# Guideline for codes

## Restricted mean survival time in cluster randomized trials with a small number of clusters : Improving variance estimation of the pseudo-values regression

Floriane Le Vilain–Abraham, Solène Desmée, Jennifer A Thompson,
Elsa Tavernier, Etienne Dantan, Agnès Caille

## Contents

# 0 Introduction

The code is divided into three folders corresponding to 3 steps:

1. The generation of the datasets
2. The statistical analysis
3. The estimation of the performance measures

The following sections describe how to used the R code for each step.

# 1 Generation of the datasets

This folder is divided in two folders: one for the scenarios under proportional-hazards (PH) assumption and on for the non-proportional hazards (NPH) assumption. Each folder contains 4 R files that could be used to generate data, either under PH or NPH assumptions.

The `main_data.R` file is R codes to generate 1000 simulated datasets of a prefined scenario. The necessary packages are: `doParalell` (line 12) and `doRNG` (line 13). The necessary R functions, saved in a R file with the same name, are loaded in the main script (`main_data.R`):

- `sim_data` (line 18): simulates a dataset and save it
- `generate_data` (line 19): generates one dataset
- `generate_cluster` (line 20): generates time-to-event data for one cluster

Before running, you have to specify in the main file (`main_data.R`):

- The directory where you saved all the R files with the necessary R functions detailed previously (line 17).
- The parameters of the scenario (line 25 in the data frame `table_parameter`).
- The directory where the 1000 simulated datasets will be saved (line 40). The 1000 datasets will be saved in a same automatically created folder. Table 1 described the content of one simulated dataset.
- The number of cores for the parallelisation (line 53).

Table 1: *Content of one simulated dataset*

| **Colnames** | **Variables** |
|---|---|
| `time` | $X_{lk}$, the observed time of the individual $l$ from the cluster $k$ |
| `arm` | arm of the clusters $k$ (= 1 for the intervention group, 0 for control group) |
| `cluster` | cluster's identifiant $k$ |
| `id_patient` | patient's identifiant $l$ |
| `status` | $\delta_{lk}$, the event indicator of the individual $l$ from the cluster $k$ ($\delta_{lk}$= 1 for death, 0 for censor) |

# 2 Statistical analysis

## 2.1 Analysis with the variance correction

All the needed functions are available in */2 - Statistical analysis/Variance corrections/*. The `main_analysis_variance.R` file is R code to analyse each dataset with the 5 variance estimators (standard, Mancl and DeRouen, Kaurmann and Carroll, Fay and Graubard and Morel *et al.*), both the normal and the Student distribution and both the independent and the exchangeable working correlation matrix. For the permutation test, see the section 2.2.

The necessary packages are: `doParalell` (line 12) and `doRNG` (line 13). The necessary R functions, saved in a R file with the same name, are loaded in the main script (`main_analysis_variance.R`):

- `sim_analysis_variance` (line 18): estimates the difference in RMST, its variance and the 95% confidence interval
- `RMST` (line 19): analyses one dataset with the pseudo-value regression
- `GEE.var.fg.R` (line 20): estimate the variance with the Fay and Graubard estimator
- `GEE.var.mbn.R` (line 21): estimate the variance with the Morel *et al.* estimator
- `GEE.var.kc.approx.R` (line 22): estimate the variance with the Kauermann and Carroll estimator

Before running, you have to specify in the main file (`main_analysis_variance.R`):

- The directory where you saved all the R files with the necessary R functions detailed previously (line 17).
- The parameters of the scenario (line 28 in the data frame `table_parameter`).
- The horizon time (line 40). Here, it is set at 365 days. All time are expressed in days.

- The directory where the 1000 simulated datasets have been saved and where the estimations of the difference in RMST, the variance and 95% confidence interval of each simulated datasets will be saved (line 46). The estimations of the difference in RMST, variance and 95% confidence interval for all the methods and for the 1000 simulated datasets will be saved in one txt file. Table 2 described the content of the txt file with all the estimations, respectively.
- The number of cores for the parallelisation (line 77).

You also have to specify in the `sim_analysis_variance.R` file, the directory where the simuled datasets have been saved (line 20).

Table 2: *Content of the txt file with the estimations of differences in RMST, variances and 95% confidence intervals for all the methods and for the 1000 simulated datasets*

| **Colnames** | **Variables** |
|---|---|
| ***Parameters used to simulate the dataset*** | |
| `d` | dataset number |
| `K` | total number of clusters |
| `m` | mean cluster size |
| `cv` | coefficient of variation of the cluster size |
| `HR` | true HR |
| `tau` | Kendall's tau |
| `censoring` | censoring rate |
| ***Analysis*** | |
| `matrix` | working correlation matrix |
| `method_correction` | variance estimator |
| `distribution` | distirbution of the test statistic |
| `delta.rmst` | estimation of the difference in RSMT |
| `var` | variance estimation |
| `ci.low` | lower bound of the 95% confidence interval |
| `ci.up` | upper bound of the 95% confidence interval |
| `t_star` | horizon time of the analysis |

## 2.2 Analysis with the permutation test

All the needed functions are available in the *Permutation test* folder (*/2 - Statistical analysis/Permutation test/*.
The `main_analysis_permutation.R` files is the R code to estimate the permutation-based confidence interval for each simulated dataset. The necessary packages are: `doParalell` (line 12) and `doRNG` (line 13). The necessary R functions, saved in a R file with the same name, are loaded in the main file (`main_analysis_permutation.R`):

- `sim_analysis_permutation` (line 18): estimates the permutation-based confidence interval for the pseudo-values regression-based methods
- `ci` (line 19): estimates the permutation-based confidence interval for one dataset
- `ci_permutation` (line 20): constructs one permutation-based confidence bound (lower or upper bound)
- `initialisation` (line 21): initialises the search procedure
- `initialisation_ci` (line 22): fonction to initialise the confidence bounds of the search procedure (estimates one permuted $\delta_{offset}$ from the permutation test $H_0 : \beta_1 = \hat{\beta}_1$)
- `allocation` (line 23): permutes the intervention allocation
- `update_bound` (line 24): updates bound using Robbins-Monro search procedure

Before running, you have to specify in the `main_analysis_permutation.R` file:

- The directory where you saved all the R files with the necessary R functions detailed previously (line 17).
- The parameters of the scenario (line 30 in the data frame `table_parameter`).
- The horizon time (line 42). Here, it is set at 365 days. All time are expressed in days.
- The directory where the estimations of the confidence intervals will be saved (line 48). The estimations of the confidence intervals for the 1000 simulated datasets will be saved in one txt file. Table 3 describes this txt file.
- The number of cores for the parallelisation (line 75).

You also have to specify in the `sim_analysis_permutation.R` file, the directory where the simuled datasets have been saved (line 21).

Table 3: *Content of txt file with the estimations of the permutation-based confidence intervals for pseudo-values regression for the 1000 simulated datasets*

| **Colnames** | **Variables** |
|---|---|
| | ***Parameters used to simulate the dataset*** |
| `d` | dataset number |
| `K` | total number of clusters |
| `m` | mean cluster size |
| `cv` | coefficient of variation of the cluster size |
| `HR` | true HR |
| `tau` | Kendall's tau |
| `censoring` | censoring rate |
| | ***Analysis*** |
| `matrix` | type of working correlation matrix used for the analysis (`"ind"` = independent, `"exch"` = exchangeable) |
| `ci_low` | permutation-based lower confidence bound |
| `ci_up` | permutation-based upper confidence bound |

# 3 Estimation of the performance measures

## 3.1 Variance estimators

This folder is divided in two folders: one for the scenarios under PH and NPH assumptions. Each folder contains 5 R files that could be used estimate the performance measures, either under PH or NPH assumptions.

All the needed functions are available in */3 - Estimation of the performance measures/Variance corrections/*. The `main_pm_variance.R` files are R codes to estimate the performance measures with the 5 variance estimators (standard, Mancl and DeRouen, Kaurmann and Carroll, Fay and Graubard and Morel *et al.*), both the normal and the Student distribution and both the independent and the exchangeable working correlation matrix. For the permutation test, see the section 3.2.

No package is necessary. The necessary R functions, saved in a R file with the same name, are loaded in the main script `main_pm_variance.R`:

- `pm_estimation_variance` (line 16): estimates the performance measures for the 5 methods
- `performance_measures` (line 17): estimates the performance measure for one methode
- `true_rmst_difference` (line 18): computes the true difference in RSMT
- `survival_function` (line 19): true survival function

Before running, you have to specify:

- The directory where you saved all the R files with the necessary R functions detailed previously (line 15).
- The directory and the name of the txt file where the estimations of the difference in RMST, variance and 95% confidence interval for all the methods and for the 1000 datasets have been saved in the step 2 (line 25).

When you run the main R script, the performance measures are summurized in a data frame. The content of this data frame is detailed in Table 4.

Table 4: *Content of the data frame obtained when running the main scripts* `main_pm_variance.R`

| Colnames | Variables |
|---|---|
| `tstar` | horizon time for the analysis |
| `K` | total number of clusters |
| `m` | mean cluster size |
| `cv` | coefficient of variation of the cluster sizes |
| `HR` | true HR |
| `tau` | Kendall's tau |
| `censoring` | censoring rate |
| `relative.error` | relative error |
| `coverage` | coverage rate |
| `rejection.rate` | type I error rate (absence of intervention effect) or power (in presence of intervention effect) |
| `D` | number of simulation iterations that converge |

### 3.2 Permutation test for the pseudo-values regression-based method: Estimation of the type I error rate and the coverage rate

All the needed functions are available in */3 - Estimation of the performance measures/Permutation test/*.

The `main_pm_permutation.R` file is R codes to estimate the coverage rate for the pseudo-values regression-based method. No package is necessary. The necessary R functions, saved in a R file with the same name, are loaded in the main script (`main_pm_permutation.R` ):

- `pm_estimation_permutation` (line 13): estimates the performance measures for the 5 methods
- `performance_measures` (line 14): estimates the performance measure for one methode
- `true_rmst_difference` (line 15): computes the true difference in RSMT
- `survival_function` (line 16): true survival function

Before running, you have to specify:

- The directory where you saved all the R files with the necessary R functions detailed previously (line 12).
- The directory and the name of the txt file where the estimations of the 95% confidence intervals for the 1000 datasets have been saved in the step 2 (line 22).

When you run the main R script, the coverage rates are summurized in a data frame. The content of this data frame is detailed in Table 5.

Table 5: *Content of the data frame obtained when running the main script* `main_pm_permutation.R`

| **Colnames** | **Variables** |
|---|---|
| `tstar` | horizon time for the analysis |
| `K` | total number of clusters |
| `m` | mean cluster size |
| `cv` | coefficient of variation of the cluster sizes |
| `HR` | true HR |
| `tau` | Kendall's tau |
| `censoring` | censoring rate |
| `matrix` | type of working correlation matrix used for the analysis |
| `rejection.rate` | type I error rate (absence of intervention effect) or power (in presence of intervention effect) |
| `coverage` | coverage rate |