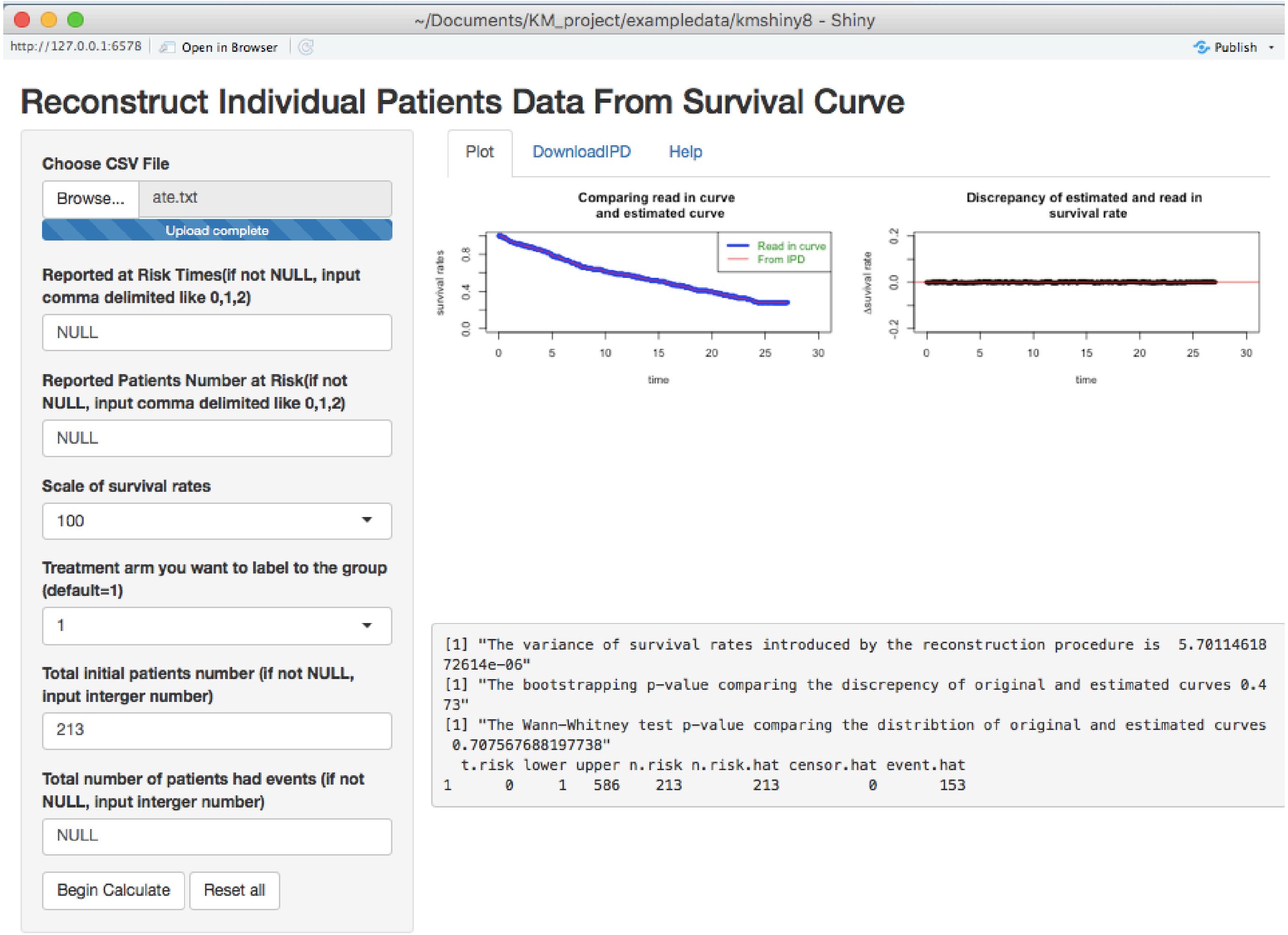
```
Total points read from Kaplan–Meier curve= 145
Points left after preprocess = 144
The indexes for each reported interval
  t.risk lower upper n.risk
1      0     1    42    213
2     10    43    66    122
3     20    67    88     80
4     30    89   109     51
5     40   110   129     30
6     50   130   144     10

              Total number of patients is  213
  The summary of the estimation is as follow
  t.risk lower upper n.risk n.risk.hat censor.hat event.hat
1      0     1    42    213        213          8        83
2     10    43    66    122        122         10        32
3     20    67    88     80         80         19        10
4     30    89   109     51         51         16         5
5     40   110   129     30         30         16         4
6     50   130   144     10         10         10         0
  The root mean square error(RMSE) of the estimations is 0.00527208
  The max absolute error of the estimation is  0.01677117
```

Figure 5: The case when patients number at risk and time interval known

3.2 When the patient number at risk and time interval do not known

Occasionally, published survival curve do not reported the patient number at risk and time interval. In this case, you can leave the most of input as the default ”NULL” by click the ”Reset all” button. Three data must be input as shown in Figure 6

- the dataset
- Total initial patients number must be input in this case. Or we can not do any estimation.
- The scale of survival rates.



Reconstruct Individual Patients Data From Survival Curve

Choose CSV File: ate.txt

Reported at Risk Times(if not NULL, input comma delimited like 0,1,2): NULL

Reported Patients Number at Risk(if not NULL, input comma delimited like 0,1,2): NULL

Scale of survival rates: 100

Treatment arm you want to label to the group (default=1): 1

Total initial patients number (if not NULL, input interger number): 213

Total number of patients had events (if not NULL, input interger number): NULL

```
[1] "The variance of survival rates introduced by the reconstruction procedure is  5.70114618
72614e-06"
[1] "The bootstrapping p-value comparing the discrepency of original and estimated curves 0.4
73"
[1] "The Wann-Whitney test p-value comparing the distribtion of original and estimated curves
 0.707567688197738"
  t.risk lower upper n.risk n.risk.hat censor.hat event.hat
1      0     1   586    213        213          0       153
```

Figure 6: The case when patients number at risk and time interval do not know

3.3 Explain of the output and figures

We output three figures for the case patient number at risk reported, and two figures for the case patient number at risk did not reported. Those curves shows the accuracy of our estimations.

We performed bootstrapping t-test to comparing the discrepancy of the original read in survival rates and the estimated survival rates. The null hypothesis is the mean discrepancy is zero.

We performed Mann-Whitney test to comparing the distribution of the original and estimated curves. P value reported.

The total variance of survival from our method come from two source. By calculating the mean square error, we have the estimation of the sample variance from the resconstruction. And for the read in precision, we generally read in more than 500 points for each curve. Then the standard error of read in is about

$$SE_{readin} = \frac{1}{500}$$

. So the total variance of survival by the reconstruction procedure is

$$Variance = MSE_{survival} + (\frac{1}{500})^2$$

3.4 Download the reconstructed patient record

Click "DownloadIPD" button, the reconstructed patient record will be shown in a table. Each row indicates one patients record, include time, status (have event label as 1, be censored label as 0), and the treatment arm. Download the .csv file for further study.

References

[1] Patricia Guyot, A. E. Ades, Mario J.N.M. Ouwens, and Nicky J. Welton. Enhanced secondary analysis of survival data: Reconstructing the data from published kaplan-meier survival curves. *BMC Medical Research Methodology*, 12, 2012.