

# Package ‘OncoDataSets’

January 20, 2025

**Type** Package

**Title** A Comprehensive Collection of Cancer Types and Cancer-Related Datasets

**Version** 0.1.0

**Maintainer** Renzo Caceres Rossi <arenzocaceresrossi@gmail.com>

**Description** Offers a rich collection of data focused on cancer research, covering survival rates, genetic studies, biomarkers, and epidemiological insights. Designed for researchers, analysts, and bioinformatics practitioners, the package includes datasets on various cancer types such as melanoma, leukemia, breast, ovarian, and lung cancer, among others. It aims to facilitate advanced research, analysis, and understanding of cancer epidemiology, genetics, and treatment outcomes.

**License** GPL-3

**URL** <https://github.com/lightbluetitan/oncodatasets>,  
<https://lightbluetitan.github.io/oncodatasets/>

**BugReports** <https://github.com/lightbluetitan/oncodatasets/issues>

**Encoding** UTF-8

**LazyData** true

**Suggests** ggplot2, dplyr, testthat (>= 3.0.0), knitr, rmarkdown

**Config/testthat/edition** 3

**RoxygenNote** 7.3.2

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** Renzo Caceres Rossi [aut, cre]

**Depends** R (>= 3.5.0)

**Repository** CRAN

**Date/Publication** 2024-12-10 21:30:17 UTC

## Contents

AflatoxinLiverCancer_df . . . . .	3
AIPulmonaryNodules_df . . . . .	4
AlcoholIntakeCancer_df . . . . .	4
BladderCancer_df . . . . .	5
BloodStorageProstate_df . . . . .	6
BrainCancerCases_df . . . . .	7
BrainCancerGeo_df . . . . .	8
BRCA1BreastCancer_df . . . . .	9
BRCA1OvarianCancer_df . . . . .	9
BRCA2BreastCancer_df . . . . .	10
BRCA2OvarianCancer_df . . . . .	11
BreastCancerWI_df . . . . .	12
CA19PancreaticCancer_df . . . . .	13
CancerSmokeCity_array . . . . .	14
cancer_in_dogs_tbl_df . . . . .	15
Carcinoma_p53_df . . . . .	15
CASP8BreastCancer_df . . . . .	16
CervicalCancer_df . . . . .	17
ChildCancer_df . . . . .	18
ColonCancerChemo_df . . . . .	19
ColorectalMiRNAs_tbl_df . . . . .	20
EndometrialCancer_df . . . . .	21
HeadNeckCarcinoma_df . . . . .	22
ICGCLiver_df . . . . .	23
LeukemiaLymphomaCases_df . . . . .	24
LeukemiaLymphomaControl_df . . . . .	24
LeukemiaLymphomaGeo_df . . . . .	25
LeukemiaRemission_df . . . . .	26
LeukemiaSurvival_df . . . . .	27
LungCancerETS_df . . . . .	27
LungNodulesDetected_df . . . . .	28
MaleMiceCancer_df . . . . .	29
Melanoma_df . . . . .	30
MiceDeathRadiation_df . . . . .	31
NCCTGLungCancer_df . . . . .	32
NodalProstate_df . . . . .	33
OncoDataSets . . . . .	34
OvarianCancer_df . . . . .	34
PancreaticMiRNAs_tbl_df . . . . .	35
ProstateMethylation_df . . . . .	36
ProstateSurgery_df . . . . .	37
ProstateSurvival_df . . . . .	38
PSAProstateCancer_df . . . . .	39
RadiationEffects_df . . . . .	40
RotterdamBreastCancer_df . . . . .	40
SkinCancerChemo_df . . . . .	42

*AflatoxinLiverCancer\_df* 3

SmallCellLung_tbl_df . . . . .	43
SmokingLungCancer_df . . . . .	43
SuspectedCancer_df . . . . .	44
UKLungCancerDeaths_df . . . . .	45
USCancerStats_df . . . . .	46
USMortalityCancer_df . . . . .	47
USRegionalMortality_df . . . . .	48
VALungCancer_list . . . . .	49
VinylideneLiverCancer_df . . . . .	49
WBreastCancer_tbl_df . . . . .	50

**Index** 52

---

AflatoxinLiverCancer\_df

*Aflatoxin Dosage and Liver Cancer in Lab Animals*

---

### Description

This dataset, AflatoxinLiverCancer\_df, is a data frame containing data from a study where varying doses of Aflatoxin B1 were administered to lab animals. The dataset records the total number of animals exposed to each dose and the number of animals that developed liver cancer.

### Usage

```
data(AflatoxinLiverCancer_df)
```

### Format

A data frame with 6 observations and 3 variables:

**dose** Dose of Aflatoxin B1 administered (integer).

**total** Total number of animals exposed to the dose (integer).

**tumor** Number of animals that developed liver cancer (integer).

### Details

The dataset name has been kept as 'AflatoxinLiverCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the faraway package. Gaylor DW (1987). \*Linear nonparametric upper limits for low dose extrapolation\*. ASA Proceedings of the Biopharmaceutical Section.

---

AIPulmonaryNodules\_df *AI for Assessment of Indeterminate Pulmonary Nodules*

---

### Description

This dataset, AIPulmonaryNodules\_df, is a data frame containing data from a study on the performance of an artificial intelligence (AI) risk stratification tool for assessing Indeterminate Pulmonary Nodules (IPNs) in chest CT scans. The dataset includes information on whether cancer was diagnosed and the AI tool's rating of the probability of cancer (from 0 to 100).

### Usage

```
data(AIPulmonaryNodules_df)
```

### Format

A data frame with 200 observations and 2 variables:

**cancer** Cancer diagnosis – whether the nodule is cancerous (1 = cancer, 0 = no cancer) (integer).

**rating** AI rating of the probability of cancer, ranging from 0 to 100 (integer).

### Details

The dataset name has been kept as 'AIPulmonaryNodules\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the R4HCR package.

---

AlcoholIntakeCancer\_df

*Alcohol Intake and Colorectal Cancer Data*

---

### Description

This dataset, AlcoholIntakeCancer\_df, is a data frame containing data related to alcohol intake and its association with colorectal cancer risk. The data includes information on alcohol intake levels (dose), the number of cancer cases, person-years of observation, and the relative risk (logrr) along with its standard error (se). The dataset consists of 48 observations with 7 variables.

### Usage

```
data(AlcoholIntakeCancer_df)
```

**Format**

A data frame with 48 observations and 7 variables:

- id** Identifier for the study (factor).
- type** Type of study (factor).
- dose** Level of alcohol intake (numeric).
- cases** Number of colorectal cancer cases (integer).
- peryears** Person-years of observation (numeric).
- logrr** Logarithm of the relative risk (numeric).
- se** Standard error of the logarithm of the relative risk (numeric).

**Details**

The dataset name has been kept as 'AlcoholIntakeCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the mixmeta package. Available at: [https://www.cengage.com/cgi-wadsworth/course\\_products\\_wp.pl?fid=M](https://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M)

---

BladderCancer_df	<i>Bladder Cancer Recurrences</i>
------------------	-----------------------------------

---

**Description**

This dataset, BladderCancer\_df, is a data frame containing data on recurrences of bladder cancer. It is commonly used to demonstrate methodology for recurrent event modelling. The dataset includes information from 340 observations and 7 variables related to bladder cancer recurrences.

**Usage**

```
data(BladderCancer_df)
```

**Format**

A data frame with 340 observations and 7 variables:

- id** Patient identifier (integer).
- rx** Treatment received: 1 = thiotepa, 2 = placebo (numeric).
- number** Number of recurrences (integer).
- size** Size of the recurrence (integer).
- stop** Time at which the event or censoring occurred (integer).
- event** Event status: 1 = recurrence, 0 = no recurrence or death (numeric).
- enum** Event enumeration (integer).

## Details

The dataset name has been kept as 'BladderCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the survival package.

---

BloodStorageProstate\_df

*Effects of Blood Storage on Prostate Cancer Study*

---

## Description

This dataset, BloodStorageProstate\_df, is a data frame containing data on 316 men who underwent radical prostatectomy and received a transfusion during or within 30 days of the surgery. The dataset includes demographic, baseline, and prognostic factors, as well as data on the time to biochemical recurrence of prostate cancer, as indicated by prostate serum antigen (PSA) levels. The main exposure of interest was the red blood cell (RBC) storage duration group, and the outcome of interest was time to PSA cancer recurrence.

## Usage

```
data(BloodStorageProstate_df)
```

## Format

A data frame with 316 observations and 20 variables:

**RBC.Age.Group** Age group of red blood cells (numeric).

**Median.RBC.Age** Median age of red blood cells (numeric).

**Age** Patient's age (numeric).

**AA** African American status (numeric).

**FamHx** Family history of prostate cancer (numeric).

**PVol** Prostate volume (numeric).

**TVol** Tumor volume (numeric).

**T.Stage** Tumor stage (numeric).

**bGS** Biopsy grade score (numeric).

**BN+** Bone metastasis status (numeric).

**OrganConfined** Organ confinement status (numeric).

**PreopPSA** Preoperative prostate serum antigen level (numeric).

**PreopTherapy** Preoperative therapy received (numeric).  
**Units** Number of blood transfusion units (numeric).  
**sGS** Surgical Gleason score (numeric).  
**AnyAdjTherapy** Any adjuvant therapy received (numeric).  
**AdjRadTherapy** Adjuvant radiation therapy received (numeric).  
**Recurrence** Cancer recurrence status (numeric).  
**Censor** Censoring status (numeric).  
**TimeToRecurrence** Time to biochemical recurrence in months (numeric).

### Details

The dataset name has been kept as 'BloodStorageProstate\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the medicaldata package. Cata et al. (2011). \*Blood Storage Duration and Biochemical Recurrence of Cancer after Radical Prostatectomy\*. Mayo Clinic Proceedings, 86(2), 120–127.

---

BrainCancerCases\_df    *New Mexico Brain Cancer Cases Data*

---

### Description

This dataset, BrainCancerCases\_df, is a data frame containing data on brain cancer cases in New Mexico. It includes information about the county, number of cases, year of diagnosis, age group, and sex of the patients. The dataset consists of 1175 observations with 5 variables.

### Usage

```
data(BrainCancerCases_df)
```

### Format

A data frame with 1175 observations and 5 variables:

**county** County of diagnosis (Factor with 31 levels).  
**cases** Number of cases (integer).  
**year** Year of diagnosis (integer).  
**agegroup** Age group of patients (integer).  
**sex** Sex of the patient (integer).

### Details

The dataset name has been kept as 'BrainCancerCases\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the rsatscan package, distributed with SaTScan software: <https://www.satscan.org>

---

BrainCancerGeo\_df      *New Mexico Brain Cancer Geography Data*

---

### Description

This dataset, BrainCancerGeo\_df, is a data frame containing geographic information related to brain cancer cases in New Mexico. It includes data on the county, latitude, and longitude of the regions where brain cancer cases have been reported. The dataset consists of 32 observations with 3 variables.

### Usage

```
data(BrainCancerGeo_df)
```

### Format

A data frame with 32 observations and 3 variables:

**county** County where the cases were recorded (Factor with 32 levels).

**lat** Latitude of the county (integer).

**long** Longitude of the county (integer).

### Details

The dataset name has been kept as 'BrainCancerGeo\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the rsatscan package, distributed with SaTScan software: <https://www.satscan.org>



---

BRCA1BreastCancer\_df *Cumulative Risk of Women Breast Cancer BRCA1 Mutation*

---

**Description**

This dataset, BRCA1BreastCancer\_df, is a data frame containing data on the cumulative risk of breast cancer in women with the BRCA1 mutation as a function of their age. The dataset includes 11 observations, with each entry representing the cumulative risk at a specific age (in years).

**Usage**

```
data(BRCA1BreastCancer_df)
```

**Format**

A data frame with 11 observations and 2 variables:

x Age of the individual in years (numeric).

y Cumulative risk of breast cancer at that age (numeric).

**Details**

The dataset name has been kept as 'BRCA1BreastCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the riskyr package.

---

BRCA1OvarianCancer\_df *Cumulative Risk of Women Ovarian Cancer BRCA1 Mutation*

---

**Description**

This dataset, BRCA1OvarianCancer\_df, is a data frame containing data on the cumulative risk of ovarian cancer in women with the BRCA1 mutation as a function of their age. The dataset includes 63 observations, with each entry representing the cumulative risk at a specific age (in years).

**Usage**

```
data(BRCA1OvarianCancer_df)
```

**Format**

A data frame with 63 observations and 2 variables:

**age** Age of the individual in years (numeric).

**cumRisk** Cumulative risk of ovarian cancer at that age (numeric).

**Details**

The dataset name has been kept as 'BRCA1OvarianCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the riskyr package. Based on Figure 2 (p. 2408) of Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K. A., Mooij, T. M., Roos-Blom, M. J., ... & BRCA1 and BRCA2 Cohort Consortium (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. JAMA, 317 (23), 2402-2416. doi: 10.1001/jama.2017.7112

---

BRCA2BreastCancer\_df *Cumulative Risk of Women Breast Cancer BRCA2 Mutation*

---

**Description**

This dataset, BRCA2BreastCancer\_df, is a data frame containing data on the cumulative risk of breast cancer in women with the BRCA2 mutation as a function of their age. The dataset includes 11 observations, with each entry representing the cumulative risk at a specific age (in years).

**Usage**

```
data(BRCA2BreastCancer_df)
```

**Format**

A data frame with 11 observations and 2 variables:

**x** Age of the individual in years (numeric).

**y** Cumulative risk of breast cancer at that age (numeric).

**Details**

The dataset name has been kept as 'BRCA2BreastCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the riskyr package.

---

BRCA2OvarianCancer\_df *Cumulative Risk of Women Ovarian Cancer BRCA2 Mutation*

---

**Description**

This dataset, BRCA2OvarianCancer\_df, is a data frame containing data on the cumulative risk of ovarian cancer in women with the BRCA2 mutation as a function of their age. The dataset includes 63 observations, with each entry representing the cumulative risk at a specific age (in years).

**Usage**

```
data(BRCA2OvarianCancer_df)
```

**Format**

A data frame with 63 observations and 2 variables:

**age** Age of the individual in years (numeric).

**cumRisk** Cumulative risk of ovarian cancer at that age (numeric).

**Details**

The dataset name has been kept as 'BRCA2OvarianCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the riskyr package. Based on Figure 2 (p. 2408) of Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K. A., Mooij, T. M., Roos-Blom, M. J., ... & BRCA1 and BRCA2 Cohort Consortium (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA*, 317 (23), 2402–2416. doi: 10.1001/jama.2017.7112

---

BreastCancerWI\_df      *Breast Cancer Wisconsin (Diagnostic)*

---

### Description

This dataset, BreastCancerWI\_df, is a data frame containing diagnostic information for 569 patients with breast cancer. The data includes features computed from digitized images of fine needle aspirates (FNA) of breast masses, as well as a diagnosis label indicating whether the mass is malignant or benign.

### Usage

```
data(BreastCancerWI_df)
```

### Format

A data frame with 569 observations and 31 variables:

**diagnosis** Diagnosis of the breast mass: malignant or benign (factor with 2 levels).

**radius\_mean** Mean radius of the mass (numeric).

**texture\_mean** Mean texture of the mass (numeric).

**perimeter\_mean** Mean perimeter of the mass (numeric).

**area\_mean** Mean area of the mass (numeric).

**smoothness\_mean** Mean smoothness of the mass (numeric).

**compactness\_mean** Mean compactness of the mass (numeric).

**concavity\_mean** Mean concavity of the mass (numeric).

**concave\_points\_mean** Mean number of concave points on the mass contour (numeric).

**symmetry\_mean** Mean symmetry of the mass (numeric).

**fractal\_dimension\_mean** Mean fractal dimension of the mass (numeric).

**radius\_sd** Standard deviation of the radius (numeric).

**texture\_sd** Standard deviation of the texture (numeric).

**perimeter\_sd** Standard deviation of the perimeter (numeric).

**area\_sd** Standard deviation of the area (numeric).

**smoothness\_sd** Standard deviation of the smoothness (numeric).

**compactness\_sd** Standard deviation of the compactness (numeric).

**concavity\_sd** Standard deviation of the concavity (numeric).

**concave\_points\_sd** Standard deviation of the number of concave points (numeric).

**symmetry\_sd** Standard deviation of the symmetry (numeric).

**fractal\_dimension\_sd** Standard deviation of the fractal dimension (numeric).

**radius\_peak** Worst (peak) value of the radius (numeric).

**texture\_peak** Worst (peak) value of the texture (numeric).  
**perimeter\_peak** Worst (peak) value of the perimeter (numeric).  
**area\_peak** Worst (peak) value of the area (numeric).  
**smoothness\_peak** Worst (peak) value of the smoothness (numeric).  
**compactness\_peak** Worst (peak) value of the compactness (numeric).  
**concavity\_peak** Worst (peak) value of the concavity (numeric).  
**concave\_points\_peak** Worst (peak) number of concave points (numeric).  
**symmetry\_peak** Worst (peak) value of the symmetry (numeric).  
**fractal\_dimension\_peak** Worst (peak) value of the fractal dimension (numeric).

### Details

The dataset name has been kept as 'BreastCancerWI\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

### Source

Data taken from the cases package. Original documentation available at: <https://archive.ics.uci.edu/ml/datasets/breast+cancer>

---

CA19PancreaticCancer\_df

*Diagnosis of Pancreatic Cancer with CA19-9 Biomarker*

---

### Description

This dataset, CA19PancreaticCancer\_df, is a data frame containing data from a diagnostic accuracy review on the CA19-9 biomarker used for diagnosing pancreatic cancer. The dataset includes the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) from various studies.

### Usage

```
data(CA19PancreaticCancer_df)
```

### Format

A data frame with 22 observations and 5 variables:

**study** Name or identifier of the study (character).  
**TP** True positives – the number of correctly identified positive cases (integer).  
**FP** False positives – the number of cases incorrectly identified as positive (integer).  
**FN** False negatives – the number of cases incorrectly identified as negative (integer).  
**TN** True negatives – the number of correctly identified negative cases (integer).

### Details

The dataset name has been kept as 'CA19PancreaticCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the R4HCR package.

---

CancerSmokeCity\_array *Lung Cancer by Smoking Status and City*

---

### Description

This dataset, CancerSmokeCity\_array, is an array containing data on lung cancer rates by smoking status and city. The data includes 32 observations organized by whether the individual smokes, their lung cancer status, and the city. The dimensions of the array are: 2 smoking statuses (smokes, does not smoke), 2 lung cancer statuses (cancer, no cancer), and 8 cities.

### Usage

```
data(CancerSmokeCity_array)
```

### Format

An array with 32 elements, with dimensions:

**Smoking** Smoking status (character): 2 categories (smokes, does not smoke).

**Lung** Lung cancer status (character): 2 categories (cancer, no cancer).

**City** City (character): 8 cities.

### Details

The dataset name has been kept as 'CancerSmokeCity\_array' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_array' indicates that the dataset is an array. The original content has not been modified in any way.

### Source

Data taken from the flatr package. Based on data in Z. Liu, Int. J. Epidemiol., 21: 197–201, 1992.

---

cancer\_in\_dogs\_tbl\_df *Cancer in Dogs and Exposure to 2,4-D Herbicide*

---

### Description

This dataset, cancer\_in\_dogs\_tbl\_df, is a tibble containing information from a study conducted in 1994. The study aimed to determine whether there is an increased risk of cancer in dogs exposed to the herbicide 2,4-Dichlorophenoxyacetic acid (2,4-D). It includes data from 491 dogs diagnosed with cancer (case group) and 945 dogs without cancer (control group).

### Usage

```
data(cancer_in_dogs_tbl_df)
```

### Format

A tibble with 1,436 observations and 2 variables:

**order** Indicates whether the dog belongs to the "case" group (with cancer) or the "control" group (without cancer) (factor with 2 levels).

**response** Indicates the dog's exposure to the herbicide 2,4-D, with levels such as "exposed" or "not exposed" (factor with 2 levels).

### Details

The dataset name has been kept as 'cancer\_in\_dogs\_tbl\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix 'tbl\_df' indicates that the dataset is a tibble. The original content has not been modified in any way.

### Source

Data taken from the openintro package. Original study: Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. Case-Control Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2,4-Dichlorophenoxyacetic Acid Herbicides. \*Journal of the National Cancer Institute\*, 83(17):1226-1231.

---

Carcinoma\_p53\_df *Mutant p53 Gene and Squamous Cell Carcinoma*

---

### Description

This dataset, Carcinoma\_p53\_df, is a data frame containing data related to the presence of the mutant p53 tumor suppressor gene and its potential role as a prognostic factor in patients with squamous cell carcinoma arising from the oropharynx cavity. The dataset includes unadjusted estimates of log hazard ratios for mutant p53 compared to normal p53 for disease-free and overall survival, along with their associated variances, collected from 6 observational studies. The dataset consists of 6 observations with 5 variables.

**Usage**

```
data(Carcinoma_p53_df)
```

**Format**

A data frame with 6 observations and 5 variables:

**study** Study identifier (integer).

**y1** Unadjusted log hazard ratio for disease-free survival (numeric).

**y2** Unadjusted log hazard ratio for overall survival (numeric).

**V1** Variance of the log hazard ratio for disease-free survival (numeric).

**V2** Variance of the log hazard ratio for overall survival (numeric).

**Details**

The dataset name has been kept as 'Carcinoma\_p53\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the mixmeta package. References:

- Jackson D, Riley R, White IR (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*. 30 (20);2481–2498.
- Tandon S, Tudur-Smith C, Riley RD, et al. (2010). A systematic review of p53 as a prognostic factor of survival in squamous cell carcinoma of the four main anatomical subsites of the head and neck. *Cancer Epidemiology, Biomarkers and Prevention*. 19 (2);574–587.
- Sera F, Armstrong B, Blangiardo M, Gasparrini A (2019). An extended mixed-effects framework for meta-analysis. *Statistics in Medicine*. 2019;38(29):5429–5444.

---

CASP8BreastCancer\_df *CASP8 Polymorphism and Breast Cancer Risk*

---

**Description**

This dataset, CASP8BreastCancer\_df, is a data frame containing results from 4 case-control studies examining the association between the CASP8 -652 6N del promoter polymorphism and breast cancer risk. The dataset includes information on the presence or absence of the polymorphism in both cases (breast cancer patients) and controls, with different genotypic combinations analyzed.

**Usage**

```
data(CASP8BreastCancer_df)
```



**Format**

A data frame with 4 observations and 7 variables:

**study** Study identifier (character).

**bc.ins.ins** Number of breast cancer cases with the ins/ins genotype (integer).

**bc.ins.del** Number of breast cancer cases with the ins/del genotype (integer).

**bc.del.del** Number of breast cancer cases with the del/del genotype (integer).

**ct.ins.ins** Number of control cases with the ins/ins genotype (integer).

**ct.ins.del** Number of control cases with the ins/del genotype (integer).

**ct.del.del** Number of control cases with the del/del genotype (integer).

**Details**

The dataset name has been kept as 'CASP8BreastCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

**Source**

Data taken from the metadat package. Frank, B., Rigas, S. H., Bermejo, J. L., Wiestler, M., Wagner, K., Hemminki, K., Reed, M. W., Sutter, C., Wappenschmidt, B., Balasubramanian, S. P., Meindl, A., Kiechle, M., Bugert, P., Schmutzler, R. K., Bartram, C. R., Justenhoven, C., Ko, Y.-D., Brüning, T., Brauch, H., Hamann, U., Pharoah, P. P. D., Dunning, A. M., Pooley, K. A., Easton, D. F., Cox, A. & Burwinkel, B. (2008). The CASP8 -652 6N del promoter polymorphism and breast cancer risk: A multicenter study. *Breast Cancer Research and Treatment*, 111(1), 139-144. <https://doi.org/10.1007/s10549-007-9752-z>

---

CervicalCancer\_df

*Cervical Cancer Screening with Smartphones*

---

**Description**

This dataset, CervicalCancer\_df, is a data frame containing data from a study evaluating the diagnostic accuracy of CIN2+ detection using a combined approach with naked-eye and digital VIA (visual inspection with acetic acid) on a Samsung Galaxy J5 smartphone, compared to traditional naked-eye inspection alone.

**Usage**

```
data(CervicalCancer_df)
```

**Format**

A data frame with 181 observations and 10 variables:

**hpv16** Presence of HPV16 (Factor with 2 levels).

**hpv1845** Presence of HPV18/45 (Factor with 2 levels).

**hpvother** Presence of other HPV strains (Factor with 2 levels).

**naked\_via** Naked-eye VIA result (Factor with 2 levels).

**smart\_via** Digital VIA result with smartphone (Factor with 2 levels).

**treatment** Treatment received (Factor with 2 levels).

**combined\_via** Combined naked-eye and digital VIA (Factor with 2 levels).

**histology** Histological diagnosis (Factor with 5 levels).

**cytology** Cytological diagnosis (Factor with 7 levels).

**CIN2plus** CIN2+ status (Factor with 2 levels).

**Details**

The dataset name has been kept as 'CervicalCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the R4HCR package. Data directly available from <https://yareta.unige.ch/archives/ffbeb6d7-b390-4755-987e-8faf85f97c67>

---

ChildCancer\_df

*Childhood Cancer Data from North Portugal*

---

**Description**

This dataset, ChildCancer\_df, is a data frame containing information on 406 children diagnosed with cancer between January 1, 1999, and December 31, 2003, in the region of North Portugal. The dataset includes complete records on the age at diagnosis, demographic details, and survival information. Due to the interval sampling, the age at diagnosis is doubly truncated by the time from birth to the beginning and end of the study.

**Usage**

```
data(ChildCancer_df)
```

**Format**

A data frame with 406 observations and 8 variables:

**X** Unspecified numerical variable (numeric).

**U** Unspecified numerical variable (numeric).

**V** Unspecified numerical variable (numeric).

**ICCGroup** Cancer group classification (numeric).

**Status** Survival status of the child: 1 = alive, 2 = deceased (numeric).

**SurvTime** Survival time in days (numeric).

**Residence** Residence type of the child: 1 = urban, 2 = rural (numeric).

**Sex** Sex of the child: 1 = male, 2 = female (numeric).

**Details**

The dataset name has been kept as 'ChildCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the DTDA package. The childhood cancer data were gathered from the IPO (Registo Oncológico do Norte) service in North Portugal, kindly provided by Doctor Maria José Bento.

---

ColonCancerChemo\_df     *Chemotherapy for Stage B/C Colon Cancer*

---

**Description**

This dataset, ColonCancerChemo\_df, is a data frame containing data from one of the first successful trials of adjuvant chemotherapy for stage B/C colon cancer. The dataset includes information from 1858 observations and 16 variables. Each patient has two records: one for recurrence and one for death.

**Usage**

```
data(ColonCancerChemo_df)
```

**Format**

A data frame with 1858 observations and 16 variables:

**id** Patient identifier (numeric).

**study** Study identifier (numeric).

**rx** Treatment received: 1 = observation, 2 = levamisole, 3 = levamisole+5-FU (factor).

- sex** Sex of the patient: 1 = male, 2 = female (numeric).
- age** Age of the patient (numeric).
- obstruct** Obstruction of the colon: 1 = yes, 0 = no (numeric).
- perfor** Perforation of the colon: 1 = yes, 0 = no (numeric).
- adhere** Adherence to nearby organs: 1 = yes, 0 = no (numeric).
- nodes** Number of positive lymph nodes detected (numeric).
- status** Survival status: 1 = alive, 2 = dead (numeric).
- differ** Tumor differentiation: 1 = well, 2 = moderate, 3 = poor (numeric).
- extent** Tumor extent: 1 = submucosa, 2 = muscle, 3 = serosa, 4 = contiguous structures (numeric).
- surg** Surgical intervention: 0 = short, 1 = long (numeric).
- node4** Presence of 4+ positive lymph nodes: 1 = yes, 0 = no (numeric).
- time** Follow-up time in days (numeric).
- etype** Event type: 1 = recurrence, 2 = death (numeric).

### Details

The dataset name has been kept as 'ColonCancerChemo\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the survival package.

---

ColorectalMiRNAs\_tbl\_df

*PubMed Data of miRNAs in Colorectal Cancer*

---

### Description

This dataset, ColorectalMiRNAs\_tbl\_df, is a tibble containing information from PubMed abstracts related to microRNAs (miRNAs) in colorectal cancer. The data provides key details such as publication metadata, article abstracts, and associated miRNAs. The dataset consists of 508 observations with 8 variables.

### Usage

```
data(ColorectalMiRNAs_tbl_df)
```

**Format**

A tibble with 508 observations and 8 variables:

**PMID** PubMed Identifier (numeric).

**Year** Publication year of the article (numeric).

**Title** Title of the PubMed article (character).

**Abstract** Abstract of the article (character).

**Language** Language of the article (character).

**Type** Type of publication, e.g., review, study (character).

**Topic** Research topic related to colorectal cancer and miRNAs (character).

**miRNA** Specific microRNAs mentioned in the publication (character).

**Details**

The dataset name has been kept as 'ColorectalMiRNAs\_tbl\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_tbl\_df' indicates that the dataset is a tibble, which is an enhanced version of a data frame in R. The original content has not been modified in any way.

**Source**

Data taken from the miRetrieve package. More information is available at: <https://pubmed.ncbi.nlm.nih.gov/>

---

EndometrialCancer\_df *Histology Grade and Risk Factors for Endometrial Cancer*

---

**Description**

This dataset, EndometrialCancer\_df, is a data frame containing information on histology grades and associated risk factors for 79 cases of endometrial cancer. The dataset provides variables related to histological grades, pathological indices, and other clinical measures. The dataset consists of 79 observations with 4 variables.

**Usage**

```
data(EndometrialCancer_df)
```

**Format**

A data frame with 79 observations and 4 variables:

**NV** Nuclear volume (integer).

**PI** Pathological index (integer).

**EH** Endometrial hyperplasia (numeric).

**HG** Histology grade (integer).

## Details

The dataset name has been kept as 'EndometrialCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the enrichwith package. The dataset was first analyzed in Heinze and Schemper (2002) and originally provided by Dr. E. Asseryanis from the Medical University of Vienna. The data was downloaded in .dat format from <https://users.stat.ufl.edu/~aa/glm/data/>, which provides datasets used in Agresti (2015).

---

HeadNeckCarcinoma\_df *Head and Neck Squamous-Cell Carcinoma Treatment*

---

## Description

This dataset, HeadNeckCarcinoma\_df, is a data frame containing results from 65 trials examining mortality risk in patients with nonmetastatic head and neck squamous-cell carcinoma receiving either locoregional treatment plus chemotherapy versus locoregional treatment alone. The dataset provides the observed minus expected number of deaths and corresponding variances in the locoregional treatment plus chemotherapy group.

## Usage

```
data(HeadNeckCarcinoma_df)
```

## Format

A data frame with 65 observations and 5 variables:

**id** Trial identifier (numeric).

**trial** Name of the trial (character).

**OmE** Observed minus expected number of deaths (numeric).

**V** Variance of the observed minus expected deaths (numeric).

**grp** Treatment group (integer).

## Details

The dataset name has been kept as 'HeadNeckCarcinoma\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the metadat package. Pignon, J. P., Bourhis, J., Domenge, C., & Designe, L. (2000). Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: Three meta-analyses of updated individual data. *Lancet*, 355(9208), 949-955. [https://doi.org/10.1016/S0140-6736\(00\)90011-4](https://doi.org/10.1016/S0140-6736(00)90011-4)

---

ICGCLiver\_df

*ICGC Liver Cancer Data from Japan*

---

**Description**

This dataset, ICGCLiver\_df, is a data frame containing liver cancer data from Japan, released by the ICGC database. The dataset includes survival time, event status, and expression levels for four genes (ANLN, CENPA, GPR182, and BCO2).

**Usage**

```
data(ICGCLiver_df)
```

**Format**

A data frame with 232 observations and 6 variables:

**time** Survival time (numeric).

**status** Event status (1 = event occurred, 0 = censored) (integer).

**ANLN** Expression level of the ANLN gene (numeric).

**CENPA** Expression level of the CENPA gene (numeric).

**GPR182** Expression level of the GPR182 gene (numeric).

**BCO2** Expression level of the BCO2 gene (numeric).

**Details**

The dataset name has been kept as 'ICGCLiver\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the ggrisk package. ICGC (International Cancer Genome Consortium) database. Liver cancer data from Japan.

---

LeukemiaLymphomaCases\_df

*North Humberside Leukemia and Lymphoma Cases*

---

### Description

This dataset, LeukemiaLymphomaCases\_df, is a data frame containing information on the number of leukemia and lymphoma cases reported in different locations within North Humberside. The dataset includes the location ID and the number of cases for each location.

### Usage

```
data(LeukemiaLymphomaCases_df)
```

### Format

A data frame with 191 observations and 2 variables:

**locationid** Location ID (integer).

**numcases** Number of leukemia and lymphoma cases (integer).

### Details

The dataset name has been kept as 'LeukemiaLymphomaCases\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the rsatscan package, distributed with SaTScan software: <https://www.satscan.org>

---

LeukemiaLymphomaControl\_df

*North Humberside Leukemia and Lymphoma Control Cases*

---

### Description

This dataset, LeukemiaLymphomaControl\_df, is a data frame containing information on the number of control cases for leukemia and lymphoma reported in different locations within North Humberside. The dataset includes the location ID and the number of control cases for each location.

### Usage

```
data(LeukemiaLymphomaControl_df)
```



**Format**

A data frame with 191 observations and 2 variables:

**locationid** Location ID (integer).

**numcontrols** Number of control cases (integer).

**Details**

The dataset name has been kept as 'LeukemiaLymphomaControl\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the rsatscan package, distributed with SaTScan software: <https://www.satscan.org>

---

LeukemiaLymphomaGeo\_df

*North Humberside Leukemia and Lymphoma Geographic Data*

---

**Description**

This dataset, LeukemiaLymphomaGeo\_df, is a data frame containing the geographical coordinates (x and y) for locations in North Humberside related to leukemia and lymphoma cases. It includes the location ID and the coordinates for each of the 191 locations.

**Usage**

```
data(LeukemiaLymphomaGeo_df)
```

**Format**

A data frame with 191 observations and 3 variables:

**locationid** Location ID (integer).

**x-coordinate** X-coordinate (integer).

**y-coordinate** Y-coordinate (integer).

**Details**

The dataset name has been kept as 'LeukemiaLymphomaGeo\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the rsatscan package, distributed with SaTScan software: <https://www.satscan.org>

---

LeukemiaRemission\_df *Impact of 6-MP on Acute Leukemia Remission Duration*

---

## Description

This dataset, `LeukemiaRemission_df`, is a data frame containing data on the duration of remission for acute leukemia patients who were randomly assigned to maintenance therapy with 6-mercaptopurine (6-MP), an active antileukemic compound, or a placebo. The dataset includes the sex, white blood cell (WBC) count, time to relapse, event status, and treatment group for the patients.

## Usage

```
data(LeukemiaRemission_df)
```

## Format

A data frame with 42 observations and 5 variables:

**sex** Sex of the patient (integer).

**wbc** White blood cell (WBC) count (numeric).

**time** Time to relapse (integer).

**event** Event status (Factor with 2 levels: 1 = relapse, 0 = no relapse).

**grp** Treatment group (Factor with 2 levels: 1 = 6-MP, 0 = placebo).

## Details

The dataset name has been kept as `'LeukemiaRemission_df'` to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the `OncoDataSets` package and assists users in identifying its specific characteristics. The suffix `'_df'` indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the `R4HCR` package. Kleinbaum, D.G. and Klein, M., 1996. *Survival Analysis: A Self-Learning Text*. Springer.

---

LeukemiaSurvival\_df      *Leukemia Remission Survival Times Placebo-Controlled RCT*

---

### Description

This dataset, LeukemiaSurvival\_df, is a data frame containing remission survival times of 42 leukemia patients enrolled in a placebo-controlled randomized controlled trial (RCT). The dataset includes information on the time to remission, patient status, sex, white blood cell count (log-transformed), and treatment regimen.

### Usage

```
data(LeukemiaSurvival_df)
```

### Format

A data frame with 42 observations and 5 variables:

**time** Time to remission in days (integer).

**status** Patient status (1 for event, 0 for censored) (integer).

**sex** Gender of the patient (numeric, 1 for male, 2 for female).

**logWBC** Log-transformed white blood cell count (numeric).

**rx** Treatment regimen (numeric, coded treatment type).

### Details

The dataset name has been kept as 'LeukemiaSurvival\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the autoReg package.

---

LungCancerETS\_df      *Passive Smoking's Lung Cancer Threat in Women*

---

### Description

This dataset, LungCancerETS\_df, is a data frame containing results from 37 studies on the risk of lung cancer in women exposed to environmental tobacco smoke (ETS) from their smoking spouse. The dataset includes data from both cohort and case-control studies, focusing on women who are lifelong nonsmokers but have been exposed to ETS.

**Usage**

```
data(LungCancerETS_df)
```

**Format**

A data frame with 37 observations and 11 variables:

**study** Study identifier (integer).

**author** Author(s) of the study (character).

**year** Year of publication (integer).

**country** Country where the study was conducted (character).

**design** Design of the study (e.g., cohort or case-control) (character).

**cases** Number of cases in the study (integer).

**or** Odds ratio estimate (numeric).

**or.lb** Lower bound of the odds ratio confidence interval (numeric).

**or.ub** Upper bound of the odds ratio confidence interval (numeric).

**yi** Effect size estimate (numeric).

**vi** Variance of the effect size estimate (numeric).

**Details**

The dataset name has been kept as 'LungCancerETS\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the metadat package. Hackshaw, A. K., Law, M. R., & Wald, N. J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal*, 315(7114), 980-988. <https://doi.org/10.1136/bmj.315.7114.980> Hackshaw, A. K. (1998). Lung cancer and passive smoking. *Statistical Methods in Medical Research*, 7(2), 119-136. <https://doi.org/10.1177/096228029800700>

---

LungNodulesDetected\_df

*Incidental or Screen-Detected Lung Nodules*

---

**Description**

This dataset, LungNodulesDetected\_df, is a data frame containing data on incidental or screen-detected lung nodules. The data includes information such as patient demographics, smoking status, nodule characteristics, and whether the nodule is malignant. The dataset was collected from patients with pulmonary nodules of up to 15mm detected on routine CT chest scans, aged 18 years or older, from 3 academic centers in the UK.

**Usage**

```
data(LungNodulesDetected_df)
```

**Format**

A data frame with 999 observations and 8 variables:

**sex** Gender of the patient, represented as a factor with 2 levels (Male, Female).

**age** Age of the patient (numeric).

**num.annotated** Number of annotated nodules (numeric).

**location** Location of the nodule, represented as a factor with 6 levels.

**spiculate** Whether the nodule is spiculated, represented as a factor with 2 levels (Yes, No).

**smoke.status** Smoking status of the patient, represented as a factor with 5 levels.

**diameter** Diameter of the nodule (numeric).

**malignant** Malignancy status of the nodule (numeric).

**Details**

The dataset name has been kept as 'LungNodulesDetected\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the R4HCR package. The dataset was collected from patients with pulmonary nodules detected on CT chest scans, aged 18 years or older, from 3 academic centers in the UK.

---

MaleMiceCancer_df	<i>Mouse Cancer Data</i>
-------------------	--------------------------

---

**Description**

This dataset, MaleMiceCancer\_df, is a data frame containing data on the occurrence of cancer in male mice. The dataset records the number of days until the occurrence of cancer under different treatment conditions. It includes 181 observations and 4 variables.

**Usage**

```
data(MaleMiceCancer_df)
```

**Format**

A data frame with 181 observations and 4 variables:

**trt** Treatment group: 1 = treatment, 2 = control (factor).

**days** Number of days until the occurrence of cancer (numeric).

**outcome** Cancer outcome: levels include 'none', 'localized', 'metastatic', and 'other' (factor).

**id** Mouse identifier (integer).

**Details**

The dataset name has been kept as 'MaleMiceCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the survival package.

---

Melanoma\_df

*Survival from Malignant Melanoma*

---

**Description**

This dataset, Melanoma\_df, is a data frame containing information about 205 patients with malignant melanoma (a type of skin cancer) who underwent a radical operation at Odense University Hospital, Denmark, between 1962 and 1977. Patients were followed up until the end of 1977. By that time, 134 patients were still alive, and 71 had died (57 due to cancer and 14 from other causes). This dataset provides detailed clinical and demographic information for studying malignant melanoma outcomes.

**Usage**

```
data(Melanoma_df)
```

**Format**

A data frame with 205 observations and 7 variables:

**time** Follow-up time in days (integer).

**status** Patient's status at the end of the study: 1 = alive, 2 = dead from cancer, 3 = dead from other causes (integer).

**sex** Sex of the patient: 1 = male, 2 = female (integer).

**age** Age of the patient at the time of surgery (integer).

**year** Year of surgery (integer).

**thickness** Tumor thickness in millimeters (numeric).

**ulcer** Presence of ulceration: 1 = no, 2 = yes (integer).

## Details

The dataset name has been kept as 'Melanoma\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the MASS package. Original study conducted at Odense University Hospital, Denmark.

---

MiceDeathRadiation\_df *Mice Deaths from Radiation*

---

## Description

This dataset, MiceDeathRadiation\_df, is a data frame containing data on deaths of RFM male mice exposed to 300 rads of x-radiation at 5–6 weeks of age. The dataset records the causes of death, which include thymic lymphoma, reticulum cell sarcoma, and other causes. Additionally, it distinguishes between mice kept in a conventional environment and those in a germ-free environment.

## Usage

```
data(MiceDeathRadiation_df)
```

## Format

A data frame with 177 observations and 4 variables:

**type** Type of environment (factor with 2 levels: conventional or germ-free).

**cause** Cause of death (factor with 3 levels: thymic lymphoma, reticulum cell sarcoma, or other).

**status** Survival status (numeric).

**y** Time to death in days (numeric).

## Details

The dataset name has been kept as 'MiceDeathRadiation\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the SMPracticals package.

---

NCCTGLungCancer\_df      *NCCTG Lung Cancer Data*

---

### Description

This dataset, NCCTGLungCancer\_df, is a data frame containing data on survival in patients with advanced lung cancer from the North Central Cancer Treatment Group (NCCTG). The data includes 228 observations and 10 variables related to clinical and performance score data for lung cancer patients.

### Usage

```
data(NCCTGLungCancer_df)
```

### Format

A data frame with 228 observations and 10 variables:

**inst** Institution code (numeric).

**time** Survival time in days (numeric).

**status** Survival status: 1 = dead, 2 = alive (numeric).

**age** Age of the patient (numeric).

**sex** Sex of the patient: 1 = male, 2 = female (numeric).

**ph.ecog** ECOG performance score (numeric).

**ph.karno** Karnofsky performance score (numeric).

**pat.karno** Patient's Karnofsky performance score (numeric).

**meal.cal** Daily calorie intake (numeric).

**wt.loss** Weight loss in kilograms (numeric).

### Details

The dataset name has been kept as 'NCCTGLungCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the nftbart package. Based on survival data from patients with advanced lung cancer from the North Central Cancer Treatment Group (NCCTG). Performance scores rate how well the patient can perform usual daily activities.



---

NodalProstate_df	<i>Nodal Involvement in Prostate Cancer</i>
------------------	---

---

## Description

This dataset, `NodalProstate_df`, is a data frame containing data on 53 patients diagnosed with prostate cancer. The dataset records several clinical and diagnostic factors to assess nodal involvement without surgery. Nodal involvement is a critical factor in determining the treatment strategy for prostate cancer patients.

## Usage

```
data(NodalProstate_df)
```

## Format

A data frame with 53 observations and 7 variables:

**m** Estimated probability of nodal involvement (numeric).

**r** Predicted nodal involvement risk (numeric).

**aged** Age group of the patient (factor with 2 levels).

**stage** Cancer stage (factor with 2 levels).

**grade** Tumor grade (factor with 2 levels).

**xray** X-ray result (factor with 2 levels).

**acid** Acid phosphatase test result (factor with 2 levels).

## Details

The dataset name has been kept as 'NodalProstate\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the `OncoDataSets` package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the `SMPracticals` package.

---

OncoDataSets	<i>OncoDataSets: A Comprehensive Collection of Cancer Types and Cancer-related DataSets</i>
--------------	---

---

### Description

This package provides a wide variety of datasets related to cancer types such as melanoma, leukemia, breast, ovarian, and lung cancer, among others.

### Details

OncoDataSets: A Comprehensive Collection of Cancer Types and Cancer-related DataSets  
 A Comprehensive Collection of Cancer Types and Cancer-related DataSets.

### Author(s)

**Maintainer:** Renzo Caceres Rossi <arenzocaceresrossi@gmail.com>

### See Also

Useful links:

- <https://github.com/lightbluetitan/oncodatasets>

---

OvarianCancer_df	<i>Ovarian Cancer Survival Data</i>
------------------	-------------------------------------

---

### Description

This dataset, OvarianCancer\_df, is a data frame containing survival data from a randomized trial comparing two treatments for ovarian cancer. It includes 26 observations and 6 variables related to patient demographics, treatment, and survival outcomes.

### Usage

```
data(OvarianCancer_df)
```

### Format

A data frame with 26 observations and 6 variables:

**futime** Follow-up time in days (numeric).

**fustat** Survival status: 1 = deceased, 0 = alive (numeric).

**age** Age of the patient in years (numeric).

**resid.ds** Residual disease: size of the largest residual tumor in centimeters (numeric).

**rx** Treatment group: 1 = standard treatment, 2 = experimental treatment (numeric).

**ecog.ps** ECOG performance status score: 0 = fully active, 1 = restricted activity, 2 = unable to carry out work activities (numeric).

## Details

The dataset name has been kept as 'OvarianCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the survival package.

---

PancreaticMiRNAs\_tbl\_df

*PubMed Data of miRNAs in Pancreatic Cancer*

---

## Description

This dataset, PancreaticMiRNAs\_tbl\_df, is a tibble containing information from PubMed abstracts related to microRNAs (miRNAs) in pancreatic cancer. The data provides key details such as publication metadata, article abstracts, and associated miRNAs. The dataset consists of 381 observations with 8 variables.

## Usage

```
data(PancreaticMiRNAs_tbl_df)
```

## Format

A tibble with 381 observations and 8 variables:

**PMID** PubMed Identifier (numeric).

**Year** Publication year of the article (numeric).

**Title** Title of the PubMed article (character).

**Abstract** Abstract of the article (character).

**Language** Language of the article (character).

**Type** Type of publication, e.g., review, study (character).

**Topic** Research topic related to pancreatic cancer and miRNAs (character).

**miRNA** Specific microRNAs mentioned in the publication (character).

## Details

The dataset name has been kept as 'PancreaticMiRNAs\_tbl\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_tbl\_df' indicates that the dataset is a tibble, which is an enhanced version of a data frame in R. The original content has not been modified in any way.

**Source**

Data taken from the miRetrieve package. More information is available at: <https://pubmed.ncbi.nlm.nih.gov/>

---

ProstateMethylation\_df

*DNA Methylation Data from Patients Prostate Cancer*

---

**Description**

This dataset, ProstateMethylation\_df, is a data frame containing pre-processed beta methylation values collected from two sample types (benign and tumor tissue) of 4 patients diagnosed with prostate cancer. The dataset can be used for analyses of methylation patterns in benign versus tumor tissues in prostate cancer cases.

**Usage**

```
data(ProstateMethylation_df)
```

**Format**

A data frame with 5067 observations and 9 variables:

**IlmnID** Unique identifier for the methylation probe (character).

**FFPE\_benign\_1** Beta methylation value for benign tissue, patient 1 (numeric).

**FFPE\_benign\_2** Beta methylation value for benign tissue, patient 2 (numeric).

**FFPE\_benign\_3** Beta methylation value for benign tissue, patient 3 (numeric).

**FFPE\_benign\_4** Beta methylation value for benign tissue, patient 4 (numeric).

**FFPE\_tumour\_1** Beta methylation value for tumor tissue, patient 1 (numeric).

**FFPE\_tumour\_2** Beta methylation value for tumor tissue, patient 2 (numeric).

**FFPE\_tumour\_3** Beta methylation value for tumor tissue, patient 3 (numeric).

**FFPE\_tumour\_4** Beta methylation value for tumor tissue, patient 4 (numeric).

**Details**

The dataset name has been kept as ProstateMethylation\_df to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the betaclust package.

---

ProstateSurgery\_df      *Prostate Cancer Surgery Study*

---

## Description

This dataset, ProstateSurgery\_df, is a data frame containing data from a study on 97 men with prostate cancer who were scheduled to undergo radical prostatectomy. The dataset includes clinical and pathological variables associated with prostate cancer.

## Usage

```
data(ProstateSurgery_df)
```

## Format

A data frame with 97 observations and 9 variables:

**lcavol** Logarithm of cancer volume (numeric).

**lweight** Logarithm of prostate weight (numeric).

**age** Patient's age in years (integer).

**lbph** Logarithm of the amount of benign prostatic hyperplasia (numeric).

**svi** Seminal vesicle invasion (binary: 0 = No, 1 = Yes; integer).

**lcp** Logarithm of capsular penetration (numeric).

**gleason** Gleason score (integer).

**pgg45** Percentage of Gleason scores 4 or 5 (integer).

**lpsa** Logarithm of prostate-specific antigen (PSA) level (numeric).

## Details

The dataset name has been kept as 'ProstateSurgery\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the faraway package.

---

ProstateSurvival\_df     *Prostate Cancer Survival Data*

---

## Description

This dataset, ProstateSurvival\_df, is a data frame containing survival times for two competing causes: time from prostate cancer diagnosis to death from prostate cancer, and time from prostate cancer diagnosis to death from other causes. The data set also contains information on several risk factors. The data in this data set are simulated from detailed competing risk survival curves and counts of numbers of patients per group presented in Lu-Yao et al. (2009).

## Usage

```
data(ProstateSurvival_df)
```

## Format

A data frame with 14,294 observations and 5 variables:

**grade** Cancer grade categorized into 2 levels (factor).

**stage** Cancer stage categorized into 3 levels (factor).

**ageGroup** Age group categorized into 4 levels (factor).

**survTime** Survival time in months from prostate cancer diagnosis (integer).

**status** Event status: 1 for death from prostate cancer, 2 for death from other causes, 0 for censored (integer).

## Details

The dataset name has been kept as 'ProstateSurvival\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the asaur package. Simulated data based on competing risk survival curves and patient counts presented in Lu-Yao et al. (2009): \*Outcomes of localized prostate cancer following conservative management\*. Journal of the American Medical Association, 302, 1202–1209.

---

PSAProstateCancer\_df *Factors associated with prostate specific antigen*

---

## Description

This dataset, PSAProstateCancer\_df, is a data frame containing data from a study by Stamey et al. (1989) to examine the association between prostate specific antigen (PSA) and several clinical measures in men about to receive a radical prostatectomy. The dataset includes 97 observations and 9 variables, each representing a factor potentially associated with PSA.

## Usage

```
data(PSAProstateCancer_df)
```

## Format

A data frame with 97 observations and 9 variables:

**lcavol** Logarithm of cancer volume (numeric).

**lweight** Logarithm of prostate weight (numeric).

**age** Age of the patient in years (integer).

**lbph** Logarithm of benign prostatic hyperplasia (numeric).

**svi** Seminal vesicle invasion (integer).

**lcp** Logarithm of cancer perineural invasion (numeric).

**gleason** Gleason score (integer).

**pgg45** Percentage of cancerous tissue with Gleason score 4 or 5 (integer).

**lpsa** Logarithm of prostate specific antigen (PSA) (numeric).

## Details

The dataset name has been kept as 'PSAProstateCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the ncvreg package. Based on data from Stamey et al. (1989), which examined the association between prostate specific antigen (PSA) and several clinical measures potentially associated with PSA in men about to receive a radical prostatectomy.

---

RadiationEffects\_df     *Radiation Dose Effects on Chromosomal Abnormality*

---

### Description

This dataset, RadiationEffects\_df, is a data frame containing data from an experiment conducted to examine the effects of gamma radiation on the number of chromosomal abnormalities observed. The data explores the relationships between radiation dose, dose rate, and chromosomal changes.

### Usage

```
data(RadiationEffects_df)
```

### Format

A data frame with 27 observations and 4 variables:

**cells** Number of cells observed (integer).

**ca** Number of chromosomal abnormalities (integer).

**doseamt** Amount of gamma radiation dose (numeric).

**doserate** Rate of gamma radiation dose (numeric).

### Details

The dataset name has been kept as 'RadiationEffects\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the faraway package. Based on the study by Purott R. and Reeder E. (1976): \*The effect of changes in dose rate on the yield of chromosome aberrations in human lymphocytes exposed to gamma radiation\*. Mutation Research, 35, 437–444.

---

RotterdamBreastCancer\_df

*Rotterdam Breast Cancer Data*

---

### Description

This dataset, RotterdamBreastCancer\_df, is a data frame containing data on 2982 patients with primary breast cancer. The data was collected as part of the Rotterdam tumor bank and was used in Royston and Altman (2013) for survival analysis and prognostic model evaluation.



**Usage**

```
data(RotterdamBreastCancer_df)
```

**Format**

A data frame with 2982 observations and 15 variables:

**pid** Patient ID (integer).

**year** Year of diagnosis (integer).

**age** Age at diagnosis in years (integer).

**meno** Menopausal status: 1 = premenopausal, 2 = postmenopausal (integer).

**size** Tumor size categorized into three levels (factor).

**grade** Tumor grade: 1 = low, 2 = intermediate, 3 = high (integer).

**nodes** Number of lymph nodes involved (integer).

**pgr** Progesterone receptor status (integer).

**er** Estrogen receptor status (integer).

**hormon** Hormonal therapy: 1 = yes, 0 = no (integer).

**chemo** Chemotherapy: 1 = yes, 0 = no (integer).

**rtime** Time to recurrence in days (numeric).

**recur** Recurrence status: 1 = recurrence, 0 = no recurrence (integer).

**dtime** Time to death in days (numeric).

**death** Death status: 1 = deceased, 0 = alive (integer).

**Details**

The dataset name has been kept as 'RotterdamBreastCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the survival package. Based on records from the Rotterdam tumor bank and used in Royston and Altman (2013) for survival analysis.

---

SkinCancerChemo\_df      *Simulated Data from Skin Cancer Chemoprevention Trial*

---

## Description

This dataset, SkinCancerChemo\_df, is a data frame containing simulated data mimicking the Skin Cancer Chemoprevention Trial as used in Chiou et al. (2017). It records tumor recurrence in patients who were part of the trial, which includes information on patient demographics, prior tumors, and the treatment they received. The dataset consists of 894 observations with 7 variables.

## Usage

```
data(SkinCancerChemo_df)
```

## Format

A data frame with 894 observations and 7 variables:

**id** Patient ID (numeric).

**time** Time to event or censoring (numeric).

**count** Number of tumor recurrences (numeric).

**age** Age of the patient at the start of the trial (numeric).

**male** Gender of the patient (1 = male, 0 = female) (numeric).

**dfmo** Indicates whether the patient received DFMO treatment (1 = yes, 0 = no) (numeric).

**priorTumor** Number of prior tumors before the trial (numeric).

## Details

The dataset name has been kept as 'SkinCancerChemo\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the `spef` package. This simulated dataset is based on the study by Chiou et al. (2017): \*Marginal and conditional cumulative incidence functions in the presence of dependent censoring\*. *Biometrics*, 73(2), 385–394.

---

SmallCellLung\_tbl\_df *Small Cell Lung Cancer Data*

---

### Description

This dataset, SmallCellLung\_tbl\_df, is a tibble containing information on the entry age and survival time of 121 patients diagnosed with small cell lung cancer (SCLC) under two different treatment regimens. The dataset provides key insights for survival analysis and treatment comparisons in patients with SCLC.

### Usage

```
data(SmallCellLung_tbl_df)
```

### Format

A tibble with 121 observations and 3 variables:

**treatment** Treatment group of the patient (factor with 2 levels).

**age** Entry age of the patient at the start of treatment (integer).

**survival** Survival time of the patient in days (integer).

### Details

The dataset name has been kept as 'SmallCellLung\_tbl\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix 'tbl\_df' indicates that the dataset is a tibble. The original content has not been modified in any way.

### Source

Data taken from the BSDA package. Originally published in: Ying, Z., Jung, S., Wei, L. 1995. Survival Analysis with Median Regression Models.

---

SmokingLungCancer\_df *Years of Smoking and Lung Cancer Deaths in Men*

---

### Description

This dataset, SmokingLungCancer\_df, is a data frame containing data on man-years of risk and observed number of lung cancer deaths among men. The data includes information about the years of smoking, pack-years, number of cigarettes smoked per day, and the number of deaths due to lung cancer.

**Usage**

```
data(SmokingLungCancer_df)
```

**Format**

A data frame with 63 observations and 4 variables:

**yrs\_smk** Years of smoking, represented as a factor with 9 levels.

**pys** Pack-years of smoking (numeric).

**num\_cigs** Number of cigarettes smoked per day, represented as a factor with 7 levels.

**deaths** Number of deaths due to lung cancer (numeric).

**Details**

The dataset name has been kept as 'SmokingLungCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the Onco-DataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the R4HCR package. Data originally from Table 24-4, page 702 of Kleinbaum et al (1988).

---

SuspectedCancer\_df      *Suspected Cancer (SCAN) Pathway*

---

**Description**

This dataset, SuspectedCancer\_df, is a data frame containing blood test results from individuals presenting with non-specific symptoms of cancer. The data was collected as part of the Suspected CANcer (SCAN) pathway, which evaluates a new standard of care for patients in primary care settings.

**Usage**

```
data(SuspectedCancer_df)
```

**Format**

A data frame with 750 observations and 8 variables:

**age** Age of the individual (numeric).

**comorbidity** Comorbidity index (numeric).

**haemoglobin** Haemoglobin level (numeric).

**albumin** Albumin level (numeric).

**alaninetrans** Alanine aminotransferase level (numeric).

**whitebloodcell** White blood cell count (numeric).

**bilirubin** Bilirubin level (numeric).

**calcium** Calcium level (numeric).

### Details

The dataset name has been kept as 'SuspectedCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the R4HCR package. Nicholson BD, Oke JL, Friedemann Smith C, et al. The Suspected CANcer (SCAN) pathway: protocol for evaluating a new standard of care for patients with non-specific symptoms of cancer. *BMJ Open* 2018;8:e018168.

---

UKLungCancerDeaths\_df *Lung Cancer Deaths among UK Physicians*

---

### Description

This dataset, UKLungCancerDeaths\_df, is a data frame containing the number of deaths due to lung cancer among British male physicians. The data is categorized by years of smoking and cigarette consumption and was originally used in Frome (1983) to analyze rates using Poisson regression models.

### Usage

```
data(UKLungCancerDeaths_df)
```

### Format

A data frame with 63 observations and 4 variables:

**years.smok** Years of smoking categorized into 9 levels (factor).

**cigarettes** Cigarette consumption categorized into 7 levels (factor).

**Time** Exposure time in person-years (numeric).

**y** Number of lung cancer deaths (numeric).

### Details

The dataset name has been kept as 'UKLungCancerDeaths\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the SMPracticals package. Based on the study by Frome, E. L. (1983): \*The analysis of rates using Poisson regression models\*. Biometrics, 39, 665–674.

---

USCancerStats\_df

*US Cancer Incidence, Mortality, and Survival Changes*

---

**Description**

This dataset, USCancerStats\_df, is a data frame containing cancer statistics for 20 solid tumor types, including incidence, mortality, and survival data. The dataset reports the absolute difference in 5-year survival between 1989-1995 and 1950-1954, as well as the percentage change in mortality and incidence from 1950 to 1996.

**Usage**

```
data(USCancerStats_df)
```

**Format**

A data frame with 20 observations and 4 variables:

**site** Tumor site (character).

**survival** Absolute difference in 5-year survival (numeric).

**mortality** Percentage change in mortality (numeric).

**incidence** Percentage change in incidence (numeric).

**Details**

The dataset name has been kept as 'USCancerStats\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the R4HCR package.

---

USMortalityCancer\_df *US Mortality Rates by Cause (Cancer) and Gender*

---

## Description

This dataset, USMortalityCancer\_df, is a data frame containing mortality rates across all ages in the USA (Nation-wide) by cause of death, sex, and rural/urban status, recorded from 2011 to 2013. It includes national aggregate rates and region-wise rates for each administrative region under the Department of Health and Human Services (HHS). The dataset consists of 40 observations with 5 variables.

## Usage

```
data(USMortalityCancer_df)
```

## Format

A data frame with 40 observations and 5 variables:

**Status** Rural or urban status (factor with 2 levels).

**Sex** Gender of the individual (factor with 2 levels).

**Cause** Cause of death (factor with 10 levels).

**Rate** Mortality rate (numeric).

**SE** Standard error of the mortality rate (numeric).

## Details

The dataset name has been kept as 'USMortalityCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

## Source

Data taken from the lattice package. This dataset is based on the study by the Rural Health Reform Policy Research Center: \*Exploring Rural and Urban Mortality Differences\*, August 2015, Bethesda, MD. Available at <https://ruralhealth.und.edu/projects/health-reform-policy-research-center/rural-urban-mortality>.

---

USRegionalMortality\_df

*US Region Mortality Rates by Cause (Cancer) and Gender*

---

### Description

This dataset, USRegionalMortality\_df, is a data frame containing mortality rates across all ages in the USA, recorded region-wise by cause of death, sex, and rural/urban status for the years 2011–2013. It includes region-wide rates for each administrative region under the Department of Health and Human Services (HHS). The dataset consists of 400 observations with 6 variables.

### Usage

```
data(USRegionalMortality_df)
```

### Format

A data frame with 400 observations and 6 variables:

**Region** Administrative region under the Department of Health and Human Services (HHS) (factor with 10 levels).

**Status** Rural or urban status (factor with 2 levels).

**Sex** Gender of the individual (factor with 2 levels).

**Cause** Cause of death (factor with 10 levels).

**Rate** Mortality rate (numeric).

**SE** Standard error of the mortality rate (numeric).

### Details

The dataset name has been kept as 'USRegionalMortality\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

### Source

Data taken from the lattice package. This dataset is based on the study by the Rural Health Reform Policy Research Center: \*Exploring Rural and Urban Mortality Differences\*, August 2015, Bethesda, MD. Available at <https://ruralhealth.und.edu/projects/health-reform-policy-research-center/rural-urban-mortality>.



---

VALungCancer\_list      *VA Lung Cancer Data Set*

---

### Description

This dataset, VALungCancer\_list, is a list containing two components: 'X' and 'y'. The data comes from a randomized trial of two treatment regimens for lung cancer. The 'X' component contains the covariates, and the 'y' component contains the survival time data. This dataset is typically used in survival analysis.

### Usage

```
data(VALungCancer_list)
```

### Format

A list with 2 components:

**X** A numeric matrix with 1137 rows and 19 columns, representing the covariates.

**y** A numeric matrix with 1137 rows and 12 columns, representing the survival time data. The columns include 'time' for the survival time and other variables related to survival analysis.

### Details

The dataset name has been kept as 'VALungCancer\_list' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_list' indicates that the dataset is a list. The original content has not been modified in any way.

### Source

Data taken from the ncvreg package. Based on data from a randomized trial of two treatment regimens for lung cancer, as presented in the classic textbook by Kalbfleisch and Prentice.

---

VinylideneLiverCancer\_df

*Effect of Vinylidene Fluoride on Liver Cancer*

---

### Description

This dataset, VinylideneLiverCancer\_df, is a data frame containing data from an experiment to investigate whether vinylidene fluoride induces liver damage. The dataset records the levels of three serum enzymes (SDH, SGOT, SGPT) under four different dosages of vinylidene fluoride. Increased serum enzyme levels are indicative of liver damage. Real data which are available on page 10 of Silvapulle and Sen (2005) and in a report prepared by Litton Bionetics Inc in 1984. These data were used in an experiment to find out whether vinylidene fluoride gives rise to liver damage.

**Usage**

```
data(VinylideneLiverCancer_df)
```

**Format**

A data frame with 40 observations and 4 variables:

**SDH** Serum enzyme SDH levels (integer).

**SGOT** Serum enzyme SGOT levels (integer).

**SGPT** Serum enzyme SGPT levels (integer).

**dose** Dose of vinylidene fluoride administered (factor with 4 levels).

**Details**

The dataset name has been kept as 'VinylideneLiverCancer\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The suffix '\_df' indicates that the dataset is a data frame. The original content has not been modified in any way.

**Source**

Data taken from the goric package. Silvapulle MJ and Sen PK (2005). \*Constrained Statistical Inference: Order, Inequality, and Shape Restrictions\*. Wiley. Litton Bionetics Inc (1984). Report on the effects of vinylidene fluoride on liver enzymes in Fischer-344 rats.

---

WBreastCancer\_tbl\_df *Women with Breast Cancer Study*

---

**Description**

This dataset, WBreastCancer\_tbl\_df, is a tibble containing data from a study among women with breast cancer. The dataset includes clinical and demographic variables for 1207 patients, providing valuable insights for breast cancer research and analysis.

**Usage**

```
data(WBreastCancer_tbl_df)
```

**Format**

A tibble with 1207 observations and 9 variables:

**id** Unique identifier for each patient (numeric).

**time** Time to the event or censoring (numeric).

**status** Event status: 1 if the event occurred, 0 if censored (numeric).

**er** Estrogen receptor status (numeric).

- age** Age of the patient at the time of diagnosis (numeric).
- histgrad** Histological grade of the tumor (numeric).
- ln\_yesno** Presence of lymph nodes: 1 if positive, 0 if negative (numeric).
- pathsd** Pathological stage of the disease (numeric).
- pr** Progesterone receptor status (numeric).

**Details**

The dataset name has been kept as 'WBreastCancer\_tbl\_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the OncoDataSets package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

**Source**

Data taken from the psfmi package.

# Index

AflatoxinLiverCancer\_df, 3  
AIPulmonaryNodules\_df, 4  
AlcoholIntakeCancer\_df, 4

BladderCancer\_df, 5  
BloodStorageProstate\_df, 6  
BrainCancerCases\_df, 7  
BrainCancerGeo\_df, 8  
BRCA1BreastCancer\_df, 9  
BRCA10varianCancer\_df, 9  
BRCA2BreastCancer\_df, 10  
BRCA20varianCancer\_df, 11  
BreastCancerWI\_df, 12

CA19PancreaticCancer\_df, 13  
cancer\_in\_dogs\_tbl\_df, 15  
CancerSmokeCity\_array, 14  
Carcinoma\_p53\_df, 15  
CASP8BreastCancer\_df, 16  
CervicalCancer\_df, 17  
ChildCancer\_df, 18  
ColonCancerChemo\_df, 19  
ColorectalMiRNAs\_tbl\_df, 20

EndometrialCancer\_df, 21

HeadNeckCarcinoma\_df, 22

ICGCLiver\_df, 23

LeukemiaLymphomaCases\_df, 24  
LeukemiaLymphomaControl\_df, 24  
LeukemiaLymphomaGeo\_df, 25  
LeukemiaRemission\_df, 26  
LeukemiaSurvival\_df, 27  
LungCancerETS\_df, 27  
LungNodulesDetected\_df, 28

MaleMiceCancer\_df, 29  
Melanoma\_df, 30  
MiceDeathRadiation\_df, 31

NCCTGLungCancer\_df, 32  
NodalProstate\_df, 33

OncoDataSets, 34  
OncoDataSets-package (OncoDataSets), 34  
OvarianCancer\_df, 34

PancreaticMiRNAs\_tbl\_df, 35  
ProstateMethylation\_df, 36  
ProstateSurgery\_df, 37  
ProstateSurvival\_df, 38  
PSAProstateCancer\_df, 39

RadiationEffects\_df, 40  
RotterdamBreastCancer\_df, 40

SkinCancerChemo\_df, 42  
SmallCellLung\_tbl\_df, 43  
SmokingLungCancer\_df, 43  
SuspectedCancer\_df, 44

UKLungCancerDeaths\_df, 45  
USCancerStats\_df, 46  
USMortalityCancer\_df, 47  
USRegionalMortality\_df, 48

VALungCancer\_list, 49  
VinylideneLiverCancer\_df, 49

WBreastCancer\_tbl\_df, 50