

Package ‘nhanesR’

August 25, 2026

Type Package

Title Download, Parse, and Analyze NHANES Data with Mortality Linkage

Version 0.1.6

Description Provides tools for downloading and organizing National Health and Nutrition Examination Survey (NHANES) public-use data files and the National Center for Health Statistics (NCHS) Public-Use Linked Mortality Files (LMF). Supports structured local caching, codebook access, survey-aware merging, and preparation of survival analysis datasets using NHANES-National Death Index (NDI) linked mortality data (follow-up through December 31, 2019). NHANES methodology is described at <https://wwwn.cdc.gov/nchs/nhanes/Default.aspx>.

License MIT + file LICENSE

Encoding UTF-8

Language en-US

Depends R (>= 4.1.0)

Imports httr2 (>= 1.0.0), haven (>= 2.5.0), readr (>= 2.1.0), rlang (>= 1.1.0), cli (>= 3.6.0), tools

Suggests foreign, survey, survival, rvest, Hmisc, testthat (>= 3.0.0), knitr, rmarkdown

Config/testthat/edition 3

VignetteBuilder knitr

URL <https://dwinsemius.github.io/nhanesR/>,
<https://github.com/dwinsemius/nhanesR>

BugReports <https://github.com/dwinsemius/nhanesR/issues>

Config/roxygen2/version 8.0.0

NeedsCompilation no

Author David Winsemius [aut, cre]

Maintainer David Winsemius <dwinsemius@comcast.net>

Repository CRAN

Date/Publication 2026-08-25 11:00:37 UTC

Contents

nhanes_cache_dir	2
nhanes_cycles	4
nhanes_download	5
nhanes_download_analyte	6
nhanes_followup_summary	8
nhanes_harmonize	8
nhanes_lmf_cycles	11
nhanes_manifest	11
nhanes_merge	12
nhanes_mortality_download	13
nhanes_mortality_link	14
nhanes_mortality_parse	16
nhanes_search_variables	17
nhanes_stack	18
nhanes_survival_prep	19
nhanes_ucod_labels	22
nhanes_variable_map	23
NH_describe	24
NH_label	25
svycph_fuse	26
svycph_set_basehaz	27
weighted_basehaz	28
Index	30

nhanes_cache_dir	<i>Get or set the nhanesR local cache directory</i>
------------------	---

Description

nhanesR stores downloaded and parsed NHANES files in a local cache to avoid redundant downloads. By default the cache is placed in the standard user data directory for your operating system (see below). Use this function to view or change the location for the current session, or set it permanently in your .Rprofile.

Usage

```
nhanes_cache_dir(path = NULL, create = TRUE)
```

Arguments

path	Optional character. New path to use as the cache directory for the current session. If NULL, returns the current setting without changing it.
create	Logical. If TRUE (default), create the directory if it does not exist.

Details

Package options:

Three options control nhanesR behavior. Set any of them in your .Rprofile to make the change permanent across sessions; changes made during a session (via `nhanes_cache_dir()` or `options()` directly) last only until the session ends.

Option	Default	Purpose
<code>nhanesR.cache_dir</code>	<code>file.path(tempdir(), "nhanesR")</code>	Root directory for all cached files
<code>nhanesR.verbose</code>	<code>TRUE</code>	Print progress messages during downloads
<code>nhanesR.timeout</code>	<code>120L</code>	HTTP request timeout in seconds

Default cache location:

By default the cache lives inside R's session-temporary directory (`tempdir()`), so `nhanesR` never writes to your home directory without your consent. Downloaded files are re-fetched in each new R session. To keep a persistent cache, set `nhanesR.cache_dir` in your `~/.Rprofile`.

Setting options permanently:

Add lines like these to your `~/.Rprofile`:

```
options(
  nhanesR.cache_dir = "/data/nhanes_cache", # shared lab server path
  nhanesR.verbose   = FALSE,               # suppress progress messages
  nhanesR.timeout   = 300L                 # 5-minute timeout for slow connections
)
```

Options set in `.Rprofile` take precedence over package defaults: `nhanesR` only sets an option at load time if it is not already defined.

Value

The current (or newly set) cache directory path, invisibly.

See Also

[nhanes_download\(\)](#) and [nhanes_download_analyte\(\)](#), whose caching behavior is controlled by the options described above.

Examples

```
# View current cache location (defaults to a subdirectory of tempdir())
nhanes_cache_dir()

# Change to a custom session-local location (CRAN-safe)

nhanes_cache_dir(file.path(tempdir(), "my_nhanes_cache"))

# Suppress download messages for this session
options(nhanesR.verbose = FALSE)

# View all current nhanesR option values
Filter(function(x) startsWith(x, "nhanesR."), names(options()))
```

nhanes_cycles	<i>List available NHANES cycles</i>
---------------	-------------------------------------

Description

Returns a data frame of all NHANES cycles known to nhanesR, including metadata about survey weights, pandemic adjustment status, and mortality linkage availability.

Usage

```
nhanes_cycles(include_iii = FALSE)
```

Arguments

include_iii Logical. Include NHANES III (1988-1994)? Default FALSE because its file naming conventions differ from continuous NHANES.

Value

A tibble with one row per cycle and columns:

cycle Character. Cycle label (e.g. "2015-2016").

begin_year, end_year Integer. Survey years.

suffix Character. Letter suffix appended to file codes.

wt_mec_2yr Character. 2-year MEC exam weight variable name.

wt_int_2yr Character. 2-year interview weight variable name.

wt_mec_4yr Character. 4-year combined weight, where available.

wt_prepan Character. Pre-pandemic weight for 2017-2020 cycle.

pandemic_adj Logical. Was this cycle pandemic-adjusted?

has_lmf_public Logical. Is a public-use LMF available?

sensor_date Character. Mortality follow-up censor date.

See Also

[nhanes_manifest\(\)](#) to see what files are available within a cycle; [nhanes_download\(\)](#) to download files; [nhanes_lmf_cycles\(\)](#) for cycles that have public-use mortality linkage.

Examples

```
nhanes_cycles()
nhanes_cycles(include_iii = TRUE)

# Extract cycle labels as a character vector for use in download functions
cycles <- nhanes_cycles()[["cycle"]]
cycles[1:10] # first ten continuous cycles (1999-2018)
```

nhanes_download	<i>Download and cache NHANES XPT data files</i>
-----------------	---

Description

Downloads one or more NHANES component files in SAS transport (XPT) format from the CDC website, parses them into R data frames, attaches variable labels, and caches the results locally as RDS files.

Usage

```
nhanes_download(file_code, cycles, refresh = FALSE, add_cycle_col = TRUE)
```

Arguments

file_code	Character. The NHANES file code(s), without suffix or extension (e.g. "DEMO", "BPX", "TRIGLY"). Case-insensitive. Can be a vector to download multiple files.
cycles	Character. One or more cycle labels (e.g. "2015-2016"). See nhanes_cycles() . If multiple cycles are given, the same file code is downloaded for each.
refresh	Logical. Re-download and re-parse even if cached? Default FALSE.
add_cycle_col	Logical. Add a cycle column to each returned data frame? Default TRUE. Required for nhanes_mortality_link() .

Details

Finding valid file codes:

File codes are the base names CDC assigns to each data file, without the cycle-letter suffix or .xpt extension. For example, the Demographics file is always "DEMO", and the blood pressure examination file is "BPX".

Use [nhanes_manifest\(\)](#) to browse all files available for a given cycle and component. The file_name column of the manifest shows the full CDC name including the cycle suffix (e.g. "TCHOL_I"); strip the trailing underscore-letter to get the base code for [nhanes_download\(\)](#):

```
m <- nhanes_manifest("2015-2016", "Laboratory")
m[, c("file_name", "description")]
```

```
# Base codes ready for nhanes_download():
sub("_[A-Z]$", "", m$file_name)
```

Note that some analyte file names changed across cycles (e.g. total cholesterol: LAB13 -> L13_B -> TCHOL_D onward). For those cases, use [nhanes_download_analyte\(\)](#) instead, which looks up the correct CDC filename for each cycle automatically via the variable catalog.

Invalid file codes:

File codes are not validated before the download attempt. If an unknown code is supplied, CDC returns HTTP 200 with an HTML error page rather than a 404. [nhanesR](#) detects this via the Content-Type header and aborts with a message directing you to [nhanes_manifest\(\)](#) to confirm the correct name.

Variable label encoding:

NHANES XPT files were produced by SAS, which writes variable labels in the system locale of the generating server — typically Latin-1 (ISO 8859-1). Some biochemistry files (notably BIOPRO and its predecessors L40_C onward) use the Latin-1 byte 0xB5 for the micro prefix in SI unit strings such as umol/L. That byte is not valid UTF-8, so any downstream code that runs regular expressions over label attributes will receive an "unable to translate ... to a wide string" warning and the label will be skipped.

nhanesR guards against this internally by passing labels through `iconv(..., to = "UTF-8", sub = "")` before pattern matching in `nhanes_harmonize()`. If you read label attributes directly in your own code, apply the same conversion:

```
labels <- iconv(
  vapply(df, function(col) attr(col, "label") %||% "", character(1L)),
  to = "UTF-8", sub = ""
)
```

Value

If a single `file_code` and single cycle are requested, a data frame. If multiple `file_codes` or cycles are requested, a named list of data frames with names of the form "`{file_code}_{cycle}`".

See Also

`nhanes_manifest()` to browse available file codes; `nhanes_download_analyte()` for analytes whose file name changed across cycles; `nhanes_cycles()` for valid cycle labels.

Examples

```
# Browse available Laboratory files for a cycle, then download by base code
m <- nhanes_manifest("2015-2016", "Laboratory")
m[, c("file_name", "description")] # see what's available
bpx <- nhanes_download("BPX", "2015-2016")

# Single file, single cycle
demo <- nhanes_download("DEMO", "2015-2016")

# Multiple cycles (returns list)
demos <- nhanes_download("DEMO", c("2013-2014", "2015-2016", "2017-2018"))

# Multiple files, single cycle
files <- nhanes_download(c("DEMO", "BPX", "TRIGLY"), "2015-2016")
```

nhanes_download_analyte

Download NHANES files for an analyte using the CDC variable catalog

Description

A smarter alternative to [nhanes_download\(\)](#) for analytes whose file name changed across cycles (e.g. total cholesterol: LAB13 -> L13_B -> L13_C -> TCHOL_D onward). Uses [nhanes_variable_map\(\)](#) to look up the correct CDC file name for each cycle, then downloads using the exact catalog name.

Usage

```
nhanes_download_analyte(
  term,
  cycles,
  component = "Laboratory",
  keep_vars = NULL,
  refresh = FALSE,
  add_cycle_col = TRUE
)
```

Arguments

term	Character. Search term passed to nhanes_variable_map() .
cycles	Character. One or more cycle labels (e.g. "1999-2000"). See nhanes_cycles() .
component	Character or NULL. NHANES component to search. Default "Laboratory".
keep_vars	Character vector or NULL. Passed to nhanes_variable_map() to disambiguate serum from urine forms of the same analyte.
refresh	Logical. Re-download even if cached? Default FALSE.
add_cycle_col	Logical. Add a cycle column to each data frame? Default TRUE.

Value

If a single cycle is requested, a data frame. If multiple cycles are requested, a named list of data frames keyed by cycle label.

See Also

[nhanes_variable_map\(\)](#) to inspect the per-cycle file/variable lookup before downloading; [nhanes_harmonize\(\)](#) to rename and stack the returned list; [nhanes_download\(\)](#) for downloading by exact file code.

Examples

```
cycles <- nhanes_cycles()[1:10, "cycle"]

# Total cholesterol -- file name changed in 1999-2004; this handles it
tchol_list <- nhanes_download_analyte("total cholesterol", cycles)

# Serum creatinine (keep_vars excludes urine creatinine)
scr_list <- nhanes_download_analyte("creatinine", cycles,
                                   keep_vars = c("LBXSCR", "LBDSCR", "LB2SCR"))
```

nhanes_followup_summary

Summarize mortality follow-up by cycle

Description

Diagnostic helper that reports median follow-up time, event rate, and maximum possible follow-up per cycle. Useful for assessing the asymmetric censoring problem when pooling cycles.

Usage

```
nhanes_followup_summary(data, cycle_col = "cycle")
```

Arguments

data	A data frame from nhanes_survival_prep() .
cycle_col	Character. Name of the cycle column. Default "cycle". Use "year" for data originating from nhanesdata.

Value

A data frame with one row per cycle.

See Also

[nhanes_survival_prep\(\)](#) which produces the required input.

Examples

```
demo <- nhanes_download("DEMO", "2015-2016")
linked <- nhanes_mortality_link(demo)
surv_data <- nhanes_survival_prep(linked, origin = "exam")
nhanes_followup_summary(surv_data)
```

nhanes_harmonize

Harmonize variable names across NHANES cycles and stack into one data frame

Description

NHANES analytes are sometimes stored under different variable names in different cycles (e.g. HDL cholesterol: LBDHDL in 1999-2002, LBXHDD in 2003-2004, LBDHDD from 2005 onward). `nhanes_harmonize()` offers two ways to resolve this:

Usage

```
nhanes_harmonize(
  data_list,
  mapping = NULL,
  unit = NULL,
  name = NULL,
  label_pattern = NULL,
  units = c("conventional", "SI", "both"),
  trim = TRUE,
  stack = TRUE
)
```

Arguments

data_list	A named list of per-cycle data frames, as returned by nhanes_download() or nhanes_download_analyte() .
mapping	A named character vector where names are the old (per-cycle) variable names and values are the single common name to use across all cycles. Multiple old names may map to the same new name. Example: <code>c(LBDHDL = "HDL_mgd1", LBXHDD = "HDL_mgd1", LBDHDD = "HDL_mgd1")</code> . Ignored when unit is provided.
unit	Character or NULL. A unit string to match against column label attributes (case-insensitive). Common values: "mg/dL", "g/dL", "U/L", "umol/L". When supplied, name must also be provided.
name	Character or NULL. The output column name to use when unit is supplied (e.g. "HDL_mgd1").
label_pattern	Character or NULL. An additional regex matched against column labels when unit is used. Use this to disambiguate when a file contains multiple columns in the same unit (e.g. "HDL" to select only the HDL column from a file that also contains total cholesterol in mg/dL).
units	Character. Controls which unit system to retain when both conventional and SI versions of the same measurement exist in a data frame (e.g. LBXTTC in mg/dL and LBDTCST in mmol/L). One of: "conventional" (default) Keep mg/dL, g/dL, U/L, etc.; drop SI duplicates. Appropriate for US-centric analyses. "SI" Keep mmol/L, g/L, umol/L, etc.; drop conventional duplicates. Appropriate for international use or journals that require SI units. "both" Retain all columns; no duplicates are removed. Detection uses label attributes rather than variable names because CDC naming is inconsistent (e.g. LBXTTC/LBDTCST). Applied before any renaming step.
trim	Logical. If TRUE (default), the returned data frame contains only SEQN, cycle, and the harmonized column(s). Set to FALSE to retain all columns (useful when the source file contains other variables you want to keep).
stack	Logical. If TRUE (default), row-bind the renamed data frames into a single data frame using nhanes_stack() . Set to FALSE to return the renamed list without stacking.

Details

- **Unit-based** (unit + name): scans column label attributes for the specified unit string (e.g. "mg/dL") and renames the matching column. No need to know the CDC variable codes in advance. Use `label_pattern` to disambiguate when a file contains multiple columns in the same unit (e.g. early cycles that bundled total cholesterol and HDL together).
- **Explicit mapping** (mapping): a named character vector of old -> new variable names. More verbose but unambiguous.

Value

If `stack = TRUE` (default), a single stacked data frame. When `unit` is used with `trim = TRUE` (the default), only `SEQN`, `cycle`, and the harmonized column are returned – ready for merging. If `stack = FALSE`, a named list of data frames.

See Also

`nhanes_download_analyte()` which produces the per-cycle list consumed by this function; `nhanes_stack()` for row-binding without renaming; `nhanes_variable_map()` to inspect variable names per cycle before choosing a mapping.

Examples

```
cycles <- nhanes_cycles()[1:10, "cycle"]

# Conventional units (default): returns SEQN + cycle + HDL_mgdl
hdl_list <- nhanes_download_analyte("HDL", cycles)
hdl <- nhanes_harmonize(hdl_list, unit = "mg/dL", name = "HDL_mgdl",
  label_pattern = "HDL")

# SI units: retain mmol/L columns, drop conventional duplicates
hdl_si <- nhanes_harmonize(hdl_list, unit = "mmol/L", name = "HDL_mmol",
  label_pattern = "HDL", units = "SI")

# Merge two analytes cleanly
tchol_list <- nhanes_download_analyte("total cholesterol", cycles)
TC <- nhanes_harmonize(tchol_list, unit = "mg/dL", name = "TC_mgdl",
  label_pattern = "total cholesterol")
lipids <- merge(hdl, TC, by = c("SEQN", "cycle"))

# Explicit mapping
hdl <- nhanes_harmonize(
  hdl_list,
  mapping = c(LBDHDL = "HDL_mgdl", LBXHDD = "HDL_mgdl",
    LBDHDD = "HDL_mgdl")
)
```

nhanes_lmf_cycles	<i>List NHANES cycles with a public-use LMF</i>
-------------------	---

Description

List NHANES cycles with a public-use LMF

Usage

```
nhanes_lmf_cycles()
```

Value

Character vector of cycle labels.

See Also

[nhanes_mortality_link\(\)](#), [nhanes_mortality_download\(\)](#)

nhanes_manifest	<i>List available files for a NHANES cycle and component</i>
-----------------	--

Description

Queries the CDC NHANES data page for a given cycle and component, returning a data frame of available files with their download URLs and documentation links.

Usage

```
nhanes_manifest(cycle, component, refresh = FALSE)
```

Arguments

cycle	Character. A cycle string, e.g. "2015-2016". See nhanes_cycles() for valid values.
component	Character. One of "Demographics", "Dietary", "Examination", "Laboratory", "Questionnaire". Case-insensitive.
refresh	Logical. Force re-query of CDC website even if cached? Default FALSE.

Details

Results are cached locally for the session to avoid repeated HTTP requests. Use refresh = TRUE to force re-query.

Value

A tibble with columns:

- cycle** Cycle label.
- component** Component name.
- file_name** File code (e.g. "DEMO_I").
- description** Plain-text description from CDC.
- xpt_url** Direct URL to the XPT data file.
- doc_url** URL to the HTML documentation/codebook page.
- date_published** Date published, if available.

See Also

[nhanes_cycles\(\)](#) for valid cycle labels; [nhanes_download\(\)](#) to download a file by its base code; [nhanes_search_variables\(\)](#) to search the variable catalog by keyword.

Examples

```
nhanes_manifest("2015-2016", "Laboratory")
nhanes_manifest("2015-2016", "Demographics")
```

nhanes_merge	<i>Merge NHANES component data frames by SEQN</i>
--------------	---

Description

Joins two or more NHANES data frames on the SEQN respondent sequence number. Validates that survey design variables (PSU, strata, weights) are present and warns when merging across components that use different weight variables.

Usage

```
nhanes_merge(..., by = "SEQN", type = c("inner", "left"), weight_var = NULL)
```

Arguments

...	Two or more data frames from nhanes_download() , each containing a SEQN column. For multi-cycle datasets, also a cycle column.
by	Character vector of join key(s). Default "SEQN". For multi-cycle data, use <code>c("SEQN", "cycle")</code> .
type	Character. Join type: "inner" retains only participants present in all files; "left" retains all participants from the first file. Default "inner" (the standard NHANES analytic approach).
weight_var	Character or NULL. If supplied, validates that this weight column exists in the merged result.

Details

Weight guidance:

The appropriate weight depends on which components are merged:

- **Demographics only** -> WTINT2YR (interview weight)
- **Any exam/lab component** -> WTMEC2YR (MEC exam weight)
- **Dietary 24-hr recall** -> WTDNR1 or WTDNR2
- **Multi-cycle pooled** -> divide the 2-year weight by the number of cycles, or use the 4-year combined weight where available

This function warns but does not enforce weight selection. Use [nhanes_cycles\(\)](#) to look up available weight variable names per cycle.

Value

A merged data frame. Duplicate columns (present in more than one input) are deduplicated, keeping the version from the first data frame where the column appears, with a warning.

See Also

[nhanes_stack\(\)](#) to row-bind per-cycle lists before merging; [nhanes_download\(\)](#) to obtain the component data frames; [nhanes_mortality_link\(\)](#) to append mortality follow-up after merging.

Examples

```
demo <- nhanes_download("DEMO", "2015-2016")
bpx <- nhanes_download("BPX", "2015-2016")
trigly <- nhanes_download("TRIGLY", "2015-2016")

analytic <- nhanes_merge(demo, bpx, trigly)

# Multi-cycle: use nhanes_stack() before merging
demo_list <- nhanes_download("DEMO", c("2013-2014", "2015-2016"))
bpx_list <- nhanes_download("BPX", c("2013-2014", "2015-2016"))
demo_pool <- nhanes_stack(demo_list)
bpx_pool <- nhanes_stack(bpx_list)
analytic <- nhanes_merge(demo_pool, bpx_pool, by = c("SEQN", "cycle"))
```

nhanes_mortality_download

Download NHANES Public-Use Linked Mortality Files

Description

Downloads the fixed-width .dat mortality files from the CDC FTP server for one or more NHANES cycles. Files are cached locally; re-downloading is skipped unless refresh = TRUE.

Usage

```
nhanes_mortality_download(
  cycles = NULL,
  refresh = FALSE,
  quiet = !getOption("nhanesR.verbose", TRUE)
)
```

Arguments

<code>cycles</code>	Character vector of cycle labels. Defaults to all cycles with a public-use LMF. See nhanes_cycles() and nhanes_lmf_cycles() .
<code>refresh</code>	Logical. Re-download even if a cached file exists? Default FALSE.
<code>quiet</code>	Logical. Suppress download messages? Default uses the <code>nhanesR.verbose</code> option.

Details

The public-use LMF provides mortality follow-up through **December 31, 2019** for NHANES 1999-2018 and NHANES III. Files were released in April 2022 and will not be updated (the 2022-linked restricted-use files require RDC access).

Value

Invisibly, a named character vector of local file paths (one per cycle). The primary side-effect is writing files to the cache directory under `mortality/dat/`.

See Also

[nhanes_mortality_parse\(\)](#), [nhanes_mortality_link\(\)](#), [nhanes_survival_prep\(\)](#)

Examples

```
# Download all available cycles
nhanes_mortality_download()

# Download specific cycles
nhanes_mortality_download(c("2013-2014", "2015-2016", "2017-2018"))
```

`nhanes_mortality_link` *Link mortality data onto an NHANES analytic dataset*

Description

Performs a left join of the parsed LMF onto a data frame containing NHANES participants, matched on the participant sequence number. Automatically handles multiple cycles by row-binding the appropriate LMF files before joining.

Usage

```
nhanes_mortality_link(
  nhanes_data,
  cycles = NULL,
  keep_vars = NULL,
  download = TRUE,
  seqn_col = "SEQN",
  cycle_col = "cycle"
)
```

Arguments

<code>nhanes_data</code>	A data frame containing NHANES participants.
<code>cycles</code>	Character vector of "YYYY-YYYY" cycle labels present in <code>nhanes_data</code> . Inferred from <code>cycle_col</code> when omitted.
<code>keep_vars</code>	Character vector of LMF variables to retain. Defaults to all: <code>c("ELIGSTAT", "MORTSTAT", "UCOD_LEADING", "DIABETES", "HYPERTEN", "PERMTH_INT", "PERMTH_EXM")</code> .
<code>download</code>	Logical. Download missing LMF files automatically? Default TRUE.
<code>seqn_col</code>	Character. Name of the participant sequence-number column in <code>nhanes_data</code> . Default "SEQN" (nhanesR / CDC standard). Use "seqn" for <code>nhanesdata</code> output.
<code>cycle_col</code>	Character. Name of the cycle column in <code>nhanes_data</code> . Default "cycle" ("YYYY-YYYY" labels). Use "year" for <code>nhanesdata</code> output, where the column contains integer start years (e.g. 1999, 2001).

Details

Data from any source – [nhanes_download\(\)](#), `nhanesA`, or `nhanesdata` – can be linked by supplying the appropriate column name arguments. For example, `nhanesdata` stores the sequence number as `seqn` (integer) and the cycle as `year` (integer start year, e.g. 1999); pass `seqn_col = "seqn"`, `cycle_col = "year"` and both are handled automatically.

Value

`nhanes_data` with LMF columns appended. Rows with no mortality record (SEQNs absent from the LMF) will have NA for all LMF columns; this should not occur for continuous NHANES 1999-2018.

See Also

[nhanes_survival_prep\(\)](#) to convert the linked data into a survival dataset; [nhanes_lmf_cycles\(\)](#) for cycles with a public-use LMF; [nhanes_stack\(\)](#) to row-bind multi-cycle data before linking.

Examples

```
demo <- nhanes_download("DEMO", "2015-2016")
demo_mort <- nhanes_mortality_link(demo)
```

nhanes_mortality_parse

Parse NHANES Linked Mortality Files into data frames

Description

Reads the fixed-width .dat files (downloading them first if needed) and returns a named list of data frames, one per cycle.

Usage

```
nhanes_mortality_parse(cycles = NULL, refresh = FALSE, download = TRUE)
```

Arguments

<code>cycles</code>	Character vector of cycle labels. Defaults to all available.
<code>refresh</code>	Logical. Re-parse even if a cached RDS exists? Default FALSE.
<code>download</code>	Logical. Auto-download missing .dat files? Default TRUE.

Details

Variable labels are attached as the "label" attribute on each column, following the haven/labelled convention.

Value

A named list of data frames. Each data frame contains:

SEQN Respondent sequence number (join key to NHANES data).

ELIGSTAT Eligibility: 1=eligible; 2=under 18; 3=insufficient data.

MORTSTAT Vital status: 0=assumed alive; 1=assumed deceased.

UCOD_LEADING Underlying cause of death (11-category ICD-10 recode).

DIABETES Diabetes mentioned on death certificate (1=yes).

HYPERTEN Hypertension mentioned on death certificate (1=yes).

PERMTH_INT Months of follow-up from interview date.

PERMTH_EXM Months of follow-up from examination date.

Note

For select records, PERMTH_INT, PERMTH_EXM, and UCOD_LEADING contain **synthetic (perturbed) values** introduced by CDC to reduce re-identification risk. MORTSTAT and ELIGSTAT are not perturbed.

See Also

[nhanes_mortality_download\(\)](#) to download the raw .dat files; [nhanes_mortality_link\(\)](#) to join parsed mortality data onto an analytic dataset.

Examples

```
lmf <- nhanes_mortality_parse(c("2015-2016", "2017-2018"))
lmf[["2015-2016"]]
```

nhanes_search_variables

Search NHANES variables by keyword

Description

Searches the CDC NHANES variable catalog for variables whose name or description matches a keyword or phrase. Results are drawn from the CDC variable list pages and cached locally to avoid repeated HTTP requests.

Usage

```
nhanes_search_variables(
  term,
  component = NULL,
  refresh = FALSE,
  summarize = TRUE
)
```

Arguments

term	Character. A keyword or phrase to search for. Matched case-insensitively against variable names and descriptions.
component	Character or NULL. Restrict search to one NHANES component: "Demographics", "Dietary", "Examination", "Laboratory", or "Questionnaire". If NULL (default), searches all components.
refresh	Logical. Re-fetch the variable catalog from CDC even if cached? Default FALSE.
summarize	Logical. If TRUE (default), collapse results to one row per unique variable name, with cycles and file names shown as comma-separated lists. Set to FALSE to return one row per variable-per-cycle.

Details

This is the recommended way to find the correct file code and variable name for an analyte across NHANES cycles. For example, total cholesterol was stored in LAB13 (1999-2000), L13_B (2001-2002), L13_C (2003-2004), and TCHOL (2005 onwards), always in variable LBXTC.

Value

When `summarize = TRUE` (default), a data frame with columns:

variable_name CDC variable code (e.g. LBXTC).

variable_desc Plain-language description.

file_names Comma-separated file codes across cycles.

cycles Comma-separated cycle labels.

n_cycles Number of cycles in which this variable appears.

When `summarize = FALSE`, one row per variable per cycle with an additional `file_name` and `component` column.

See Also

[nhanes_variable_map\(\)](#) for a per-cycle file/variable lookup table ready for use with [nhanes_download_analyte\(\)](#); [nhanes_manifest\(\)](#) to browse files rather than variables.

Examples

```
# Find total cholesterol across all cycles (summarized)
nhanes_search_variables("total cholesterol")

# Raw one-row-per-cycle output
nhanes_search_variables("total cholesterol", summarize = FALSE)

# Restrict to laboratory component
nhanes_search_variables("alanine", component = "Laboratory")

# Search for HDL cholesterol
nhanes_search_variables("HDL")
```

nhanes_stack

Stack NHANES data across multiple cycles

Description

Row-binds the same component across multiple cycles, enforcing that the `cycle` column is present and handling variable name changes across cycles.

Usage

```
nhanes_stack(..., fill = TRUE)
```

Arguments

... Named or unnamed data frames, each representing one cycle's data for the same component.

fill Logical. If TRUE (default), columns present in some but not all cycles are filled with NA in cycles where they are absent. If FALSE, only columns common to all cycles are retained.

Value

A single data frame with all cycles stacked. A cycle column is always included.

See Also

[nhanes_harmonize\(\)](#) which calls this internally and also renames variables across cycles; [nhanes_merge\(\)](#) to join components by SEQN; [nhanes_mortality_link\(\)](#) which expects a stacked data frame as input.

Examples

```
demos <- nhanes_download("DEMO", c("2013-2014", "2015-2016", "2017-2018"))
stacked <- nhanes_stack(demos)
```

nhanes_survival_prep *Prepare an NHANES-LMF dataset for survival analysis*

Description

Takes a linked NHANES-mortality data frame (from [nhanes_mortality_link\(\)](#)) and returns a dataset ready for use with [survival::Surv\(\)](#), with:

Usage

```
nhanes_survival_prep(
  data,
  origin = c("exam", "interview"),
  time_unit = c("months", "years"),
  cause = NULL,
  weight_var = NULL,
  seqn_col = "SEQN",
  cycle_col = "cycle"
)
```

Arguments

<code>data</code>	A data frame from <code>nhanes_mortality_link()</code> .
<code>origin</code>	Character. Follow-up origin: "interview" (uses PERMTH_INT) or "exam" (uses PERMTH_EXM). For analyses involving lab or examination data, use "exam".
<code>time_unit</code>	Character. "months" (default, as stored in LMF) or "years" (divides by 12).
<code>cause</code>	Character or NULL. If supplied, creates a cause-specific event indicator <code>event_cause</code> that is 1 only when UCOD_LEADING matches this code and MORTSTAT == 1. See <code>nhanes_ucod_labels()</code> for valid codes. For all-cause mortality, leave NULL.
<code>weight_var</code>	Character. Name of the survey weight column to carry through. The column is renamed to <code>survey_weight</code> in the output. If NULL, no weight is attached.
<code>seqn_col</code>	Character. Name of the participant sequence-number column. Default "SEQN". Use "seqn" for nhanesdata output.
<code>cycle_col</code>	Character. Name of the cycle column. Default "cycle". Use "year" for nhanesdata output.

Details

- Ineligible participants (ELIGSTAT != 1) removed with a warning
- A time column (in months or years) from the specified follow-up origin
- An event column (0/1) from MORTSTAT
- Optional cause-specific event indicator based on UCOD_LEADING
- Survey weight column renamed and validated

Value

A data frame with additional columns:

time Follow-up time in specified units.

event All-cause mortality indicator (0/1).

event_cause Cause-specific indicator, if cause supplied.

survey_weight Survey weight, renamed from `weight_var`.

.eligstat_dropped Attribute recording how many rows were removed.

Eligibility filtering

Participants with ELIGSTAT != 1 are **automatically removed** with a warning. This includes those under 18 at time of survey (ELIGSTAT == 2) and those with insufficient identifying data for linkage (ELIGSTAT == 3). The number dropped is reported and attached as an attribute.

Asymmetric follow-up across cycles

All public-use LMF files censor at December 31, 2019. Participants enrolled in later cycles (e.g. 2017-2018) have substantially shorter maximum follow-up than those enrolled in 1999-2000. This function warns when multiple cycles are detected, as this asymmetry must be accounted for in any pooled analysis.

Survey weights

NHANES provides three families of survey weight, each correcting for a different sampling stage. Choosing the wrong weight produces biased point estimates and incorrect standard errors.

Weight	When to use
WTINT2YR	Interview-only data (questionnaires, no lab/exam)
WTMEC2YR	Any examination or laboratory component
WTSAF2YR	Analytes from the fasting subsample (triglycerides, glucose, insulin, calculated LDL)

The fasting subsample weight (WTSAF2YR) is a *statistical* probability weight – not a body-weight measurement – that accounts for the additional random subsampling of participants asked to fast before their blood draw. Fasting participants are a minority of all MEC attendees; using WTMEC2YR for fasting analytes ignores this extra subsampling step and will give incorrect population estimates.

For pooled multi-cycle analyses divide the 2-year weight by the number of cycles pooled, or use the pre-computed 4-year weight WTMEC4YR where available. See the NHANES analytic guidelines for details.

Perturbed variables

PERMTH_INT, PERMTH_EXM, and UCOD_LEADING contain synthetic values for select records (CDC data perturbation to reduce re-identification risk). MORTSTAT is not perturbed. Cause-specific analyses using UCOD_LEADING should be interpreted with this in mind.

See Also

[nhanes_mortality_link\(\)](#) which produces the input for this function; [nhanes_followup_summary\(\)](#) to check follow-up time by cycle; [nhanes_ucod_labels\(\)](#) for cause-of-death codes accepted by cause.

Examples

```
demo_list <- nhanes_download("DEMO", c("2013-2014", "2015-2016"))
demo <- nhanes_stack(demo_list)
demo_mort <- nhanes_mortality_link(demo)

# All-cause mortality, exam origin, MEC 2-year weight
surv_data <- nhanes_survival_prep(
  demo_mort,
  origin = "exam",
  time_unit = "years",
  weight_var = "WTMEC2YR"
)

# Cause-specific: cardiovascular (code "001")
surv_data_cvd <- nhanes_survival_prep(
  demo_mort,
  origin = "exam",
  cause = "001",
  weight_var = "WTMEC2YR"
```

```

)

# Use with survival package
library(survival)
library(survey)
design <- svydesign(
  id      = ~SDMVPSU,
  strata  = ~SDMVSTRA,
  weights = ~survey_weight,
  nest    = TRUE,
  data    = surv_data
)

```

nhanes_ucod_labels	<i>Lookup table for UCOD_LEADING cause-of-death codes</i>
--------------------	---

Description

Returns the ICD-10 recode table used in the public-use LMF UCOD_LEADING variable, including code, plain-language label, and ICD-10 chapter ranges.

Usage

```
nhanes_ucod_labels()
```

Value

A data frame with columns code, label, icd10_range.

See Also

[nhanes_survival_prep\(\)](#) where the cause argument accepts these codes.

Examples

```
nhanes_ucod_labels()
```

nhanes_variable_map	<i>Build a per-cycle variable map for an analyte</i>
---------------------	--

Description

Wraps [nhanes_search_variables\(\)](#) to return a single-row-per-cycle lookup table showing which variable name and file to use for a given analyte across NHANES cycles. Useful for analytes whose variable names changed between cycles (e.g. HDL cholesterol: LBDHDL -> LBXHDD -> LBDHDD).

Usage

```
nhanes_variable_map(
  term,
  component = "Laboratory",
  keep_vars = NULL,
  include_reliability = FALSE,
  refresh = FALSE
)
```

Arguments

term	Character. Search term passed to nhanes_search_variables() .
component	Character or NULL. NHANES component to search. Default "Laboratory".
keep_vars	Character vector or NULL. If provided, only variables whose names appear in this vector are retained before the per-cycle deduplication step. Useful for disambiguating serum vs. urine forms of the same analyte (e.g. serum vs. urinary creatinine).
include_reliability	Logical. Include NHANES reliability substudy files, in which a random subset of participants had a second blood draw to measure within-subject variability? Default FALSE. Reliability files are identified by the LB2 variable prefix and _2_ file name pattern (e.g. 113_2_b, 113_2_r). These files should not be pooled with the main analytic dataset as they would duplicate participants.
refresh	Logical. Re-fetch the variable catalog? Default FALSE.

Details

When multiple variables match within a cycle (e.g. mg/dL and mmol/L versions), the function prefers the non-SI variable. Comment-code variables (suffix LC/LCN or "comment" in description) are always dropped.

Value

A data frame with columns cycle, variable_name, and file_name, one row per cycle in which the analyte was measured. Returns zero rows if nothing matches.

See Also

[nhanes_search_variables\(\)](#) for the underlying keyword search; [nhanes_download_analyte\(\)](#) which uses this map to resolve filenames automatically.

Examples

```
# HDL cholesterol across all cycles
nhanes_variable_map("HDL")

# Serum creatinine only (exclude urine variables)
nhanes_variable_map("creatinine",
  keep_vars = c("LBXSCR", "LBDSCR", "LB2SCR"))

# Urinary albumin only
nhanes_variable_map("albumin", keep_vars = c("URXUMA", "UR2UMA", "UR1MA"))
```

NH_describe	<i>NHANES-annotated variable descriptions</i>
-------------	---

Description

Attaches CDC plain-language descriptions as column labels via [NH_label](#), then calls [describe](#). This is a convenience wrapper; for repeated use prefer [NH_label\(\)](#) once so that labels persist across all subsequent Hmisc operations.

Usage

```
NH_describe(x, descriptions = NULL, all_weights = FALSE, ...)
```

Arguments

x	A data frame of NHANES data, typically from nhanes_download_analyte .
descriptions	Optional lookup passed through to NH_label . See that function for accepted forms.
all_weights	Logical. If FALSE (default), columns whose names match REP[0-9]+\$ (balanced repeated replication weights such as WTMREP01–WTMREP52) are excluded from the output. Set to TRUE to include all weight columns.
...	Additional arguments passed to describe .

Details

Replicate weights (variables matching REP[0-9]+\$, e.g. WTMREP01–WTMREP52 and WTIREP01–WTIREP52) are suppressed by default because they appear in NHANES DEMO files but are not needed for Taylor-series linearization variance estimation, which is the standard approach for NHANES analysis. Set `all_weights = TRUE` to include them.

Value

An object of class "describe" with CDC descriptions embedded as variable labels.

See Also

[NH_label](#) to attach labels to a data frame for persistent use; [describe](#) for the underlying engine.

Examples

```
tc <- nhanes_download_analyte("total cholesterol", "2015-2016")
invisible(NH_describe(tc))

# Include replicate weights in the output
demo_list <- nhanes_download("DEMO", nhanes_cycles()[1:10, "cycle"])
demo <- nhanes_stack(demo_list)
invisible(NH_describe(demo, all_weights = TRUE))

# Supply descriptions from a prior nhanes_search_variables() call
vars <- nhanes_search_variables("cholesterol")
invisible(NH_describe(tc, descriptions = vars))
```

NH_label

Attach CDC variable descriptions as Hmisc-style labels

Description

Looks up the plain-language CDC description for each column in a NHANES data frame and stores it as an "label" attribute on the column. Hmisc reads these attributes automatically in [describe](#), [Hmisc::summary\(\)](#), [Hmisc::html\(\)](#), and other label-aware functions, so labelling once makes descriptions available everywhere.

Usage

```
NH_label(x, descriptions = NULL)
```

Arguments

- | | |
|--------------|---|
| x | A data frame of NHANES data, typically from nhanes_download_analyte . |
| descriptions | Optional lookup for variable descriptions. May be: <ul style="list-style-type: none"> • NULL (default): descriptions are loaded from the locally cached variable catalog. Run any nhanes_search_variables call first to populate the cache. • A data.frame with columns <code>variable_name</code> and <code>variable_desc</code>, such as the output of nhanes_search_variables. • A named character vector mapping variable names to descriptions. |

Value

x with "label" attributes set on each column that could be matched to a CDC description. Columns with no catalog match are returned unchanged.

See Also

[NH_describe](#) for a one-step labelled describe; [nhanes_search_variables](#) to browse the variable catalog; [label](#) for the Hmisc label convention.

Examples

```
tc <- nhanes_download_analyte("total cholesterol", "2015-2016")
tc <- NH_label(tc)

# CDC descriptions now appear in all Hmisc label-aware output
invisible(Hmisc::describe(tc))
```

svycph_fuse

Fuse svycoxph and cph objects for rms ecosystem compatibility

Description

Takes a fitted cph object and a fitted svycoxph object with the same formula, and returns a modified cph-class object with survey-correct coefficients and variance-covariance matrix while preserving the rms \$Design structure needed for `anova.rms()`, `Predict()`, `summary.rms()`, and related generics.

Usage

```
svycph_fuse(fit_cph, fit_svy)
```

Arguments

`fit_cph` A cph object fitted with `x=TRUE`, `y=TRUE`, `surv=TRUE`.
`fit_svy` A svycoxph object fitted with the same formula and data.

Details

`anova.rms()` constructs Wald tests as $(L\beta)^\top (LVL^\top)^{-1} (L\beta)$ where L is a contrast matrix derived from \$Design and V is \$var. Substituting the survey-corrected V from svycoxph yields design-correct Wald statistics while preserving the rms term-identification machinery.

Degrees of freedom in `anova.rms()` are based on contrast matrix rank, not survey PSU count. For fully correct F-test df, use `survey::regTermTest()` directly on `fit_svy`.

Value

A modified cph object with survey-correct inference. The `$var` slot contains the sandwich variance-covariance matrix from `fit_svy`; `$coefficients` contains the weighted partial likelihood estimates. The `$Design` and all structural slots are preserved from `fit_cph`.

References

- Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79(1), 139–147.
- Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika*, 87(1), 37–47.

See Also

[weighted_basehaz](#), [svycph_set_basehaz](#)

svycph_set_basehaz	<i>Substitute weighted baseline hazard into a fused cph object</i>
--------------------	--

Description

Replaces the baseline hazard estimate in a fused cph object with the survey-weighted version from `weighted_basehaz()`, enabling `survplot()` to produce design-correct survival curves.

Usage

```
svycph_set_basehaz(fit_fused, h0)
```

Arguments

<code>fit_fused</code>	A <code>svycph_fused</code> object from <code>svycph_fuse()</code> .
<code>h0</code>	A data frame returned by <code>weighted_basehaz()</code> .

Value

The modified fused object with weighted baseline hazard.

See Also

[svycph_fuse](#), [weighted_basehaz](#)

weighted_basehaz	<i>Compute survey-weighted cumulative baseline hazard with variance</i>
------------------	---

Description

Implements the weighted Breslow estimator and its linearization-based variance following Lin (2000). The point estimate weights each event's contribution by its survey weight; the variance uses the influence function of the weighted estimator combined with the survey design's stratified cluster structure.

Usage

```
weighted_basehaz(
  fit_svy,
  design,
  centered = TRUE,
  se_type = c("lin", "greenwood")
)
```

Arguments

fit_svy	A fitted svycoxph object.
design	The svydesign object used to fit fit_svy.
centered	Logical. If TRUE (default), the baseline hazard corresponds to a person at the weighted mean of each covariate, matching the centered=TRUE default of survival::basehaz().
se_type	Character. Which variance estimator to use for std.err: "lin" (default) Lin (2000) design-based linearization variance. Measures sensitivity to PSU selection; appropriate for population-level design inference. Produces very small SEs for large-population surveys like NHANES. "greenwood" Survey-weighted Greenwood formula: $\sum_{t_k \leq t} n^w(t_k) / [Y^w(t_k)]^2$. Measures statistical precision from the weighted event count; gives confidence bands of interpretable width for survplot().

Details

The weighted Breslow increment at each event time t_k is:

$$d\hat{H}_0^w(t_k) = \frac{\sum_{i:t_i=t_k, \delta_i=1} w_i}{\sum_{j \in \mathcal{R}(t_k)} w_j \exp(\mathbf{X}_j^\top \hat{\beta})}$$

The influence function of $d\hat{H}_0^w(t_k)$ for observation i is (Lin 2000, eq. 2.3):

$$\phi_i(t_k) = \frac{I(t_i = t_k, \delta_i = 1)}{Y^w(t_k)} - \frac{n^w(t_k)}{[Y^w(t_k)]^2} I(t_i \geq t_k) \exp(\mathbf{X}_i^\top \hat{\beta})$$

where $Y^w(t_k) = \sum_{j \in \mathcal{R}(t_k)} w_j \exp(\mathbf{X}_j^\top \hat{\beta})$ and $n^w(t_k) = \sum_{i:t_i=t_k, \delta_i=1} w_i$.

The cumulative influence $\Phi_i(t) = \sum_{t_k \leq t} \phi_i(t_k)$ is used to construct the linearization variance estimate (Lin 2000, eq. 2.4):

$$\widehat{\text{Var}}(\hat{H}_0^w(t)) = \sum_h \frac{n_h}{n_h - 1} \sum_{\alpha \in h} (e_{h\alpha}(t) - \bar{e}_h(t))^2$$

where $e_{h\alpha}(t) = \sum_{i \in \text{PSU } \alpha} \Phi_i(t)$ is the PSU-level total of influence functions within stratum h .

Note: this variance conditions on $\hat{\beta}$ and does not propagate uncertainty from coefficient estimation. For large samples this contribution is negligible relative to the design variance.

Value

A data frame with columns:

time	Event times.
hazard	Weighted cumulative baseline hazard $H_0(t)$.
surv	Baseline survival $\exp(-H_0(t))$.
se_H0	Standard error of $H_0(t)$ on the hazard scale.
std.err	Standard error of $\log(H_0(t))$, for direct substitution into the <code>\$std.err</code> slot of a fused <code>cph</code> object. Computed from <code>se_H0</code> via the delta method: $\text{SE}(\log H) = \text{SE}(H)/H$.

Choosing se_type

For NHANES-scale populations the Lin design variance is orders of magnitude smaller than the Greenwood-weighted variance because the former captures only PSU-selection uncertainty (very small for rare events), while the latter captures statistical uncertainty from the weighted event count. The ratio is approximately proportional to the square root of the mean survey weight. Use `se_type = "greenwood"` when the goal is `survplot()` confidence bands that convey statistical reliability; use `"lin"` when the goal is design-consistent variance for formal population inference.

References

Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika*, 87(1), 37–47.

See Also

[svycph_fuse](#), [svycph_set_basehaz](#)

Index

describe, [24](#), [25](#)

label, [26](#)

NH_describe, [24](#), [26](#)
NH_label, [24](#), [25](#), [25](#)
nhanes_cache_dir, [2](#)
nhanes_cycles, [4](#)
nhanes_cycles(), [5–7](#), [11–14](#)
nhanes_download, [5](#)
nhanes_download(), [3](#), [4](#), [7](#), [9](#), [12](#), [13](#), [15](#)
nhanes_download_analyte, [6](#), [24](#), [25](#)
nhanes_download_analyte(), [3](#), [5](#), [6](#), [9](#), [10](#),
[18](#), [24](#)
nhanes_followup_summary, [8](#)
nhanes_followup_summary(), [21](#)
nhanes_harmonize, [8](#)
nhanes_harmonize(), [6](#), [7](#), [19](#)
nhanes_lmf_cycles, [11](#)
nhanes_lmf_cycles(), [4](#), [14](#), [15](#)
nhanes_manifest, [11](#)
nhanes_manifest(), [4–6](#), [18](#)
nhanes_merge, [12](#)
nhanes_merge(), [19](#)
nhanes_mortality_download, [13](#)
nhanes_mortality_download(), [11](#), [17](#)
nhanes_mortality_link, [14](#)
nhanes_mortality_link(), [5](#), [11](#), [13](#), [14](#), [17](#),
[19–21](#)
nhanes_mortality_parse, [16](#)
nhanes_mortality_parse(), [14](#)
nhanes_search_variables, [17](#), [25](#), [26](#)
nhanes_search_variables(), [12](#), [23](#), [24](#)
nhanes_stack, [18](#)
nhanes_stack(), [9](#), [10](#), [13](#), [15](#)
nhanes_survival_prep, [19](#)
nhanes_survival_prep(), [8](#), [14](#), [15](#), [22](#)
nhanes_ucod_labels, [22](#)
nhanes_ucod_labels(), [20](#), [21](#)
nhanes_variable_map, [23](#)
nhanes_variable_map(), [7](#), [10](#), [18](#)
survival::Surv(), [19](#)
svycph_fuse, [26](#), [27](#), [29](#)
svycph_set_basehaz, [27](#), [27](#), [29](#)
weighted_basehaz, [27](#), [28](#)