Package 'IPDfromKM'

November 11, 2020

Type Package

Title Map Digitized Survival Curves Back to Individual Patient Data

Version 0.1.10

Date 2020-10-20

Description An implementation to reconstruct individual patient data from Kaplan-Meier (K-M) survival curves, visualize and assess the accuracy of the reconstruction, then perform secondary analysis on the reconstructed data. We involve a simple function to extract the coordinates form the published K-M curves. The function is developed based on Poisot T. 's digitize package (2011) <doi:10.32614/RJ-2011-004> . For more complex and tangled together graphs, digitizing software, such as 'DigitizeIt' (for MAC or windows) or 'ScanIt'(for windows) can be used to get the coordinates. Additional information should also be involved to increase the accuracy, like numbers of patients at risk (often reported at 5-10 time points under the x-axis of the K-M graph), total number of patients, and total number of events. The package implements the modified iterative K-M estimation algorithm (modified-iKM) improved upon the approach proposed by Guyot (2012) <doi:10.1186/1471-2288-12-9> with some modifications.

Depends R (>= 3.6.0), ggplot2, dplyr, survival, gridExtra,readbitmap

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

NeedsCompilation no

Author Na Liu [aut, cre],
J.Jack Lee [aut, ths],
Yanhong Zhou [ctb]

Maintainer Na Liu <nliu1104@gmail.com>

Repository CRAN

Date/Publication 2020-11-11 09:50:06 UTC

R topics documented:

`getIPD` *Reconstruct individual patient data (IPD) from Scanned Kaplan-Meier(K-M) curves*

Description

After the raw dataset is processed using the `preprocess` function, we can use the `getIPD()` function to reconstruct the IPD. Here the total number of events (tot.events) is an optional input; and the treatment arm can be arbitrarily assigned to label the patients' treatment group (Typically, 0 for the control group, and 1 for the treatment group).

The output is the reconstructed IPD in the form of a three-column table (i.e.,time, patient status, and treatment group ID).

In addition, in order to evaluate the accuracy of our reconstruction process, we will calculate the survival probabilities at each read-in time points based on the reconstructed IPD, then compare them with the corresponding read-in survival probabilities. The test statistics are also included in the output.

Usage

```
getIPD(prep,armID=1,tot.events=NULL)
```

Arguments

`prep`	the class object returned from the `preprocess()` function.
`armID`	the arbitrary lable used as the group indicator for the reconstructed IPD. Typically 0 for the control group and 1 for the treatment group.
`tot.events`	the total number of events. This may not be available for some published curves, thus this input is optional.

Value

`getIPD()` returns a list object, including four items as follows.

IPD: the estimated individual patient in a three-column table (i.e.time, status, and treatment group indicator).

Points: the data frame shows estimations of parameters at each read-in time points.

riskmat: the data frame shows index of read-in points within each time interval, as well as the estimated numbers of censored patients and events within each time interval.

kstest: the test statistics and p value of Kolmogorov-Smirnov test when comparing the distributions of estimated and read-in K-M curves. The null hypothesis is the read-in and estimated survival probabilities are from the same distribution.

precision: a list shows the root mean squre error(RMSE), mean absolute error and max absolute error which measure the differences between the estimated and read-in survival probabilities.

endpts: the number of patients remaining at the end of trial.

References

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; 1:9.

Examples

```
# Radiationdata$radio is a dataset exported from ScanIt software ================
radio <- Radiationdata$radio

# Load time points when the patients numbers =======
# at risk reported (i.e. trisk in month) =========
trisk <- Radiationdata$trisk

# Load the numbers of patients at risk reported (i.e. nrisk) ======
# at the time points (trisk) ========
nrisk.radio <- Radiationdata$nrisk.radio

# Use the trisk and nrisk as input for preprocess and reconstruction ============
pre_radio_1 <- preprocess(dat=Radiationdata$radio, trisk=trisk,
             nrisk=nrisk.radio,totalpts=NULL,maxy=100)
est_radio_1 <- getIPD(prep=pre_radio_1,armID=0,tot.events=NULL)

# Output include reconstructed individual patients data ==========================
head(est_radio_1$IPD)

# When trisk and nrisk were not available, then we must input ====================
# the initial number of patients   ===============================================
pre_radio_2 <- preprocess(dat=Radiationdata$radio, totalpts=213,maxy=100)
est_radio_2 <- getIPD(prep=pre_radio_2,armID=0,tot.events=NULL)

# Output include reconstructed individual patients data ==========================
head(est_radio_2$IPD)
```

getpoints *Extract the coordinates from Kaplan-Meier(K-M) curves by mouse-clicks*

Description

The `getpoints()` function extracts the coordinates from K-M curves by mouse-clicks. The K-M curves should be in the format of bitmap images(in JPEG,PNG,BMP,JPG or TIFF), and the use of .png file is highly recommended, since it can greatly shorten the processing time in R.

In addition to the image itself, the input of the `getpoints()` function includes two x-coordinates (x1 and x2) and two y-coordinates to decide the location and scale of the curve. Once the image is read into R and displayed in the plots window, firstly the user need to click on the four points on the x-axis and y-axis according to the input, and in the order of (x1,x2,y1,and y2); secondly, the user need to collect the points coordinates by mouse-clicks on the curve. To get desirable estimation, we suggest collecting 80-100 points on each curve, and including the points where the survival probability drops. The output of this function is a two-column dataset of coordinates extracted from the K-M curve.

Usage

```
getpoints(f,x1,x2,y1,y2)
```

Arguments

`f`	the bitmap image(in JPEG,PNG,BMP,JPG or TIFF formate) of the K-M curves. The input can be either the pathway to the image file, or the bitmap digital image itself.
`x1`	two points needed to decide the postion and scale of the x-axis. Here x1 is the actual x-coordinate of the right point on x-axis
`x2`	two points needed to decide the postion and scale of the x-axis. Here x2 is the actual x-coordinate of the left point on x-axis
`y1`	two points needed to decide the postion and scale of the y-axis. Here y1 is the actual y-coordinate of the lower point on y-axis
`y2`	two points needed to decide the postion and scale of the y-axis. Here y2 is the actual y-coordinate of the upper point on y-axis

Value

`getpoints()` returns a two-column dataset of coordinates extracted from a K-M curve.

References

Poisot T. The digitize package: extracting numerical data from scatterplots. The R Journal. 2011 Jun 1;3(1):25-6.

Examples

```
str(imgexp)


## Extract the coordinates from Kaplan-Meier(K-M) curves by mouse-clicks.
## The K-M curve should be in the format of bitmap images. The input f should be either
## the pathway to the image file, or the bitmap digital image itself.
## Example: extract coordinates from the sample bitmap digital image (imgexp)
plot.new()
rasterImage(imgexp, 0, 0, 1, 1)
## User need to use mouse-clicks to decide the positions of coordinates,
## and the points want to extract.
df <- getpoints(imgexp,0,60,0,100)
head(df)
## the extracted dataset df can be used to estimate IPD by other functions in the package
trisk <- Radiationdata$trisk
nrisk.radio <- Radiationdata$nrisk.radio
pre_radio <- preprocess(dat=df, trisk=trisk,
             nrisk=nrisk.radio,totalpts=NULL,maxy=100)
est_radio <- getIPD(prep=pre_radio,armID=0,tot.events=NULL)
```

---

imgexp *A bitmap digital image*

Description

The sample dataset is a bitmap digital image from a published Kaplan Meier curves.
It is the same image we used to extract the sample `Radiationdata`. We can use `getpoints()` function to extract the coordinates from the image.

Usage

```
imgexp
```

Format

numeric dataset

References

Bonner JA, Harari PM, Giralt J, Azarnia N, Shin DM, Cohen RB, Jones CU, Sur R, Raben D, Jassem J, Ove R, Kies MS, Baselga J, Youssoufian H, Amellal N, Rowinsky EK, Ang KK: Radiotherapy plus Cetuximab for Squamous-Cell Carcinoma of the Head and Neck. N Engl J Med. 2006, 354: 567-78. 10.1056/NEJMoa053422.

Examples

```
str(imgexp)
plot.new()
rasterImage(imgexp, 0, 0, 1, 1)
```

`plot.getKM` *Graph and compare the K-M curve from reconstructed IPD with the read-in coordinates*

Description

Graph the survival curve based on the reconstructed IPD, and compare it with the input coordinates. The output includes three graphs: (1) The estimated K-M curve versus read-in; (2) The estimated numbers of patients at risk versus reported; and (3) The estimated survival probabilities minus read-in survival probabilities over time.

Usage

```
## S3 method for class 'getKM'
plot(x, ...)
```

Arguments

`x`	the object returned by other functions.
`...`	ignored arguments

References

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; 1:9.

Examples

```
# Radiationdata$radio is a dataset exported from ScanIt software ================
radio <- Radiationdata$radio

# Load time points when the patients number ========
# at risk reported (i.e. trisk in month) =========
trisk <- Radiationdata$trisk

# Load the numbers of patients at risk reported (i.e. nrisk) ========
# at the time points (trisk) =============
nrisk.radio <- Radiationdata$nrisk.radio
```

```
##### Use the trisk and nrisk as input ==========
pre_radio_1 <- preprocess(dat=Radiationdata$radio, trisk=trisk,nrisk=nrisk.radio,maxy=100)
est_radio_1 <- getIPD(prep=pre_radio_1,armID=0,tot.events=NULL)
# Output include reconstructed individual patients data
head(est_radio_1$IPD)
# Plot
plot(est_radio_1)

##### When trisk and nrisk were not available, then we must input ========
##### the initial number of patients                              ========
pre_radio_2 <- preprocess(dat=Radiationdata$radio, totalpts=213,maxy=100)
est_radio_2 <- getIPD(prep=pre_radio_2,armID=0,tot.events=NULL)
# Output include reconstructed individual patients data
head(est_radio_2$IPD)
# Plot
plot (est_radio_2)
```

`preprocess` *Preprocess the read-in coordinates*

Description

Preprocess the raw coordinates into an appropriate format for reconstruct IPD. Returns include the clean dataset and a table displaying the index of read-in points within each time interval.

Usage

```
preprocess(dat,trisk=NULL,nrisk=NULL,totalpts=NULL,maxy=100)
```

Arguments

`dat`	a two-column dataset with the first column being times, and the second the survival probabilities extracted from a published K-M curve using `getpoints` function, or software such as ScanIt or DigitizeIt.
`trisk`	a vector containing risk time points (i.e., times points at which the number of patients at risk are reported). This often can be found under the x-axis of a K-M curve. The default value is NULL.
`nrisk`	a vector containing the numbers of patients at risk reported at the risk time points. This often can be found under the x-axis of a K-M curve. The default value is NULL.
`totalpts`	the initial number of patients, with a default value of NULL. However, when both trisk and nrisk are NULL, this number is required for the estimation.
`maxy`	the scale of survival probability. Set maxy=100 when the probabilities are reported in percentages (e.g., 70%). Set maxy=1 when the probabilities are reported using decimal numbers (e.g, 0.7).

Details

The `preprocess()` function process the coordinates dataset extrated from a published K-M curve using `getpoints` function, or software such as DigitizeIt or ScanIt.
In most of published Kaplan-Meier curves, we can also find several numbers of patients at risk under the x-axis. These numbers at risk, and the time reported them, should be manually input in the form of vectors (nrisk and trisk). However, when these information is not available, we can leave the "trisk" and "nrisk" parameter as "NULL". In this case, the initial number of patients "totalpts" should be input.

Sample dataset can be found in `Radiationdata`.

Value

`preprocess()` returns a list object, including four items as follows.

preprocessdat: the two-column(i.e.,time, survival) table after preprocessing

intervalIndex: a table displaying the index of read-in points within each time interval.

endpts: the number of patients remaining at the end of the trial.

inputdat: the read-in dataset.

References

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; 1:9.

Examples

```
# Radiationdata$radio is a dataset exported from ScanIt software ================
radio <- Radiationdata$radio

# Load time points when the patients number =======
# at risk reported (i.e. trisk in month) ======
trisk <- Radiationdata$trisk

# Load the numbers of patients at risk reported (i.e. nrisk) =======
# at the time points (trisk) ======
nrisk.radio <- Radiationdata$nrisk.radio

# Use the trisk and nrisk as input for preprocess and reconstruction ============
pre_radio_1 <- preprocess(dat=Radiationdata$radio, trisk=trisk,
             nrisk=nrisk.radio,totalpts=NULL,maxy=100)
est_radio_1 <- getIPD(prep=pre_radio_1,armID=0,tot.events=NULL)
```

```
# Output include reconstructed individual patients data ========================
head(est_radio_1$IPD)

# When trisk and nrisk were not available, then we must input ====================
# the initial number of patients   ===============================================
pre_radio_2 <- preprocess(dat=Radiationdata$radio, totalpts=213,maxy=100)
est_radio_2 <- getIPD(prep=pre_radio_2,armID=0,tot.events=NULL)

# Output include reconstructed individual patients data ========================
head(est_radio_2$IPD)
```

`Radiationdata` *Two-column coordinates(X,Y) extracted from published Kaplan Meier curves by ScanIt software*

Description

The datasets are extracted from a published Kaplan Meier image by ScanIt. Locoreginal control events were studied in 424 head and neck cancer patients: 213 in Radiotherapy treatment group and 211 in the Radiotherapy plus cetuximab group. There are 145 pairs of coordinates extracted from the radiation treatment arm, and 136 pairs of coordinates are extracted from the radiation plus arm. For both datasets, the first columns are the times, and the second columns are the survival probabilities in percentage. For each group, numbers of patients at risk were reported at the months of 0, 10, 20, 30, 40, and 50. Three vectors (i.e., trisk, nrisk.radio and nrisk.radioplus) record these numbers.

Usage

```
Radiationdata
```

Format

List of two dataframes (radio and radioplus) and three vectors (i.e., trisk, nrisk.radio and nrisk.radioplus)

References

Bonner JA, Harari PM, Giralt J, Azarnia N, Shin DM, Cohen RB, Jones CU, Sur R, Raben D, Jassem J, Ove R, Kies MS, Baselga J, Youssoufian H, Amellal N, Rowinsky EK, Ang KK: Radiotherapy plus Cetuximab for Squamous-Cell Carcinoma of the Head and Neck. N Engl J Med. 2006, 354: 567-78. 10.1056/NEJMoa053422.

Examples

```
## the sample datasets
radio <- Radiationdata$radio
radioplus <- Radiationdata$radioplus
trisk <- Radiationdata$trisk
nrisk_radio <- Radiationdata$nrisk.radio
nrisk_radioplus <- Radiationdata$nrisk.radioplus
plot(radio,xlab="time",ylab="survival rates",type="l",
    lty=2,col="cyan4",xlim=c(1,70),main="Curves extracted by ScanIt software")
lines(radioplus,type="l",col="red4",lty=1)
legend("topright", c("Radiotherapy", "Radiotherapy plus cetuximab"),
      col = c("cyan4","red4"),lty=c(2,1),text.col = "green4",bty = "n")
text(40,80,"Reported Hazard Ratio with 95% CI:")
text(40,75,"0.68 (0.52,0.89)")
## reconstruct the IPD from the sample dataset
pre_radio <- preprocess(dat=radio, trisk=trisk,nrisk=nrisk_radio,maxy=100)
est_radio <- getIPD(prep=pre_radio,armID=0,tot.events=NULL)
pre_radio_plus <- preprocess(dat=radioplus, trisk=trisk,nrisk=nrisk_radioplus,maxy=100)
est_radio_plus <- getIPD(prep=pre_radio_plus,armID=1,tot.events=NULL)
surv2 <- survreport(ipd1=est_radio$IPD,ipd2=est_radio_plus$IPD,arms=2,
                    interval=8,s=c(0.75,0.5,0.25),showplots=TRUE)
print(surv2)
```

`summary.getKM` *Print the summary of the IPD estimation*

Description

Generate descriptive summary for objects returned by other functions

Usage

```
## S3 method for class 'getKM'
summary(object, ...)
```

Arguments

`object`	the object returned by other functions.
`...`	ignored arguments

Details

`summary()` prints the objects returned by other functions.

References

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; 1:9.

Examples

```
# Radiationdata$radio is a dataset exported from ScanIt software ================
radio <- Radiationdata$radio

# Load time points when the patients number =========
# at risk reported (i.e. trisk in month) =========
trisk <- Radiationdata$trisk

# Load the numbers of patients at risk reported (i.e. nrisk) ======
# at the time points (trisk) ========
nrisk.radio <- Radiationdata$nrisk.radio

# Use the trisk and nrisk as input for preprocess and reconstruction ============
pre_radio_1 <- preprocess(dat=Radiationdata$radio, trisk=trisk,
             nrisk=nrisk.radio,totalpts=NULL,maxy=100)
est_radio_1 <- getIPD(prep=pre_radio_1,armID=0,tot.events=NULL)

# Output include reconstructed individual patients data ==========================
head(est_radio_1$IPD)
summary(est_radio_1)

# When trisk and nrisk were not available, then we must input ====================
# the initial number of patients ================================================
pre_radio_2 <- preprocess(dat=Radiationdata$radio, totalpts=213,maxy=100)
est_radio_2 <- getIPD(prep=pre_radio_2,armID=0,tot.events=NULL)

# Output include reconstructed individual patients data ===========================
head(est_radio_2$IPD)
summary(est_radio_2)
```

`survreport` *Survival analysis on the reconstructed IPD*

Description

Graph the Kaplan-Meier curves and the cumulative hazard curves for the reconstructed IPD (from the output of `getIPD` function). Also report the survival times with confidence intervals for a given vector of survival probabilities, as well as the landmark survival probabilities of interest.(for example, if set interval=6, the survival probability will be reported at every six months)

Usage

```
survreport(ipd1,ipd2=NULL,arms=1,interval=6,s=c(0.75,0.5,0.25),showplots=TRUE)
```

Arguments

ipd1	a three-column (i.e., time, status, and treatment indicator)table of IPD for treatment 1.
ipd2	a three-column (i.e., time, status, and treatment indicator)table of IPD for treatment 2.
arms	number of treatment group. Can be either 1 or 2.
interval	length of the time interval for which the landmark survival probabilities are of interest. The default is at every 6 months.
s	a vector with survival probabilities for which the corresponding survival times are reported. e.g., s=0.5 means that the median survival time is desired.
showplots	indicate if the survival plots are displayed or not in the plot window

Value

`survreport()` returns a list object.

References

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; 1:9.

Examples

```
### Get data from the sample dataset=======================
radio <- Radiationdata$radio
radioplus <- Radiationdata$radioplus
trisk <- Radiationdata$trisk
nrisk_radio <- Radiationdata$nrisk.radio
nrisk_radioplus <- Radiationdata$nrisk.radioplus
### Estimate the IPD for the Radiotherapy treatment group ====================
pre_radio <- preprocess(dat=radio, trisk=trisk,nrisk=nrisk_radio,maxy=100)
est_radio <- getIPD(prep=pre_radio,armID=0,tot.events=NULL)
### Estimate the IPD for the Radiotherapy plus treatment group ====================
pre_radio_plus <- preprocess(dat=radioplus, trisk=trisk,nrisk=nrisk_radioplus,maxy=100)
est_radio_plus <- getIPD(prep=pre_radio_plus,armID=1,tot.events=NULL)
### survival report for one arm ===================
surv1 <- survreport(ipd1=est_radio$IPD,arms=1,interval=6,s=c(0.75,0.5,0.25),showplots=FALSE)
print(surv1)
surv1 <- survreport(ipd1=est_radio$IPD,arms=1,interval=10,s=seq(0,1,0.2),showplots=TRUE)
print(surv1)
### survival report for two arms ===================
surv2 <- survreport(ipd1=est_radio$IPD,ipd2=est_radio_plus$IPD,arms=2,
                    interval=8,s=c(0.75,0.5,0.25),showplots=TRUE)
```

```
print(surv2)
```

Index