

# Package: cstidy (via r-universe)

June 1, 2026

**Title** Helpful Functions for Cleaning Surveillance Data

**Version** 2025.10.27

**Description** Helpful functions for the cleaning and manipulation of surveillance data, especially with regards to the creation and validation of panel data from individual level surveillance data.

**Depends** R (>= 3.5.0)

**License** MIT + file LICENSE

**URL** <https://www.csids.no/cstidy/>, <https://github.com/csids/cstidy>

**BugReports** <https://github.com/csids/cstidy/issues>

**Encoding** UTF-8

**LazyData** true

**Imports** data.table, magrittr, ggplot2, csdata, cstime, crayon, digest, stringr, methods

**Suggests** testthat, knitr, rmarkdown, rstudioapi, glue, gt, dplyr, purrr

**RoxygenNote** 7.2.3

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** Richard Aubrey White [aut, cre] (ORCID: <https://orcid.org/0000-0002-6747-1726>), CSIDS [cph]

**Maintainer** Richard Aubrey White <hello@rwhite.no>

**Config/pak/sysreqs** libicu-dev

**Repository** <https://raubreywhite.r-universe.dev>

**Date/Publication** 2025-10-27 11:30:02 UTC

**RemoteUrl** <https://github.com/cran/cstidy>

**RemoteRef** HEAD

**RemoteSha** dd6aeb49bd396531e712e3952f9129588b08dc69

## Contents

expand_time_to . . . . .	2
generate_test_data . . . . .	3
heal_time_csfmt_rts_data_v1 . . . . .	4
heal_time_csfmt_rts_data_v2 . . . . .	4
identify_data_structure . . . . .	5
nor_covid19_cases_by_time_location_csfmt_rts_v1 . . . . .	6
nor_covid19_icu_and_hospitalization_csfmt_rts_v1 . . . . .	7
remove_class_csfmt_rts_data . . . . .	8
set_csfmt_rts_data_v1 . . . . .	8
set_csfmt_rts_data_v2 . . . . .	11
unique_time_series . . . . .	14

<b>Index</b>	<b>16</b>
--------------	-----------

---

expand_time_to	<i>Expand time to</i>
----------------	-----------------------

---

## Description

Attempts to expand the dataset to include more time

A time series is defined as a unique combination of:

- granularity\_time
- granularity\_geo
- country\_iso3
- location\_code
- border
- age
- sex
- \*\_id
- \*\_tag

## Usage

```
expand_time_to(
  x,
  max_isoyear = NULL,
  max_isoyearweek = NULL,
  max_date = NULL,
  ...
)
```

**Arguments**

x	An object of type <a href="#">csfmt_rts_data_v2</a>
max_isoyear	Maximum isoyear
max_isoyearweek	Maximum isoyearweek
max_date	Maximum date
...	Not used.

**Value**

csfmt\_rts\_data\_v2, a larger dataset that includes more rows corresponding to more time.

**See Also**

Other `csfmt_rts_data`: [identify\\_data\\_structure\(\)](#), [remove\\_class\\_csfmt\\_rts\\_data\(\)](#), [set\\_csfmt\\_rts\\_data\\_v1\(\)](#), [set\\_csfmt\\_rts\\_data\\_v2\(\)](#), [unique\\_time\\_series\(\)](#)

---

generate\_test\_data      *Generate test data*

---

**Description**

Generates some test data

**Usage**

```
generate_test_data(fmt = "csfmt_rts_data_v2")
```

**Arguments**

fmt	Data format (csfmt_rts_data_v2)
-----	---------------------------------

**Value**

csfmt\_rts\_data\_v2, a dataset containing fake data.

**Examples**

```
cstidy::generate_test_data("csfmt_rts_data_v2")
```

---

heal\_time\_csfmt\_rts\_data\_v1

*Provides corresponding healed times (deprecated)*

---

**Description**

Provides corresponding healed times (deprecated)

**Usage**

```
heal_time_csfmt_rts_data_v1(x, cols, granularity_time = "date")
```

**Arguments**

**x** A vector containing either dates, isoyearweek, or isoyear.  
**cols** Columns to restrict the output to.  
**granularity\_time** date, isoyearweek, or isoyear, depending on the values contained in x.

**Value**

data.table, a dataset with time columns corresponding to the values given in x.

---

heal\_time\_csfmt\_rts\_data\_v2

*Provides corresponding healed times*

---

**Description**

Provides corresponding healed times

**Usage**

```
heal_time_csfmt_rts_data_v2(x, cols, granularity_time = "date")
```

**Arguments**

**x** A vector containing either dates, isoyearweek, or isoyear.  
**cols** Columns to restrict the output to.  
**granularity\_time** date, isoyearweek, or isoyear, depending on the values contained in x.

**Value**

data.table, a dataset with time columns corresponding to the values given in x.

---

`identify_data_structure`*Hash the data structure of a dataset for a given column*

---

## Description

Reduces the data structure of a column inside a dataset into something that describes

## Usage

```
identify_data_structure(x, col, ...)
```

```
## S3 method for class 'csfmt_rts_data_v2'
```

```
identify_data_structure(x, col, ...)
```

```
## S3 method for class '`tbl_Microsoft SQL Server`'
```

```
identify_data_structure(x, col, ...)
```

## Arguments

<code>x</code>	An object
<code>col</code>	Column name to hash
<code>...</code>	Arguments passed to or from other methods

## Value

`csfmt_rts_data_structure_hash_v2`, a summary object.

## See Also

Other `csfmt_rts_data`: [expand\\_time\\_to\(\)](#), [remove\\_class\\_csfmt\\_rts\\_data\(\)](#), [set\\_csfmt\\_rts\\_data\\_v1\(\)](#), [set\\_csfmt\\_rts\\_data\\_v2\(\)](#), [unique\\_time\\_series\(\)](#)

## Examples

```
cstidy::generate_test_data() %>%  
  cstidy::set_csfmt_rts_data_v2() %>%  
  cstidy::identify_data_structure("deaths_n") %>%  
  plot()
```

---

nor\_covid19\_cases\_by\_time\_location\_csfmt\_rts\_v1

*Covid-19 data for PCR-confirmed cases in Norway (nation and county)*

---

### Description

This data comes from the Norwegian Surveillance System for Communicable Diseases (MSIS). The date corresponds to when the PCR-test was taken.

### Usage

nor\_covid19\_cases\_by\_time\_location\_csfmt\_rts\_v1

### Format

A csfmt\_rts\_data\_v1 with 11028 rows and 18 variables:

**granularity\_time** day/isoweek

**granularity\_geo** nation, county

**country\_iso3** nor

**location\_code** norge, 11 counties

**border** 2020

**age** total

**isoyear** Isoyear of event

**isoweek** Isoweek of event

**isoyearweek** Isoyearweek of event

**season** Season of event

**seasonweek** Seasonweek of event

**calyear** Calyear of event

**calmonth** Calmonth of event

**calyearmonth** Calyearmonth of event

**date** Date of event

**covid19\_cases\_testdate\_n** Number of confirmed covid19 cases

**covid19\_cases\_testdate\_pr100000** Number of confirmed covid19 cases per 100.000 population

### Details

The raw number of cases and cases per 100.000 population are recorded.

This data was extracted on 2022-05-04.

### Source

[https://github.com/folkehelseinstituttet/surveillance\\_data/blob/master/covid19/\\_DOCUMENTATION\\_data\\_covid19\\_msis\\_by\\_time\\_location.txt](https://github.com/folkehelseinstituttet/surveillance_data/blob/master/covid19/_DOCUMENTATION_data_covid19_msis_by_time_location.txt)

---

nor\_covid19\_icu\_and\_hospitalization\_csfmt\_rts\_v1

*Norwegian Covid-19 data for ICU and hospitalization*

---

## Description

This data was extracted on 2022-05-04.

## Usage

nor\_covid19\_icu\_and\_hospitalization\_csfmt\_rts\_v1

## Format

A csfmt\_rts\_data\_v1 with 919 rows and 18 variables:

**granularity\_time** day/isoweek

**granularity\_geo** nation

**country\_iso3** nor

**location\_code** norge

**border** 2020

**age** total

**isoyear** Isoyear of event

**isoweek** Isoweek of event

**isoyearweek** Isoyearweek of event

**season** Season of event

**seasonweek** Seasonweek of event

**calyear** Calyear of event

**calmonth** Calmonth of event

**calyearmonth** Calyearmonth of event

**date** Date of event

**icu\_with\_positive\_pcr\_n** Number of new admissions to the ICU with a positive PCR test

**hospitalization\_with\_covid19\_as\_primary\_cause\_n** Number of new hospitalizations with Covid-19 as the primary cause

## Source

[https://github.com/folkehelseinstituttet/surveillance\\_data/blob/master/covid19/\\_DOCUMENTATION\\_data\\_covid19\\_hospital\\_by\\_time.txt](https://github.com/folkehelseinstituttet/surveillance_data/blob/master/covid19/_DOCUMENTATION_data_covid19_hospital_by_time.txt)

---

```
remove_class_csfmt_rts_data
  Remove class csfmt_rts_data_*
```

---

**Description**

Remove class csfmt\_rts\_data\_\*

**Usage**

```
remove_class_csfmt_rts_data(x)
```

**Arguments**

x                    data.table

**Value**

No return value, called for the side effect of removing the csfmt\_rts\_data class from x.

**See Also**

Other csfmt\_rts\_data: [expand\\_time\\_to\(\)](#), [identify\\_data\\_structure\(\)](#), [set\\_csfmt\\_rts\\_data\\_v1\(\)](#), [set\\_csfmt\\_rts\\_data\\_v2\(\)](#), [unique\\_time\\_series\(\)](#)

**Examples**

```
x <- cstdy::generate_test_data() %>%
  cstdy::set_csfmt_rts_data_v2()
class(x)
cstdy::remove_class_csfmt_rts_data(x)
class(x)
```

---

```
set_csfmt_rts_data_v1 Convert data.table to csfmt_rts_data_v1 (deprecated)
```

---

**Description**

set\_csfmt\_rts\_data\_v1 converts a data.table to csfmt\_rts\_data\_v1 by reference. csfmt\_rts\_data\_v1 creates a new csfmt\_rts\_data\_v1 (not by reference) from either a data.table or data.frame.

**Usage**

```
set_csfmt_rts_data_v1(x, create_unified_columns = TRUE, heal = TRUE)
```

```
csfmt_rts_data_v1(x, create_unified_columns = TRUE, heal = TRUE)
```

**Arguments**

x	The data.table to be converted to csfmt_rts_data_v1
create_unified_columns	Do you want it to create unified columns?
heal	Do you want to impute missing values on creation?

**Value**

An extended data.table, which has been modified by reference and returned (invisibly).

No return value, called for side effect of replacing the current data.table with a csfmt\_rts\_data\_v1 in place.

Returns a duplicated csfmt\_rts\_data\_v1.

**Smart assignment**

csfmt\_rts\_data\_v1 contains the smart assignment feature for time and geography.

When the **variables in bold** are assigned using :=, the listed variables will be automatically imputed.

**location\_code:**

- granularity\_geo
- country\_iso3

**isoyear:**

- granularity\_time
- isoweek
- isoyearweek
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

**isoyearweek:**

- granularity\_time
- isoyear
- isoweek
- season
- seasonweek
- calyear
- calmonth

- calyearmonth
- date

**date:**

- granularity\_time
- isoyear
- isoweek
- isoyearweek
- season
- seasonweek
- calyear
- calmonth
- calyearmonth

**Unified columns**

csfmt\_rts\_data\_v1 contains 16 unified columns:

- granularity\_time
- granularity\_geo
- country\_iso3
- location\_code
- border
- age
- sex
- isoyear
- isoweek
- isoyearweek
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

**See Also**

Other csfmt\_rts\_data: [expand\\_time\\_to\(\)](#), [identify\\_data\\_structure\(\)](#), [remove\\_class\\_csfmt\\_rts\\_data\(\)](#), [set\\_csfmt\\_rts\\_data\\_v2\(\)](#), [unique\\_time\\_series\(\)](#)

---

set\_csfmt\_rts\_data\_v2 *Convert data.table to csfmt\_rts\_data\_v2*

---

### Description

set\_csfmt\_rts\_data\_v2 converts a `data.table` to `csfmt_rts_data_v2` by reference. `csfmt_rts_data_v2` creates a new `csfmt_rts_data_v2` (not by reference) from either a `data.table` or `data.frame`.

### Usage

```
set_csfmt_rts_data_v2(x, create_unified_columns = TRUE, heal = TRUE)
```

```
csfmt_rts_data_v2(x, create_unified_columns = TRUE, heal = TRUE)
```

### Arguments

x	The <code>data.table</code> to be converted to <code>csfmt_rts_data_v2</code>
create_unified_columns	Do you want it to create unified columns?
heal	Do you want to impute missing values on creation?

### Details

For more details see the vignette: `vignette("csfmt_rts_data_v2", package = "cstidy")`

### Value

An extended `data.table`, which has been modified by reference and returned (invisibly).

No return value, called for side effect of replacing the current `data.table` with a `csfmt_rts_data_v2` in place.

Returns a duplicated `csfmt_rts_data_v2`.

### Smart assignment

`csfmt_rts_data_v2` contains the smart assignment feature for time and geography.

When the **variables in bold** are assigned using `:=`, the listed variables will be automatically imputed.

#### **location\_code:**

- **granularity\_geo**
- **country\_iso3**

#### **isoyear:**

- **granularity\_time**
- **isoweek**

- isoyearweek
- isoquarter
- isoyearquarter
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

**isoyearweek:**

- granularity\_time
- isoyear
- isoweek
- isoquarter
- isoyearquarter
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

**date:**

- granularity\_time
- isoyear
- isoweek
- isoyearweek
- isoquarter
- isoyearquarter
- season
- seasonweek
- calyear
- calmonth
- calyearmonth

## Unified columns

csfmt\_rts\_data\_v2 contains 16 unified columns:

- granularity\_time
- granularity\_geo
- country\_iso3
- location\_code
- border
- age
- sex
- isoyear
- isoweek
- isoyearweek
- isoquarter
- isoyearquarter
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

## See Also

Other csfmt\_rts\_data: [expand\\_time\\_to\(\)](#), [identify\\_data\\_structure\(\)](#), [remove\\_class\\_csfmt\\_rts\\_data\(\)](#), [set\\_csfmt\\_rts\\_data\\_v1\(\)](#), [unique\\_time\\_series\(\)](#)

## Examples

```
# Create some fake data as data.table
d <- cstdy::generate_test_data(fmt = "csfmt_rts_data_v2")
d <- d[1:5]

# convert to csfmt_rts_data_v2 by reference
cstdy::set_csfmt_rts_data_v2(d, create_unified_columns = TRUE)

#
d[1, isoyearweek := "2021-01"]
d
d[2, isoyear := 2019]
d
d[3, date := as.Date("2020-01-01")]
d
d[4, c("isoyear", "isoyearweek") := .(2021, "2021-01")]
d
```

```
d[5, c("location_code") := .("norge")]
d

# Investigating the data structure of one column inside a dataset
cstidy::generate_test_data() %>%
  cstidy::set_csfmt_rts_data_v2() %>%
  cstidy::identify_data_structure("deaths_n") %>%
  plot()
# Investigating the data structure via summary
cstidy::generate_test_data() %>%
  cstidy::set_csfmt_rts_data_v2() %>%
  summary()
```

---

unique\_time\_series      *Unique time series*

---

## Description

Attempts to identify the unique time series that exist in this dataset.

A time series is defined as a unique combination of:

- granularity\_time
- granularity\_geo
- country\_iso3
- location\_code
- border
- age
- sex
- \*\_id
- \*\_tag

## Usage

```
unique_time_series(x, set_time_series_id = FALSE, ...)
```

## Arguments

`x`                    An object of type `csfmt_rts_data_v2`

`set_time_series_id`            If TRUE, then `x` will have a new column called 'time\_series\_id'

`...`                    Not used.

## Value

data.table, a dataset that lists all the unique time series in `x`.

**See Also**

Other `csfmt_rts_data`: [expand\\_time\\_to\(\)](#), [identify\\_data\\_structure\(\)](#), [remove\\_class\\_csfmt\\_rts\\_data\(\)](#), [set\\_csfmt\\_rts\\_data\\_v1\(\)](#), [set\\_csfmt\\_rts\\_data\\_v2\(\)](#)

# Index

- \* **csfmt\_rts\_data**
  - expand\_time\_to, [2](#)
  - identify\_data\_structure, [5](#)
  - remove\_class\_csfmt\_rts\_data, [8](#)
  - set\_csfmt\_rts\_data\_v1, [8](#)
  - set\_csfmt\_rts\_data\_v2, [11](#)
  - unique\_time\_series, [14](#)
- \* **datasets**
  - nor\_covid19\_cases\_by\_time\_location\_csfmt\_rts\_v1, [6](#)
  - nor\_covid19\_icu\_and\_hospitalization\_csfmt\_rts\_v1, [7](#)
- csfmt\_rts\_data\_v1
  - (set\_csfmt\_rts\_data\_v1), [8](#)
- csfmt\_rts\_data\_v2, [3](#), [14](#)
- csfmt\_rts\_data\_v2
  - (set\_csfmt\_rts\_data\_v2), [11](#)
- expand\_time\_to, [2](#), [5](#), [8](#), [10](#), [13](#), [15](#)
- generate\_test\_data, [3](#)
- heal\_time\_csfmt\_rts\_data\_v1, [4](#)
- heal\_time\_csfmt\_rts\_data\_v2, [4](#)
- identify\_data\_structure, [3](#), [5](#), [8](#), [10](#), [13](#), [15](#)
- nor\_covid19\_cases\_by\_time\_location\_csfmt\_rts\_v1, [6](#)
- nor\_covid19\_icu\_and\_hospitalization\_csfmt\_rts\_v1, [7](#)
- remove\_class\_csfmt\_rts\_data, [3](#), [5](#), [8](#), [10](#), [13](#), [15](#)
- set\_csfmt\_rts\_data\_v1, [3](#), [5](#), [8](#), [8](#), [13](#), [15](#)
- set\_csfmt\_rts\_data\_v2, [3](#), [5](#), [8](#), [10](#), [11](#), [15](#)
- unique\_time\_series, [3](#), [5](#), [8](#), [10](#), [13](#), [14](#)