

Package ‘RESIDE’

October 1, 2026

Title Rapid Easy Synthesis to Inform Data Extraction

Version 0.4.0

Description Assists researchers with planning analysis prior to obtaining data from Trusted Research Environments (TREs), also known as safe havens. Marginal distributions of one or more related data frames can be exported from a TRE and imported elsewhere, where data can be synthesised from them, with or without user specified correlations, by sampling from a multivariate cumulative distribution (copula). The International Stroke Trial (IST) is included as an example dataset under the ODC-By licence, Sandercock et al. (2011) <doi:10.7488/ds/104>, Sandercock et al. (2011) <doi:10.1186/1745-6215-12-101>.

License GPL (>= 3)

Encoding UTF-8

VignetteBuilder knitr

Suggests testthat (>= 3.0.0), lifecycle, knitr, rmarkdown, DT, survival, pharmaversesdtm

Depends R (>= 4.1.0)

Imports dplyr, magrittr, bestNormalize, RDP, methods, tibble, simstudy

LazyData true

Config/testthat/edition 3

URL <https://hehta.github.io/RESIDE/>

Config/roxygen2/version 8.1.0

NeedsCompilation no

Author Ryan Field [aut, cre] (ORCID: <<https://orcid.org/0000-0002-4424-9890>>),
David McAllister [aut] (ORCID: <<https://orcid.org/0000-0003-3550-1764>>),
Claudia Geue [ctb] (ORCID: <<https://orcid.org/0000-0003-2243-0733>>)

Maintainer Ryan Field <ryan.field@glasgow.ac.uk>

Repository CRAN

Date/Publication 2026-09-30 22:30:02 UTC

Contents

RESIDE-package	2
correlation	3
export_empty_cor_matrix	4
export_marginal_distributions	4
get_marginal_distributions	6
import_cor_matrix	7
import_marginal_distributions	8
IST	9
print.RESIDE	13
print.summary.RESIDE	14
summary.RESIDE	15
synthesise_data	16
Index	18

RESIDE-package	<i>RESIDE: Rapid Easy Synthesis to Inform Data Extraction</i>
----------------	---

Description

Assists researchers with planning analysis prior to obtaining data from Trusted Research Environments (TREs), also known as safe havens. Marginal distributions of one or more related data frames can be exported from a TRE and imported elsewhere, where data can be synthesised from them, with or without user specified correlations, by sampling from a multivariate cumulative distribution (copula). The International Stroke Trial (IST) is included as an example dataset under the ODC-By licence, Sandercock et al. (2011) [doi:10.7488/ds/104](https://doi.org/10.7488/ds/104), Sandercock et al. (2011) [doi:10.1186/1745621512101](https://doi.org/10.1186/1745621512101).

Details

[Experimental]

The RESIDE Package

This work was supported by the UKRI Strength in Places Fund (SIPF) Competition, #' project number 107140. The project title is SIPF The Living Laboratory driving economic growth in Glasgow through real world implementation of precision medicine.

Author(s)

Maintainer: Ryan Field <ryan.field@glasgow.ac.uk> ([ORCID](#))

Authors:

- Ryan Field <ryan.field@glasgow.ac.uk> ([ORCID](#))
- David McAllister <david.mcallister@glasgow.ac.uk> ([ORCID](#))

Other contributors:

- Claudia Geue <claudia.geue@glasgow.ac.uk> ([ORCID](#)) [contributor]

See Also

Useful links:

- <https://hehta.github.io/RESIDE/>

correlation

Create a correlation object

Description

A helper function to create a correlation object

Usage

```
correlation(x, y, rho, ...)
```

Arguments

x	The name of the first variable
y	The name of the second variable
rho	The correlation between the two variables
...	Additional arguments to specify data frame names and factor names See details for more information on the additional arguments.

Details

This function is a helper function to create a correlation object that can be used to specify correlations between variables when synthesising data using the [synthesise_data](#) function. Additional Arguments:

- df_name: The name of the data frame containing both variables
- df_name.x: The name of the data frame containing the first variable
- df_name.y: The name of the data frame containing the second variable
- factor_name.x: The name of the factor variable for the first variable
- factor_name.y: The name of the factor variable for the second variable

Value

A list containing the correlation information

Examples

```
correlation("age", "bmi", 0.5)
```

export_empty_cor_matrix

Export an empty correlation matrix (removed)

Description

This function has been removed. Correlations should now be specified using the [correlation](#) function.

Usage

```
export_empty_cor_matrix(...)
```

Arguments

... Ignored, retained for backwards compatibility.

Details

Previously this function exported an empty correlation matrix as a csv file. Correlations are now supplied directly to [synthesise_data](#) as a list of objects created with the [correlation](#) function.

Value

No return value, always throws an error.

See Also

[correlation](#)

Examples

```
try(export_empty_cor_matrix())
```

export_marginal_distributions

Export Marginal Distributions

Description

Export the marginal distributions to CSV files

Usage

```
export_marginal_distributions(  
  marginals,  
  folder_path,  
  create_folder = FALSE,  
  force = FALSE  
)
```

Arguments

marginals	an Object of type RESIDE from get_marginal_distributions
folder_path	path to folder where to save files.
create_folder	if the folder does not exist should it be created, Default: FALSE
force	if the folder already contains marginal distribution files should they be removed, Default: FALSE

Details

Exports each of the marginal distributions to CSV files within a given folder, along with the continuous quantiles.

Value

No return value, called for exportation of files.

See Also

[get_marginal_distributions](#)

Examples

```
marginal_distributions <- get_marginal_distributions(  
  IST,  
  variables = c("SEX", "AGE", "RSBP", "RATRIAL")  
)  
export_marginal_distributions(  
  marginal_distributions,  
  folder_path = file.path(tempdir(), "marginals"),  
  create_folder = TRUE,  
  force = TRUE  
)
```

`get_marginal_distributions`*Generate Marginal Distributions for a given data frame*

Description

Generate Marginal Distributions from a given data frame with options to specify which variables to use.

Usage

```
get_marginal_distributions(  
  df,  
  subject_identifier = "",  
  variables = c(),  
  print = FALSE,  
  retype = TRUE  
)
```

Arguments

<code>df</code>	Data frame or a "list" of data frames to get the marginal distributions from
<code>subject_identifier</code>	(Optional) Subject identifier required if a list of data frames is provided, Default: ""
<code>variables</code>	(Optional) variable (columns) to select, Default: c()
<code>print</code>	Whether to print the marginal distributions to the console, Default: FALSE
<code>retype</code>	Whether to re-type the data frame, Default: TRUE

Details

A function to generate marginal distributions from a given data frame, depending on the variable type the marginals will differ, for binary variables a mean and number of missing is generated for continuous variables, they are first transformed and both mean and sd of the transformed variables are stored along with the quantile mapping for back transformation. For categorical variables, the number of each category is stored, missing values are categorise as "missing".

Value

A list of marginal distributions of an S3 RESIDE Class

See Also

[export_marginal_distributions](#)

Examples

```
marginal_distributions <- get_marginal_distributions(
  IST,
  variables = c(
    "SEX",
    "AGE",
    "ID14",
    "RSBP",
    "RATRIAL"
  )
)
```

import_cor_matrix	<i>Import a correlation matrix (removed)</i>
-------------------	--

Description

This function has been removed. Correlations should now be specified using the [correlation](#) function.

Usage

```
import_cor_matrix(...)
```

Arguments

... Ignored, retained for backwards compatibility.

Details

Previously this function imported a correlation matrix from a csv file. Correlations are now supplied directly to [synthesise_data](#) as a list of objects created with the [correlation](#) function.

Value

No return value, always throws an error.

See Also

[correlation](#)

Examples

```
try(import_cor_matrix())
```

```
import_marginal_distributions
```

Import Marginal Distributions

Description

Import the marginal distribution as exported from a Trusted Research Environment (TRE)

Usage

```
import_marginal_distributions(  
  folder_path = ".",  
  binary_variables_file = "",  
  categorical_variables_file = "",  
  continuous_variables_file = "",  
  summary_file = "summary.csv"  
)
```

Arguments

`folder_path` Where the marginal distribution files are located, Default: '.' see details.

`binary_variables_file`
 filename for the binary_variables file, Default: "" see details.

`categorical_variables_file`
 filename for the categorical variables file , Default: "" see details.

`continuous_variables_file`
 filename for the continuous variables file, Default: "" see details.

`summary_file` filename for the summary file, Default: 'summary.csv' see details.

Details

This function will import marginal distributions as generated within a Trusted Research Environment (TRE) using the function [export_marginal_distributions](#). The `folder_path` allows the path of the files provided by the TRE to be imported, this will default to the current working directory. The file parameters will provide the default file names if no filenames are specified.

Value

Returns an object of a RESIDE class

See Also

[synthesise_data](#)

Examples

```
# Export marginal distributions to a temporary folder
folder_path <- file.path(tempdir(), "marginals")
export_marginal_distributions(
  get_marginal_distributions(
    IST,
    variables = c("SEX", "AGE", "RSBP", "RATRIAL")
  ),
  folder_path = folder_path,
  create_folder = TRUE,
  force = TRUE
)
# Import the marginal distributions
marginals <- import_marginal_distributions(folder_path = folder_path)
```

 IST

IST Dataset

Description

The International Stroke Trial Dataset

Usage

IST

Format

A data frame with 19435 rows and 112 columns:

AGE Randomisation data: Age in years

CMPLASP Other data and derived variables: Compliant for aspirin

CMPLHEP Other data and derived variables: Compliant for heparin

CNTRYNUM Other data and derived variables: Country code

COUNTRY Other data and derived variables: Abbreviated country code

DALIVE Recurrent stroke within 14 days: Discharged alive from hospital

DALIVED Recurrent stroke within 14 days: Date Discharged alive from hospital

DAP Data collected on 14 day/discharge form about treatments given in hospital: Non trial antiplatelet drug (Y/N)

DASP14 Data collected on 14 day/discharge form about treatments given in hospital: Aspirin given for 14 days or till death or discharge (Y/N)

DASPLT Data collected on 14 day/discharge form about treatments given in hospital: Discharged on long term aspirin (Y/N)

DAYLOCAL Randomisation data: Estimate of local day of week (assuming RDATE is Oxford)

- DCAA** Data collected on 14 day/discharge form about treatments given in hospital: Calcium antagonists (Y/N)
- DCAREND** Data collected on 14 day/discharge form about treatments given in hospital: Carotid surgery (Y/N)
- DDEAD** Other events within 14 days: Dead on discharge form
- DDEADC** Other events within 14 days: Cause of death (1-Initial stroke/2-Recurrent stroke (ischaemic or unknown /3-Recurrent stroke (haemorrhagic)/4-Pneumonia /5-Coronary heart disease/6-Pulmonary embolism /7-Other vascular or unknown/8-Non-vascular/0-unknown)
- DDEADD** Date of dead on discharge form (yyyy/mm/dd); NOTE: this death is not necessarily within 14 days of randomisation
- DDEADX** Other events within 14 days: Comment on death
- DDIAGHA** Final diagnosis of initial event: Haemorrhagic stroke
- DDIAGISC** Final diagnosis of initial event: Ischaemic stroke
- DDIAGUN** Final diagnosis of initial event: Indeterminate stroke
- DEAD1** Indicator variables for specific causes of death: Initial stroke
- DEAD2** Indicator variables for specific causes of death: Recurrent ischaemic/unknown stroke
- DEAD3** Indicator variables for specific causes of death: Recurrent haemorrhagic stroke
- DEAD4** Indicator variables for specific causes of death: Pneumonia
- DEAD5** Indicator variables for specific causes of death: Coronary heart disease
- DEAD6** Indicator variables for specific causes of death: Pulmonary embolism
- DEAD7** Indicator variables for specific causes of death: Other vascular or unknown
- DEAD8** Indicator variables for specific causes of death: Non vascular
- DGORM** Data collected on 14 day/discharge form about treatments given in hospital: Glycerol or manitol (Y/N)
- DHAEMD** Data collected on 14 day/discharge form about treatments given in hospital: Haemodilution (Y/N)
- DHH14** Data collected on 14 day/discharge form about treatments given in hospital: Medium dose heparin given for 14 days etc in pilot (combine with above)
- DIED** Other data and derived variables: Indicator variable for death (1=died; 0=did not die)
- DIVH** Data collected on 14 day/discharge form about treatments given in hospital: Non trial intravenous heparin (Y/N)
- DLH14** Data collected on 14 day/discharge form about treatments given in hospital: Low dose heparin given for 14 days or till death/discharge (Y/N)
- DMAJNCH** Data collected on 14 day/discharge form about treatments given in hospital: Major non-cerebral haemorrhage (Y/N)
- DMAJNCHD** Data collected on 14 day/discharge form about treatments given in hospital: Date of Major non-cerebral haemorrhage (yyyy/mm/dd)
- DMAJNCHX** Data collected on 14 day/discharge form about treatments given in hospital: Comment of Major non-cerebral haemorrhage
- DMH14** Data collected on 14 day/discharge form about treatments given in hospital: Date of Major non-cerebral haemorrhage (yyyy/mm/dd)

DNOSTRK Final diagnosis of initial event: Not a stroke

DNOSTRKX Final diagnosis of initial event: Comment on Not a stroke

DOAC Data collected on 14 day/discharge form about treatments given in hospital: Other anticoagulants (Y/N)

DPE Other events within 14 days: Pulmonary embolism

DPED Other events within 14 days: Date of Pulmonary embolism (yyyy/mm/dd)

DPLACE Other events within 14 days: Discharge destination (A-Home /B-Relatives home /C-Residential care /D-Nursing home /E-Other hospital departments /U-Unknown)

DRSH Recurrent stroke within 14 days: Haemorrhagic stroke

DRSHD Recurrent stroke within 14 days: Date of Haemorrhagic stroke (yyyy/mm/dd)

DRSISC Recurrent stroke within 14 days: Ischaemic recurrent stroke

DRSISCD Recurrent stroke within 14 days: Date of Ischaemic recurrent stroke (yyyy/mm/dd)

DRSUNK Recurrent stroke within 14 days: Unknown type

DRSUNKD Recurrent stroke within 14 days: Date of Unknown type (yyyy/mm/dd)

DSCH Data collected on 14 day/discharge form about treatments given in hospital: Non trial subcutaneous heparin (Y/N)

DSIDE Data collected on 14 day/discharge form about treatments given in hospital: Other side effect (Y/N)

DSIDED Data collected on 14 day/discharge form about treatments given in hospital: Date of Other side effect

DSIDEX Data collected on 14 day/discharge form about treatments given in hospital: Comment of Other side effect

DSTER Data collected on 14 day/discharge form about treatments given in hospital: Steroids (Y/N)

DTHROMB Data collected on 14 day/discharge form about treatments given in hospital: Thrombolysis (Y/N)

DVT14 Indicator variables for specific causes of death: Indicator of deep vein thrombosis on discharge form

EXPD14 Other data and derived variables: Predicted probability of death at 14 days

EXPD6 Other data and derived variables: Predicted probability of death at 6 month

EXPDD Other data and derived variables: Predicted probability of death/dependence at 6 month

FAP Data collected at 6 months: On antiplatelet drugs

FDEAD Data collected at 6 months: Dead at six month follow-up (Y/N)

FDEADC Data collected at 6 months: Cause of death (1-Initial stroke /2-Recurrent stroke (ischaemic or unknown) /3-Recurrent stroke (haemorrhagic) /4-Pneumonia /5-Coronary heart disease /6-Pulmonary embolism /7-Other vascular or unknown /8-Non-vascular /0-unknown)

FDEADD Data collected at 6 months: Date of death; NOTE: this death is not necessarily within 6 months of randomisation

FDEADX Data collected at 6 months: Comment on death

FDENNIS Data collected at 6 months: Dependent at 6 month follow-up (Y/N)

FLASTD Data collected at 6 months: Date of last contact

FOAC Data collected at 6 months: On anticoagulants

FPLACE Data collected at 6 months: Place of residence at 6 month follow-up (A-Home /B-Relatives home /C-Residential care /D-Nursing home /E-Other hospital departments /U-Unknown)

FRECOVER Data collected at 6 months: Fully recovered at 6 month follow-up (Y/N)

FU1_COMP Other data and derived variables: Date discharge form completed

FU1_RECD Other data and derived variables: Date discharge form received

FU2_DONE Other data and derived variables: Date 6 month follow-up done

H14 Indicator variables for specific causes of death: Cerebral bleed/haemorrhagic stroke within 14 days; this is slightly wider definition than DRSH and is used for analysis of cerebral bleeds

HOSPNUM Randomisation data: Hospital number

HOURLocal Randomisation data: Local time – hours

HTI14 Indicator variables for specific causes of death: Indicator of haemorrhagic transformation within 14 days

ID14 Other data and derived variables: Indicator of death at 14 days

ISC14 Indicator variables for specific causes of death: Indicator of ischaemic stroke within 14 days

MINLOCAL Randomisation data: Local time – minutes

NCB14 Indicator variables for specific causes of death: Indicator of any non-cerebral bleed within 14 days

NCCODE Other data and derived variables: Coding of compliance (see Table 3) [doi:10.1186/174562151324](https://doi.org/10.1186/174562151324)

NK14 Indicator variables for specific causes of death: Indicator of indeterminate stroke within 14 days

OCCODE Other data and derived variables: Six month outcome (1-dead /2-dependent /3-not recovered /4-recovered /8 or 9 – missing status)

ONDRUG Data collected on 14 day/discharge form about treatments given in hospital: Estimate of time in days on trial treatment

PE14 Indicator variables for specific causes of death: Indicator of pulmonary embolism within 14 days

RASP3 Randomisation data: Aspirin within 3 days prior to randomisation (Y/N)

RATRIAL Randomisation data: Atrial fibrillation (Y/N); not coded for pilot phase - 984 patients

RCONSC Randomisation data: Conscious state at randomisation (F - fully alert, D - drowsy, U - unconscious)

RCT Randomisation data: CT before randomisation (Y/N)

RDATE Randomisation data: Date of randomisation

RDEF1 Randomisation data: Face deficit (Y/N/C=can't assess)

RDEF2 Randomisation data: Arm/hand deficit (Y/N/C=can't assess)

RDEF3 Randomisation data: Leg/foot deficit (Y/N/C=can't assess)

RDEF4 Randomisation data: Dysphasia (Y/N/C=can't assess)

RDEF5 Randomisation data: Hemianopia (Y/N/C=can't assess)

RDEF6 Randomisation data: Visuospatial disorder (Y/N/C=can't assess)
RDEF7 Randomisation data: Brainstem/cerebellar signs (Y/N/C=can't assess)
RDEF8 Randomisation data: Other deficit (Y/N/C=can't assess)
RDELAY Randomisation data: Delay between stroke and randomisation in hours
RHEP24 Randomisation data: Heparin within 24 hours prior to randomisation (Y/N)
RSBP Randomisation data: Systolic blood pressure at randomisation (mmHg)
RSLEEP Randomisation data: Symptoms noted on waking (Y/N)
RVISINF Randomisation data: Infarct visible on CT (Y/N)
RXASP Randomisation data: Trial aspirin allocated (Y/N)
RXHEP Randomisation data: Trial heparin allocated (M/L/N) \[M is coded as H=high in pilot\
SET14D Other data and derived variables: Know to be dead or alive at 14 days (1=Yes, 0=No);
 this does not necessarily mean that we know outcome at 6 monts – see OCCODE for this
SEX Randomisation data: M=male; F=female
STRK14 Indicator variables for specific causes of death: Indicator of any stroke within 14 days
STYPE Randomisation data: Stroke subtype (TACS/PACS/POCS/LACS/other)
TD Other data and derived variables: Time of death or censoring in days
TRAN14 Indicator variables for specific causes of death: Indicator of major non-cerebral bleed
 within 14 days ...

Details

Obtained from Sandercock, Peter; Niewada, Maciej; Czlonkowska, Anna. (2011). International Stroke Trial database (version 2), [dataset]. University of Edinburgh. Department of Clinical Neurosciences. doi:10.7488/ds/104 Under ODC-by licence

Author(s)

Sandercock P et al. <Peter.Sandercock@ed.ac.uk>

References

doi:10.7488/ds/104

print.RESIDE

print.RESIDE

Description

S3 override for print RESIDE

Usage

```
## S3 method for class 'RESIDE'
print(x, ...)
```

Arguments

`x` an object of class `RESIDE`

`...` Other parameters, `full = TRUE` prints every category of the categorical variables, Default: `FALSE`

Details

S3 Override for `RESIDE` Class, prints the overall summary followed by the marginal distributions of each data frame. By default categorical variables with more than 10 categories only print the 10 most common categories, use `full = TRUE` to print every category.

Value

The `RESIDE` object, invisibly. Called to print to the terminal.

Examples

```
print(
  marginal_distributions <- get_marginal_distributions(
    IST,
    variables = c(
      "SEX",
      "AGE",
      "ID14",
      "RSBP",
      "RATRIAL"
    )
  )
)
print(marginal_distributions, full = TRUE)
```

```
print.summary.RESIDE  print.summary.RESIDE
```

Description

S3 override for `print.summary.RESIDE`

Usage

```
## S3 method for class 'summary.RESIDE'
print(x, ...)
```

Arguments

`x` an object of class `summary.RESIDE`

`...` Other parameters currently none are used

Details

S3 Override for summary.RESIDE Class, prints the overall summary followed by a table summarising each data frame.

Value

The summary.RESIDE object, invisibly. Called to print to the terminal.

See Also

[summary.RESIDE](#)

summary.RESIDE	<i>summary.RESIDE</i>
----------------	-----------------------

Description

S3 override for summary RESIDE

Usage

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

Arguments

object	an object of class RESIDE
...	Other parameters currently none are used

Details

S3 Override for RESIDE Class, a higher level summary than [print.RESIDE](#). For each data frame it gives the number of rows, subjects and variables, the number of each type of variable, the number of date variables and the number of variables with missing data.

Value

An object of class summary.RESIDE, a list containing overall, a data frame of the overall summary, and data_frames, a data frame with a row for each data frame.

See Also

[print.RESIDE](#)

Examples

```
summary(
  get_marginal_distributions(
    IST,
    variables = c(
      "SEX",
      "AGE",
      "ID14",
      "RSBP",
      "RATRIAL"
    )
  )
)
```

synthesise_data

Synthesise data from marginal distributions

Description

Allows the synthesis of data from marginal distributions obtained from a Trusted Research Environment (TRE)

Usage

```
synthesise_data(marginals, correlation_matrix = NULL, correlations = NULL, ...)
```

```
synthesize_data(marginals, correlation_matrix = NULL, correlations = NULL, ...)
```

Arguments

<code>marginals</code>	an object of class <code>RESIDE</code>
<code>correlation_matrix</code>	No longer supported, use <code>correlations</code> . Default: <code>NULL</code>
<code>correlations</code>	A list of correlations created with the correlation function, Default: <code>NULL</code>
<code>...</code>	Additional parameters currently none are used.

Details

This function will synthesise a dataset from marginals imported using [import_marginal_distributions](#). By default the dataset will not contain correlations, however user specified correlations can be added using the `correlations` parameter, see [correlation](#). Categorical variables are correlated using a single category, specified with `factor_name.x` or `factor_name.y`. Correlated variables are synthesised together, one row per subject, and joined to each data frame by subject. Correlated variables therefore take a single value per subject within each data frame. It is not possible to entirely maintain the marginal distributions when specifying correlations.

Value

a data frame of simulated data, or a named list of data frames for marginals from multiple data frames.

See Also

[correlation](#)

Examples

```
marginals <- get_marginal_distributions(
  IST,
  variables = c("SEX", "AGE", "RSBP", "RATRIAL")
)
df <- synthesise_data(marginals)
df_cor <- synthesise_data(
  marginals,
  correlations = list(
    correlation("AGE", "RSBP", 0.3),
    correlation("SEX", "AGE", -0.2, factor_name.x = "M")
  )
)
```

Index

* datasets

IST, [9](#)

correlation, [3](#), [4](#), [7](#), [16](#), [17](#)

export_empty_cor_matrix, [4](#)

export_marginal_distributions, [4](#), [6](#), [8](#)

get_marginal_distributions, [5](#), [6](#)

import_cor_matrix, [7](#)

import_marginal_distributions, [8](#), [16](#)

IST, [9](#)

print.RESIDE, [13](#), [15](#)

print.summary.RESIDE, [14](#)

RESIDE (RESIDE-package), [2](#)

RESIDE-package, [2](#)

summary.RESIDE, [15](#), [15](#)

synthesise_data, [3](#), [4](#), [7](#), [8](#), [16](#)

synthesize_data (synthesise_data), [16](#)