

# Package ‘NHSRdatasets’

July 21, 2025

**Type** Package

**Title** NHS and Healthcare-Related Data for Education and Training

**Date** 2021-03-09

**Version** 0.3.0

**Maintainer** Chris Mainey <c.mainey@nhs.net>

**Description** Free United Kingdom National Health Service (NHS) and other healthcare, or population health-related data for education and training purposes. This package contains synthetic data based on real healthcare datasets, or cuts of open-licenced official data. This package exists to support skills development in the NHS-R community: <<https://nhsrcommunity.com/>>.

**License** CC0

**Language** en-GB

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.1.1

**Depends** R (>= 3.5.0)

**BugReports** <https://github.com/nhs-r-community/NHSRdatasets/issues>

**Imports** tibble

**Suggests** dplyr, caret, e1071, readr, lattice, magrittr, varhandle, rsample, ggplot2, lme4, MASS, ModelMetrics, lme4, rmarkdown, scales, ggrepel, lubridate, tidyr, forcats, knitr, janitor, stringi, synthpop

**VignetteBuilder** knitr

**URL** <https://github.com/nhs-r-community/NHSRdatasets>,  
<https://nhs-r-community.github.io/NHSRdatasets/>

**NeedsCompilation** no

**Author** Gary Hutson [aut],  
Tom Jemmett [aut] (ORCID: <<https://orcid.org/0000-0002-6943-2990>>),  
Chris Mainey [aut, cre] (ORCID:

<<https://orcid.org/0000-0002-3018-6171>>),  
 Zoë Turner [aut],  
 NHS-R community [cph]

**Repository** CRAN

**Date/Publication** 2021-03-14 00:00:06 UTC

## Contents

ae_attendances . . . . .	2
LOS_model . . . . .	3
ons_mortality . . . . .	4
stranded_data . . . . .	5
synthetic_news_data . . . . .	6
<b>Index</b>	<b>8</b>

---

ae_attendances	<i>NHS England Accident &amp; Emergency Attendances and Admissions</i>
----------------	--

---

## Description

Reported attendances, 4 hour breaches and admissions for all A&E departments in England for the years 2016/17 through 2018/19 (Apr-Mar). The data has been tidied to be easily usable within the tidyverse of packages.

## Usage

```
data(ae_attendances)
```

## Format

Tibble with six columns

**period** The month that this data relates to

**org\_code** The **ODS** code for this provider

**type** The **department type**. either 1, 2 or other

**attendances** the number of patients who attended this department in this month

**breaches** the number of patients who breaches the 4 hour target in this month

**admissions** the number of patients admitted from A&E to the hospital in this month

## Details

Data sourced from **NHS England Statistical Work Areas** which is available under the **Open Government Licence v3.0**

**Source**

**NHS England Statistical Work Areas**

**Examples**

```
data(ae_attendances)
library(dplyr)
library(ggplot2)
library(scales)

# Create a plot of the performance for England over time
ae_attendances %>%
  group_by(period) %>%
  summarise_at(vars(attendances, breaches), sum) %>%
  mutate(performance = 1 - breaches / attendances) %>%
  ggplot(aes(period, performance)) +
  geom_hline(yintercept = 0.95, linetype = "dashed") +
  geom_line() +
  geom_point() +
  scale_y_continuous(labels = percent) +
  labs(title = "4 Hour performance over time")

# Now produce a plot showing the performance of each trust
ae_attendances %>%
  group_by(org_code) %>%
  # select organisations that have a type 1 department
  filter(any(type == "1")) %>%
  summarise_at(vars(attendances, breaches), sum) %>%
  arrange(desc(attendances)) %>%
  mutate(performance = 1 - breaches / attendances,
         overall_performance = 1 - sum(breaches) / sum(attendances),
         rank = rank(-performance, ties.method = "first") / n()) %>%
  ggplot(aes(rank, performance)) +
  geom_vline(xintercept = c(0.25, 0.5, 0.75), linetype = "dotted") +
  geom_hline(yintercept = 0.95, colour = "red") +
  geom_hline(aes(yintercept = overall_performance), linetype = "dotted") +
  geom_point() +
  scale_y_continuous(labels = percent) +
  theme_minimal() +
  theme(panel.grid = element_blank(),
        axis.text.x = element_blank()) +
  labs(title = "4 Hour performance by trust",
       subtitle = "Apr-16 through Mar-19",
       x = "", y = "")
```

**Description**

Artificially generated hospital data. Fictional patients at 10 fictional hospitals, with LOS, Age and Date status data Data were generate to learn Generalized Linear Models (GLM) concepts, modelling either Death or LOS.

**Usage**

```
data(LOS_model)
```

**Format**

Data frame with five columns

**ID** A fictional patient ID number

**Organisation** A factor representing one of ten fictional hospital trusts, e.g. Trust1

**Age** Age in years of each fictional patient

**LOS** In-hospital length of stay in days. The difference between admission and discharge date in dates

**Death** Binary for death status: 0 = survived, 1= died in hospital

**Source**

Generated by Chris Mainey <chris.mainey@uhb.nhs.uk>, Feb-2019

**Examples**

```
data(LOS_model)

model1 <- glm(Death ~ Age + LOS, data=LOS_model, family="binomial")
summary(model1)

# Now with an Age, LOS, and Age*LOS interaction.
model2<- glm(Death ~ Age * LOS, data=LOS_model, family="binomial")
summary(model2)
```

---

ons\_mortality

*Deaths registered weekly in England and Wales, provisional*

---

**Description**

Provisional counts of the number of deaths registered in England and Wales, by age, sex and region, in the latest weeks for which data are available.

**Usage**

```
data(ons_mortality)
```

**Format**

Data frame with five columns

**category\_1** character, containing the names of the groups for counts, e.g. "Total deaths", "all ages".

**category\_2** character, subcategory of names of groups where necessary, e.g. details of region: "East", details of age bands "15-44".

**counts** numeric, numbers of deaths in whole numbers and average numbers with decimal points.  
To retain the integrity of the format this column data is left as character.

**date** date, format is yyyy-mm-dd; all dates are a Friday.

**week\_no** integer, each week in a year is numbered sequentially.

**Details**

Source and licence acknowledgement This data has been made available through Office of National Statistics under the Open Government Licence <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>

**Source**

Collected by Zoë Turner <zoe.turner2@nottshc.nhs.uk>, Apr-2020 from <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfigures>

**Examples**

```
data(ons_mortality)

library(dplyr)
library(tidyr)

wideForm <- ons_mortality %>%
  select(-week_no) %>%
  pivot_wider(names_from = date,
              values_from = counts
  )
```

---

stranded_data	<i>Stranded Patient (Patients flagged as having a greater than 7 day LOS)</i> <i>Model</i>
---------------	---

---

**Description**

This model is to be used as a machine learning classification model, for supervised learning. The binary outcome is stranded vs not stranded patients.

**Usage**

```
data(stranded_data)
```

**Format**

Tibble with nine columns (1 x outcome and 8 predictors)

**stranded.label** Outcome variable - whether the patient is stranded or not

**age** Patient age on admission

**care.home.referral** Whether than have been referred from a care home

**medicallysafe** Medically safe for discharge - means the patient is assessed as safe, but has not been discharged yet

**hcop** Indicates whether they have been triaged from a Health Care for Older People specialty

**mental\_health\_care** Flag to indicate whether they need mental health support and care

**periods\_of\_previous\_care** Count of the number of previous spells of care

**admit\_date** Date they were admitted to hospital

**frailty\_index** An initial index assessment to say if the patient is frail or not. This is needed for alignment of service provision.

**Source**

Synthetically generated by Gary Hutson <g.hutson@nhs.net>, Mar-2021.

**Examples**

```
library(magrittr)
library(dplyr)
data("stranded_data")
stranded_data %>%
  glimpse()
```

---

synthetic_news_data	<i>Synthetic National Early Warning Scores Data</i>
---------------------	---

---

**Description**

Synthetic NEWS data to show as the results of the NHSR\_synpop package. These datasets have been synthetically generated by this package to be utilised in the NHSRDatasets package.

**Usage**

```
data(synthetic_news_data)
```

**Format**

Tibble with twelve columns

**male** character string containing gender code

**age** age of patient

**NEWS** National Early Warning Score (NEWS)

**syst** Systolic BP - Systolic BP result

**dias** Diastolic Blood Pressure - result on NEWS scale

**temp** Temperature of patient

**pulse** Pulse of the patient

**resp** Level of response from the patient

**sat** SATS(Oxygen Saturation Levels) of the patient

**sup** Suppressed Oxygen score

**alert** Level of alertness of patient

**died** Indicator to monitor patient death

**Source**

Generated by Dr. Muhammed Faisal and created by Gary Hutson <g.hutson@nhs.net>, Mar-2021

**Examples**

```
library(magrittr)
library(dplyr)
data("synthetic_news_data")
synthetic_news_data %>%
  glimpse()
```

# Index

- \* **England**
  - ons\_mortality, [4](#)
- \* **Provisional**
  - ons\_mortality, [4](#)
- \* **Wales**
  - ons\_mortality, [4](#)
- \* **a&e**
  - ae\_attendances, [2](#)
- \* **datasets**
  - ae\_attendances, [2](#)
  - LOS\_model, [3](#)
  - ons\_mortality, [4](#)
  - stranded\_data, [5](#)
  - synthetic\_news\_data, [6](#)
- \* **deaths**
  - ons\_mortality, [4](#)
- \* **hospital**
  - ae\_attendances, [2](#)
  - LOS\_model, [3](#)
- \* **mortality**
  - ons\_mortality, [4](#)
- \* **regression**
  - LOS\_model, [3](#)
- \* **stranded\_model**
  - stranded\_data, [5](#)
- \* **synthetic\_news\_data**
  - synthetic\_news\_data, [6](#)

ae\_attendances, [2](#)

LOS\_model, [3](#)

ons\_mortality, [4](#)

stranded\_data, [5](#)

synthetic\_news\_data, [6](#)