Example - Run CRIME_PISE

0. Dependency

SHAP analysis will be performed using `shapper`. Please follow the instructions in the following website to install `shapper`: https://github.com/ModelOriented/shapper

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(survival)
library(prodlim)
library(ggplot2)
library(randomForestSRC)
```

```
 randomForestSRC 3.2.3

 Type rfsrc.news() to see new features, changes, and bug fixes.
```

```
library(timeROC)
library(shapper)
```

Load helper functions in `funcs.R`

```
source("funcs.R")
```

```
out.path = "result"
if (!file.exists(out.path)) {
  dir.create(out.path)
}
```

1. Load Data

In the `data` folder, `data_train.csv` was the dataset used to train the Random Survival Forest (with competing risk) model. It includes the features and outcomes of 230 patients admitted between 2014 and 2020. `data_test.csv` includes 50 testing patients not included in the model training process. These patients were admitted between 2021 and 2022.

Feature Names

1. `subjID` unique, de-identified subject (aka patient) ID
2. **Outcomes**: `event` The first event of interest occured: **PISE**=first unprovoked seizure post-stroke before mortality, **Death**=post-stroke death event before PISE, **Censored**=censorship. `event_sd` Number-coded event: 0=Censored, 1=PISE, 2=Death. `time` Time (in years) from the stroke last known normal datetime to the event recorded.
3. **SeLECT**(Galovic et al. 2018) **score**: `severity` Severity of stroke: 0 represents admission NIHSS ≤ 3, 1 represents admission NIHSS $\in [4-10]$. 2 represents admission NIHSS ≥ 11. `large_artery`, 1 represents large artery atherosclerosis etiology. `early_seizure` 3 represents the occurance of clinical seizure ≤ 7 days post-stroke. `cortical_involvement` 2 represents the involvement of cortex. `mca_involvement` 1 represents the involvement of MCA territory. `select_score` Total SeLECT score.
4. `age` Age at stroke onset
5. `cardioembolism` Cardioembolism etiology
6. `pre_morbid` Pre-stroke modified Rankin Scale score
7. `afib` History of atrial fibrillation prior to the stroke onset
8. `aca` Involvement of ACA territory in the final stroke infarct
9. `nih_3d` NIHSS score at 3-days (72-hours) post-stroke

10. `hand_volume` Manually traced final infarct volume
11. `total_ea_95` Epileptiform abnormality (EA) burden calculated using SPaRCNet(Ge et al. 2021)
12. `fft_theta_all_global` Global theta power calculated using Persyst. `rhm_theta_all_global` Global theta rhythmic activities calculated using Persyst. `fft_total_all_asym` Total power asymmetry between left and right hemispheres. `rhm_total_all_asym` Total rhythmic activity asymmetry between left and right hemispheres.

```
df.train = read.csv(file = './data/data_train.csv')
df.test = read.csv(file = './data/data_test.csv')
```

Load CRIME_PISE model (Random Survival Forest with Competing Risk).

```
load("CRIME_PISE.RData")
```

2. Make Prediction

Given a set of feature input, `CRIME_PISE` would predict the risk score for PISE - `PISE_score` (i.e., PISE before PISE) and Death `Death_score` (i.e., death before PISE).

```
# Training out-of-bag prediction
df.train$PISE_score = CRIME_PISE$predicted.oob[,1]
df.train$Death_score = CRIME_PISE$predicted.oob[,2]
```

```
# Testing prediction
df.test$PISE_score = predict(CRIME_PISE, df.test)$predicted[,1]
df.test$Death_score = predict(CRIME_PISE, df.test)$predicted[,2]
```

Patients were stratified into the 2x2 groups based on whether the predicted out-of-bag PISE (or Death) risk score is above (or below) the **training cohort median** (i.e., `t.PISE`, `t.Death`).

1. **Event-Free**: below-median for both predicted PISE and Death risk scores.
2. **PISE**: above-median for predicted PISE and below-median for predicted Death risk scores.
3. **Death**: above-median for predicted Death and below-median for predicted PISE risk scores.
4. PISE or Death: above-median for both predicted PISE and Death risk scores. To avoid ambiguity in label assignment, patients in this group were further classified as "PISE" if the predicted PISE risk score was greater than that of Death, and vice versa.

```
t.PISE = quantile(df.train$PISE_score, 0.5)
t.Death = quantile(df.train$Death_score, 0.5)

# predicting the labels
df.test$pred_label = crime.predict(
  t.PISE = t.PISE,
  t.Death = t.Death,
  PISE.score = df.test$PISE_score,
  Death.score = df.test$Death_score
)
```

Save the predicted score and label for each patient, together with their features into `data_test_pred.csv` file in the result folder.

```
write.csv(df.test, paste(out.path, "data_test_pred.csv", sep="/"))
```

3. Estimate Individual Patient CIF

The prediction of cumulative incidence function (CIF) of PISE and Death outcomes overtime, as well as feature contribution to the outcome prediction, would allow users to make final predictions in a more flexible manner, as opposed to the threshold method implemented in Section 2.

Here, we take patient in the first row of the `data_test.csv` for example:

```
i = 1 # first patient
```

We can first look at the actual vs predicted label of the target patient:

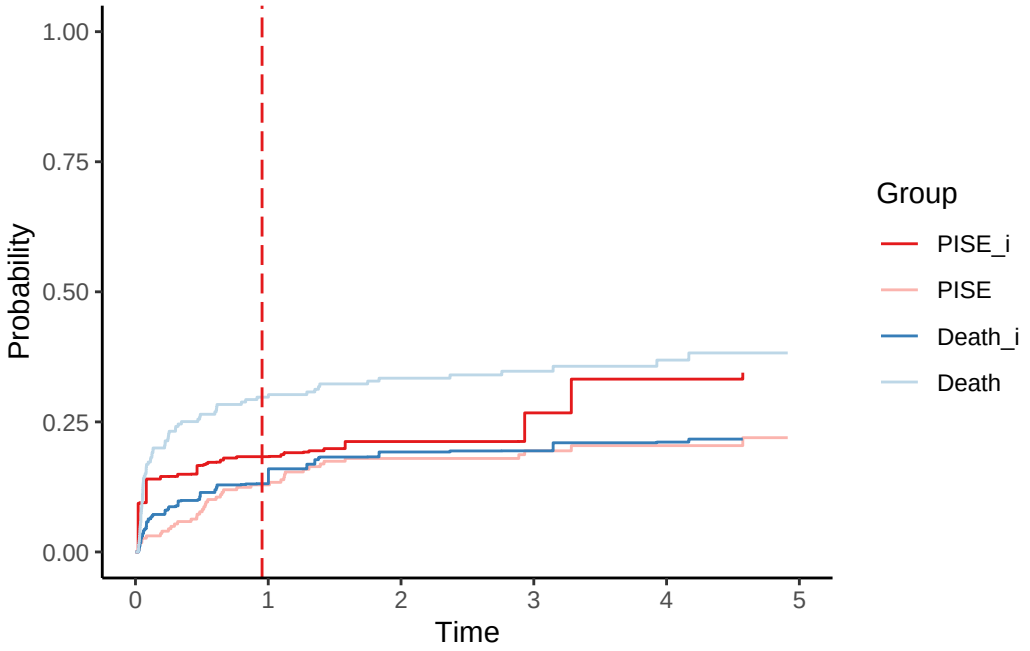
```
lab_i = df.test$event[i]
time_i = df.test$time[i]
plab_i = df.test$pred_label[i]

obs_i = df.test[i, CRIME_PISE$xvar.names]
pred_i = predict(CRIME_PISE, obs_i)

print(paste(
  "The patient was ", lab_i,
  " at ", round(time_i, 2), " years post-stroke.",
  " The patient was predicted to be ", plab_i, ".", sep=""))
```

```
[1] "The patient was PISE at 0.95 years post-stroke. The patient was predicted to be PISE."
```

`PISE` (or `Death`) denotes the population derived Aalen-Johansenestimates of CIF in the **training cohort**. These two curves were for reference only.

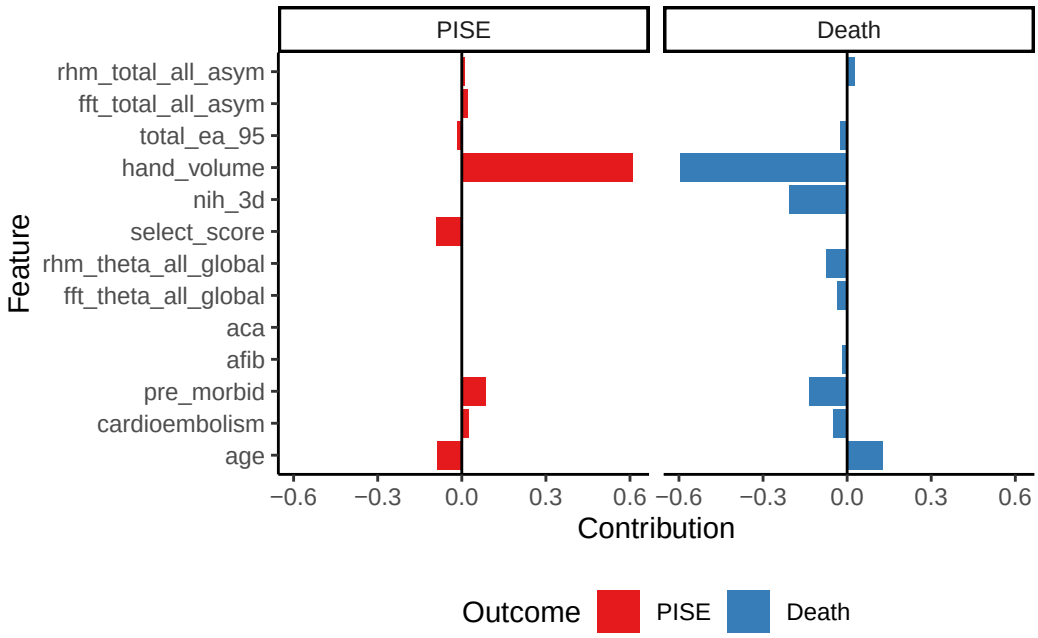
`PISE_i` (or `Death_i`) denotes the model predicted CIFs for the **target patient**.

```
plot_individual_cif(pred_i, lab_i, time_i, df.train$time, df.train$event)
```



For this patient, `hand_volume` (i.e., infarct volume) is the major positive contributor to the PISE risk.

```
shap.df = make_shap_df(mdl=CRIME_PISE, obs_i=obs_i)
plot_individual_shap(shap.df)
```



4. Population-level Evaluation

We also evaluate the model performance at the 1st-year post-stroke in the entire testing dataset. Patients without >1 year follow-up will be excluded.

1. `AUC_PISE` (or `AUC_Death`) First-year AUC for the PISE (or Death) outcome.
2. `sensitivity`, `specificity`
3. `ppv` Positive predictive value. `npv` Negative predictive value.
4. `n_pos` The number of positive predictions made by the model. `n_pos_PISE` The number of true PISE amongst model-predicted positive cases. `n_pos_Death` The number of patients who died before PISE amongst model-predicted positive cases. `n_pos_EventFree` The number of patients without any event (during the 1st-year post-stroke) amongst model-predicted positive cases.

```
eval.metrics = crime.evaluate(
  yr=1, # evaluate at 1st year post-stroke
  t=df.test$time,
  y=df.test$event,
  yhat=df.test$pred_label,
  PISE.score=df.test$PISE_score,
```

```
    Death.score=df.test$Death_score
  )

  t(data.frame(eval.metrics))
```

```
                     [,1]
AUC_PISE        0.7467903
AUC_Death       0.7597528
sensitivity     0.7000000
specificity     0.7368421
ppv             0.4117647
npv             0.9032258
n_pos          17.0000000
n_pos_PISE      7.0000000
n_pos_Death     4.0000000
n_pos_EventFree 6.0000000
```

```
  write.csv(
    t(data.frame(eval.metrics)),
    paste(out.path, "data_test_eval.csv", sep="/")
  )
```


Galovic, Marian, Nico Döhler, Barbara Erdélyi-Canavese, Ansgar Felbecker, Philip Siebel, Julian Conrad, Stefan Evers, et al. 2018. "Prediction of Late Seizures After Ischaemic Stroke with a Novel Prognostic Model (the SeLECT Score): A Multivariable Prediction Model Development and Validation Study." *The Lancet Neurology* 17 (2): 143–52. https://doi.org/10.1016/s1474-4422(17)30404-0.

Ge, Wendong, Jin Jing, Sungtae An, Aline Herlopian, Marcus Ng, Aaron F. Struck, Brian Appavu, et al. 2021. "Deep Active Learning for Interictal Ictal Injury Continuum EEG Patterns." *Journal of Neuroscience Methods* 351 (March): 108966. https://doi.org/10.1016/j.jneumeth.2020.108966.