Package ‘ukc19’

December 19, 2025

Title Datasets from the UK COVID-19 Outbreak

Version 0.0.3

Description Provides easy access to a curated selection of pre-processed data sets relevant to the COVID-19 outbreak in the UK for teaching and demonstration purposes.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.3.9007

Depends R (>= 3.5)

LazyData true

Language en-GB

URL `https://ai4ci.github.io/ukc19/`, `https://github.com/ai4ci/ukc19`

Imports dplyr

NeedsCompilation no

Author Robert Challen [aut, cre] (ORCID: <`https://orcid.org/0000-0002-5504-7768`>),
AI4CI Hub; UKRI AI Programme and EPSRC (EP/Y028392/1) [fnd, cph] (url: https://gtr.ukri.org/projects?ref=EP%2FY028392%2F1)

Maintainer Robert Challen <`rob.challen@bristol.ac.uk`>

Repository CRAN

Date/Publication 2025-12-19 15:20:02 UTC

Contents

`covid_challenge` *COVID-19 viral load following challenge*

Description

Viral load from nasal swabs of subset of positive participants from COVID-19 human challenge study, as detected by Quantitative PCR. Values were mined from the vector files of the figures. The Y-axis values are approximate as had to be manually read from the scale.

Usage

```
data("covid_challenge")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 629 rows and 3 columns.

Details

Data extracted from Killingley et al, 2022, figure 2 "Viral shedding after a short incubation period peaks rapidly after human SARS-CoV-2 challenge". Panel A (middle left sub panel).

For datasets compiled from existing literature, Scientific Data's policy is that compilers (creators of the secondary compilation dataset and authors of the associated Data Descriptor) are not required by the journal to ask permission from the original authors to extract small amounts of numerical information or other fields. Expected practice is to attribute the original work via citation.

`id` **(chr)**	id a unique ID for participant
`log10_viral_load` **(dbl)**	log 10 viral load in copies per millilitre detected
`time` **(dbl)**	time of the sample in days from exposure.

Source

https://www.nature.com/articles/s41591-022-01780-9/figures/2

References

B. Killingley et al., ‘Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge in young adults’, Nat Med, vol. 28, no. 5, pp. 1031–1041, May 2022, doi: 10.1038/s41591-022-01780-9.

Examples

```
dplyr::glimpse(covid_challenge)
```

`covid_variants` *COG-UK counts of genomic variants*

Description

Weekly counts of identified variants for the whole of England.

Usage

```
data("covid_variants")
```

Format

An object of class `grouped_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 479 rows and 5 columns.

Details

Counts of COVID-19 variants from the COGUK COVID-19 sequencing project. Positive samples were selected based on viral load on initial PCR testing and sent onward for testing. Prioritisation and over-sampling of cases with S-gene target failure happened so this data is not unbiased.

From late March 2023 onward, due to the low number of sequenced samples, the UK SARS-CoV-2 sequencing surveillance data is not updated on the Wellcome Sanger Institute COVID-19 Genomic surveillance dashboard. Due to changes since the end of mass COVID-19 testing in the UK since April 2022 - the Wellcome Sanger Institute COVID-19 Genomic surveillance dashboard only includes a subset of UK SARS-CoV-2 sequencing surveillance data and should not be used to estimate frequency of SARS-CoV-2 variants circulating. Not all samples sequenced and deposited in public databases are presented here. This data is not de-duplicated on a patient level - and may include targeted sequencing that may introduce biases.

`covid_variants` **dataframe with 479 rows and 5 columns:**

`date` **(date)**	The date - unclear if this was of the sample or result
`class` **(fct)**	The variant description as a name and pango lineage
`who_class` **(fct)**	The WHO short name
`count` **(dbl)**	The number of sequences of this variant identified on this date
`denom` **(dbl)**	The total number of sequences of all variants identified on this date

Source

https://covid19.sanger.ac.uk/lineages/raw Contains Ordnance Survey data © Crown copyright and database right 2019 Contains UK Health Security Agency data © Crown copyright and database right 2020 Office for National Statistics licensed under the Open Government Licence v.3.0

Examples

```
dplyr::glimpse(covid_variants)
```

`covid_variants_ltla` *COG-UK counts of genomic variants by lower tier local authority*

Description

Counts of COVID-19 variants from the COGUK COVID-19 sequencing project. Positive samples were selected based on viral load on initial PCR testing and sent onward for testing. Prioritisation and over-sampling of cases with S-gene target failure happened so this data is not unbiased.

Usage

```
data("covid_variants_ltla")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 55785 rows and 8 columns.

Details

Weekly counts of identified variants by Lower tier local authority (2019 names) This dataset has implicit zeros. The full range of areas can be got from the `geography` data set with: `geography %>% dplyr::filter(codeType == "LAD19")`

From late March 2023 onward, due to the low number of sequenced samples, the UK SARS-CoV-2 sequencing surveillance data is not updated on the Wellcome Sanger Institute COVID-19 Genomic surveillance dashboard. Due to changes since the end of mass COVID-19 testing in the UK since April 2022 - the Wellcome Sanger Institute COVID-19 Genomic surveillance dashboard only includes a subset of UK SARS-CoV-2 sequencing surveillance data and should not be used to estimate frequency of SARS-CoV-2 variants circulating. Not all samples sequenced and deposited in public databases are presented here. This data is not de-duplicated on a patient level - and may include targeted sequencing that may introduce biases.

`covid_variants_ltla` **dataframe with 55785 rows and 8 columns:**

`date` **(date)**	The date - unclear if this was of the sample or result
`code` **(chr)**	The ONS geographical region code
`codeType` **(chr)**	The type of ONS geographical code

`name` **(chr)** The ONS geographical region name

`who_class` **(fct)** The WHO short name

`count` **(dbl)** The number of sequences of this variant identified on this date

`denom` **(dbl)** The total number of sequences of all variants identified on this date

Source

https://covid19.sanger.ac.uk/lineages/raw Contains Ordnance Survey data © Crown copyright and database right 2019 Contains UK Health Security Agency data © Crown copyright and database right 2020 Office for National Statistics licensed under the Open Government Licence v.3.0

Examples

```
dplyr::glimpse(covid_variants_ltla)
```

`du_serial_interval` *Serial interval from publicly reported cases*

Description

Data on which initial serial interval estimates were performed by Du et al, 2020.

Usage

```
data("du_serial_interval")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 752 rows and 3 columns.

Details

"This is a publication of the U.S. Government. This publication is in the public domain and is therefore without copyright. All text from this work may be reprinted freely. Use of these materials should be properly cited."

`du_serial_interval` **dataframe with 752 rows and 3 columns:**

`id` **(dbl)** Unique case id

`symptom_onset` **(dbl)** Time of symptom onset as an integer

`infector_id` **(dbl)** Case id of infector where known

Source

https://github.com/MeyersLabUTexas/COVID-19

References

Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers, 'Serial Interval of COVID-19 among Publicly Reported Confirmed Cases', Emerg Infect Dis, vol. 26, no. 6, pp. 1341–1343, Jun. 2020, doi: 10.3201/eid2606.200357.

Examples

```
dplyr::glimpse(du_serial_interval)
```

`early_global_combined` *John Hopkins data from the early outbreak*

Description

Mined out the commit history of COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University this dataset has early outbreak trajectories (21st Jan 2020 up to March 8th 2020) for a wide range of geographies, for confirmed cases, deaths and recovered cases. These trajectories are based on reported date, but are occasionally revised which will vary from region to region and maybe between different statistics, which show up as infrequent changes in published estimates over time.

Usage

```
data("early_global_combined")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 104036 rows and 9 columns.

Details

This data set is originally licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) by the Johns Hopkins University on behalf of its Center for Systems Science in Engineering. Copyright Johns Hopkins University 2020.

`country` **(chr)**	The country
`province` **(chr)**	Sub-national division
`lat` **(dbl)**	Latitude
`long` **(dbl)**	Longitude
`reported_date` **(date)**	Date of the observation based on reports of cases on this date.
`total_cases` **(dbl)**	Cumulative cases
`published_date` **(date)**	Date the observation was published on the JHU github.
`total_deaths` **(dbl)**	Cumulative deaths
`total_recovered` **(dbl)**	Cumulative recovered

Source

https://github.com/CSSEGISandData/COVID-19

Examples

```
dplyr::glimpse(early_global_combined)
```

`england_cases_by_5yr_age`

England only COVID-19 case counts stratified by 5-year age bands

---

Description

A dataset of the daily count of COVID-19 cases by age group in England downloaded from the UKHSA coronavirus API, and formatted for use in `ggoutbreak`. A denominator is calculated which is the overall positive count for all age groups. This data set can be used to calculate group-wise incidence and absolute growth rates and group wise proportions and relative growth rates by age group.

Usage

```
data("england_cases_by_5yr_age")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 26790 rows and 8 columns.

Details

You may want `england_covid_positivity` instead which includes the test denominator. The denominator here is the total number of positive tests across all age groups and not the number of tests taken or population size.

`england_cases_by_5yr_age` **dataframe with 26790 rows and 8 columns:**

`name` **(chr)** The region name

`code` **(chr)** The region code

`codeType` **(chr)** The ONS geographical region code type (including year)

`date` **(date)** The date

`class` **(chr)** the age group in 5 year age bands

`count` **(dbl)** the test positives for each age group

`denom` **(dbl)** the test positives across all age groups

`population` **(dbl)** the population size for this age group

Source

https://ukhsa-dashboard.data.gov.uk/covid-19-archive-data-download

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(england_cases_by_5yr_age)
```

`england_covid_positivity`

England only COVID-19 case counts with total test numbers

Description

The daily count of COVID-19 new PCR positive cases in England. The denominator the overall number of PCR tests conducted. This gives us a proportion of positive tests which can be used to correct for testing effort.

Usage

```
data("england_covid_positivity")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 1413 rows and 6 columns.

Details

`england_covid_positivity` **dataframe with 2048 rows and 6 columns:**

`name` **(chr)**	The region name
`code` **(chr)**	The region code
`codeType` **(chr)**	The ONS geographical region code type (including year)
`date` **(date)**	The date
`count` **(dbl)**	the count of PCR test positives
`denom` **(dbl)**	the total count of PCR tests conducted on that day

Source

`https://ukhsa-dashboard.data.gov.uk/covid-19-archive-data-download`

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(england_covid_positivity)
```

`ganyani_clusters` *COVID-19 cluster outbreaks data from Tianjin and Singapore*

Description

Data from which serial interval and generation time estimates were performed by Ganyani et al, 2020

Usage

```
data("ganyani_clusters")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 196 rows and 6 columns.

Details

Original article licensed under Creative Commons 4.0. Data was cleansed and formatted for R.

`ganyani_clusters` **dataframe with 196 rows and 6 columns:**

`id` **(dbl)**	a unique id for a person (unique within the `source`)
`contacts` **(list dbl)**	list of known contacts in the cluster
`cluster_id` **(dbl)**	id of a cluster (unique within the `source`)
`symptom_onset` **(date)**	symptom onset date
`known_primary_case` **(lgl)**	flag if this person is know to be the primary case in the cluster
`source` **(chr)**	geographical source of the data

Source

`https://github.com/cecilekremer/COVID19`

References

Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, Hens N. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Euro Surveill. 2020 Apr;25(17):2000257. doi: 10.2807/1560-7917.ES.2020.25.17.2000257. PMID: 32372755; PMCID: PMC7201952.

Examples

```
dplyr::glimpse(ganyani_clusters)
```

`geography` *UK geographic codes an CTRY, RGN and LAD level*

Description

Geographic codes and names from the ONS for administrative regions of the UK relevant to the COVID-19 response. There are multiple entries for lower tier local authority codes as these changed during the course of the pandemic.

Usage

```
data("geography")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 1512 rows and 3 columns.

Details

`geography` **dataframe with 1512 rows and 3 columns:**

`name` **(chr)**	The region name
`code` **(chr)**	The region code
`codeType` **(chr)**	The ONS geographical region code type (including year)

Source

`https://geoportal.statistics.gov.uk/`

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(geography)
```

`ltla_cases` *UK-wide COVID-19 case counts stratified by Lower tier local authority*

Description

A dataset of the daily count of COVID-19 cases by Lower tier local authority in the UK downloaded from the UKHSA coronavirus API, and formatted for use in `ggoutbreak`.

Usage

```
data("ltla_cases")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 512050 rows and 6 columns.

Details

`ltla_cases` **dataframe with 512050 rows and 6 columns:**

`name` **(chr)**	The region name
`code` **(chr)**	The region code
`codeType` **(chr)**	The ONS geographical region code type (including year)
`date` **(date)**	The date
`count` **(dbl)**	the test positives for each LTLA
`population` **(dbl)**	the population size for this geography

Source

`https://ukhsa-dashboard.data.gov.uk/covid-19-archive-data-download`

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(ltla_cases)
```

`nhs_app` *NHS digital contact tracing activity*

Description

Summary data collected as part of the NHS digital contact tracing app monitoring. This describes the number of alerts issued, and venue "check-ins".

Usage

```
data("nhs_app")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 137 rows and 3 columns.

Details

`date` **(date)**	The date
`alerts` **(int)**	Number of alerts
`visits` **(int)**	Number of check-ins

Source

https://www.gov.uk/government/publications/nhs-covid-19-app-statistics

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(nhs_app)
```

`ons_infection_survey` *ONS COVID-19 infection survey*

Description

The COVID-19 ONS infection survey took a random sample of the population and provides an estimate of the prevalence of COVID-19 that is theoretically free from ascertainment bias. This data set is the output of the model based on underlying data.

Usage

```
data("ons_infection_survey")
```

Format

An object of class `grouped_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 9820 rows and 8 columns.

Details

- `code` **(chr)** The ONS geographical region code
- `codeType` **(chr)** The type of ONS geographical code
- `name` **(chr)** The ONS geographical region name
- `date` **(date)** A date
- `prevalence.0.5` **(dbl)** the median proportion of people in the region testing positive for COVID-19
- `prevalence.0.025` **(dbl)** the lower CI of the proportion of people in the region testing positive for COVID-19
- `prevalence.0.975` **(dbl)** the upper CI of the proportion of people in the region testing positive for COVID-19
- `denom` **(int)** the sample size on which this estimate was made (daily rate inferred from weekly sample sizes.)

Source

https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/coronaviruscovid19infectionsurveydata

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(ons_infection_survey)
```

`pcr_test_sensitivity` *COVID PCR test sensitivity over time*

Description

Model output from Binny et al, 2023, describing the sensitivity of COVID PCR tests over the course of an infection.

Usage

```
data("pcr_test_sensitivity")
```

Format

An object of class `list` of length 2.

Details

`pcr_test_sensitivity` **named list with 2 items:**

`modelled` **(df modelled*)**	Original model output from supplementary
`resampled` **(df resampled*)**	resampled and reformatted data

`df modelled` **dataframe with 501 rows and 4 columns:**

`days_since_infection` **(dbl)**	days since infection
`median` **(dbl)**	median sensitivity
`lower_95` **(dbl)**	lower 95% CI of sensitivity
`upper_95` **(dbl)**	upper 95% CI of sensitivity

`df resampled` **dataframe with 5100 rows and 3 columns:**

`tau` **(dbl)**	days since infection
`probability` **(dbl)**	the sensitivity as a probability of detection
`boot` **(int)**	a bootstrap identifier

Source

https://pmc.ncbi.nlm.nih.gov/articles/instance/9796165/bin/jiac317_supplementary_data.zip

References

Rachelle N Binny, Patricia Priest, Nigel P French, Matthew Parry, Audrey Lustig, Shaun C Hendy, Oliver J Maclaren, Kannan M Ridings, Nicholas Steyn, Giorgia Vattiato, Michael J Plank, Sensitivity of Reverse Transcription Polymerase Chain Reaction Tests for Severe Acute Respiratory Syndrome Coronavirus 2 Through Time, The Journal of Infectious Diseases, Volume 227, Issue 1, 1 January 2023, Pages 9–17, https://doi.org/10.1093/infdis/jiac317

`spim_consensus` *SPI-M-O consensus reproduction number and growth rate estimates*

Description

A set of consensus estimates for the reproduction number and growth rate of the COVID-19 epidemic in England, produced by the SPI-M-O subgroup of SAGE

Usage

```
data("spim_consensus")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 113 rows and 5 columns.

Details

`spim_consensus_rt` **dataframe with 113 rows and 5 columns:**

`date` **(date)** the date

`rt.low` **(dbl)** the lower estimate of the reproduction number

`rt.high` **(dbl)** the upper estimate of the reproduction number

`growth.low` **(dbl)** the lower estimate of the exponential growth rate

`growth.high` **(dbl)** the higher estimate of the exponential growth rate

Source

`https://www.gov.uk/guidance/the-r-value-and-growth-rate`

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(spim_consensus)
```

`timeline` *Timeline of events*

Description

Major events in the UK COVID-19 pandemic, limited to lock-downs, vaccination roll-out and first identification of major variants.

Usage

```
data("timeline")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 19 rows and 3 columns.

Details

`label` **(chr)**	The event
`start` **(date)**	The start date
`end` **(date)**	The end date if a period

Source

https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_the_United_Kingdom

Examples

```
dplyr::glimpse(timeline)
```

`uk_population_2019` *Country, regional, and sub-national total population estimates*

Description

ONS National and sub-national mid-year population estimates for the UK and its constituent countries by administrative area, age and sex (including components of population change, median age and population density).

Usage

```
data("uk_population_2019")
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 398 rows and 4 columns.

Details

Mid-2019: April 2019 local authority district codes edition of this dataset. This is UK wide and covers country, regions and LTLA (2019 boundaries)

`uk_population_2019` **dataframe with 398 rows and 4 columns:**

`name` **(chr)**	The region name
`code` **(chr)**	The region code
`codeType` **(chr)**	The ONS geographical region code type (including year)
`population` **(dbl)**	the count of the population in that age group

Source

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(uk_population_2019)
```

`uk_population_2019_by_10yr_age`

Country, regional, and sub-national population estimates by 10 year age groups

---

Description

ONS National and sub-national mid-year population estimates for the UK and its constituent countries by administrative area, age and sex (including components of population change, median age and population density).

Usage

```
data("uk_population_2019_by_10yr_age")
```

Format

An object of class `grouped_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 3980 rows and 6 columns.

Details

Mid-2019: April 2019 local authority district codes edition of this dataset, this is UK wide and covers country, regions and LTLA (2019 boundaries)

Stratified by 10 year age groups

`uk_population_2019_by_10yr_age` **dataframe with 3980 rows and 6 columns:**

`name` **(chr)** The region name

`code` **(chr)** The region code

`codeType` **(chr)** The ONS geographical region code type (including year)

`class` **(chr)** The age group in 10 year age bands

`population` **(dbl)** the count of the population in that age group

`baseline_proportion` **(dbl)** the proportion of the total regional population that is in an age group

Source

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(uk_population_2019_by_10yr_age)
```

`uk_population_2019_by_5yr_age`

Country, regional, and sub-national population estimates by 5 year age groups

---

Description

ONS National and sub-national mid-year population estimates for the UK and its constituent countries by administrative area, age and sex (including components of population change, median age and population density).

Usage

```
data("uk_population_2019_by_5yr_age")
```

Format

An object of class `grouped_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 7562 rows and 6 columns.

Details

Mid-2019: April 2019 local authority district codes edition of this dataset, this is UK wide and covers country, regions and LTLA (2019 boundaries)

Stratified by 5 year age groups

`uk_population_2019_by_5yr_age` **dataframe with 7562 rows and 6 columns:**

- `name` **(chr)** The region name
- `code` **(chr)** The region code
- `codeType` **(chr)** The ONS geographical region code type (including year)
- `class` **(chr)** The age group in 5 year age bands
- `population` **(dbl)** the count of the population in that age group
- `baseline_proportion` **(dbl)** the proportion of the total regional population that is in an age group

Source

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

Originally licensed under the Open Government Licence v3.0

Examples

```
dplyr::glimpse(uk_population_2019_by_5yr_age)
```

`viral_shedding` *COVID-19 Viral shedding data*

Description

Data from van Kampen et al, 2019, describing duration of viral shedding from symptom onset in patients with COVID-19.

Usage

```
data("viral_shedding")
```

Format

An object of class `list` of length 2.

Details

`viral_shedding` **named list with 2 items:**

`original` **(df original*)**	original description
`resampled` **(df resampled*)**	resampled description

`df original` **dataframe with 690 rows and 4 columns:**

`duration of symptoms in days` **(dbl)**	duration of symptoms in days
`RNA copies per mL` **(chr)**	RNA copies per mL
`PRNT titer` **(chr)**	PRNT titer
`virus culture result` **(chr)**	virus culture result

`df resampled` **dataframe with 2600 rows and 3 columns:**

`tau` **(int)**	time from symptom onset to measurement
`probability` **(dbl)**	probability of detected viral excretion
`boot` **(int)**	a bootstrap identifier

Source

https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-020-20568-4/MediaObjects/41467_2020_20568_MOESM4_ESM.xlsx

References

van Kampen, J.J.A., van de Vijver, D.A.M.C., Fraaij, P.L.A. et al. Duration and key determinants of infectious virus shedding in hospitalized patients with coronavirus disease-2019 (COVID-19). Nat Commun 12, 267 (2021). https://doi.org/10.1038/s41467-020-20568-4

Index