Package 'imprinting'

May 8, 2026

Type Package

Title Calculate Birth Year-Specific Probabilities of Immune Imprinting to Influenza

Version 0.1.1

Description Reconstruct birth-year specific probabilities of immune imprinting to influenza A, using the methods of Gostic et al. (2016) <doi:10.1126/science.aag1322>. Plot, save, or export the calculated probabilities for use in your own research. By default, the package calculates subtype-specific imprinting probabilities, but with user-provided frequency data, it is possible to calculate probabilities for arbitrary kinds of primary exposure to influenza A, including primary vaccination and exposure to specific clades, strains, etc.

License MIT + file LICENSE

Encoding UTF-8

Imports dplyr (>= 1.0.9), tidyr (>= 1.2.0), cowplot (>= 1.0.0), ggplot2, tidyselect

Depends R (>= 3.5.0)

Suggests knitr, rmarkdown, testthat (>= 3.0.0), devtools

Config/testthat/edition 3

RoxygenNote 7.2.2

VignetteBuilder knitr

URL https://cobeylab.github.io/imprinting/

BugReports https://github.com/cobeylab/imprinting/issues

NeedsCompilation no

Author Katelyn Gostic [aut], Alex Byrnes [ctb, cre]

Maintainer Alex Byrnes <abyrnes@uchicago.edu>

Repository CRAN

Date/Publication 2022-12-16 22:50:02 UTC

Contents

`get_country_cocirculation_data` *Get data on the relative circulation of each influenza A subtype*

Description

`get_country_cocirculation_data()` imports data on the fraction of influenza A cases in a specific country and year that were caused by each influenza A subtype (H1N1, H2N2, or H3N2), or group (group 1 or group 2). Group 1 contains H1N1 and H2N2, and group 2 contains H2N2.

Usage

```
get_country_cocirculation_data(
  country,
  max_year,
  min_samples = 30,
  output_format = "tibble"
)
```

Arguments

`country`	country of interest. Run `show_available_countries()` for a list of valid inputs.
`max_year`	last year of interest. Results will be generated from 1918:max_year.
`min_samples`	if fewer than `min_samples` (default 30) are reported in the country and year of interest, the function will substitute data from the corresponding WHO region.
`output_format`	can be 'tibble' (the default) or 'matrix' (used mainly for convenience within other functions)

Details

The data come from three sources:

- Historical assumptions: From 1918-1956, we assume only H1N1 circulated. From 1957-1967, we assume only H2N2 circulated. From 1968-1976, we assume only H3N2 circulated.
- Thompson et al. JAMA, 2003: From 1977-1996 we pull data on the relative dominance of H1N1 and H3N2 from Table 1 of Thompson et al. 2003, which reports surveillance data collected in the United States.
- From 1997-present, we pull in country or region-specific data from WHO Flu Mart on the fraction of specimens collected in routine influenza surveillance that test positive for each subtype. Country-specific data are the default. Regional data, then global data are used if the number of country or region-specific specimens is insufficient. `get_template_data()` imports the data for 1918-1996. `get_country_inputs_1997_to_present()` and `get_regional_inputs_1997_to_present()` import the data for 1997 on.

Value

A matrix with rows showing the calendar year, the fraction of influenza A-positive specimens of each subtype (rows `A/H1N1`, `A/H2N2`, and `A/H3N2`), and of each HA group (rows `group 1`, and `group 2`). Row `A` should always be 1, as it shows the sum of subtype-specific fractions. Row `B` is a placeholder whose values are all `NA`.

See Also

doi:10.1126/science.aag1322Gostic et al. Science, (2016) for detailed methods.

Examples

```
get_country_cocirculation_data("United States", "2019")
get_country_cocirculation_data("Laos", "2022", min_samples = 40)
```

`get_country_inputs_1997_to_present`
Return raw flu surveillance data for a specific country

Description

Load and return a tibble containing raw influenza surveillance data for the country of interest. Data are from WHO Flu Mart.

Usage

```
get_country_inputs_1997_to_present(country, max_year)
```

Arguments

country	name of country. Run show_available_countries() for a list of options.
max_year	results will be output for all available years up to max_year.

Value

A tibble with the following columns:

- Country: name of WHO region
- Year: calendar year of surveillance
- n_processed: total specimens processed

Examples

```
get_country_inputs_1997_to_present("Aruba", 1998)
get_country_inputs_1997_to_present("Honduras", 2022)
```

---

get_country_intensity_data

Get the relative intensity of influenza A circulation

Description

get_country_intensity data() returns data on the annual intensity of influenza circulation in each calendar year. Following doi:10.1126/science.aag1322Gostic et al. Science, (2016), we define 1 as the average intensity. Seasons with intensities greater than 1 have more flu A circulation than average, and seasons with intensities less than 1 are mild.

Usage

```
get_country_intensity_data(country, max_year, min_specimens = 50)
```

Arguments

country	country of interest. Run show_available_countries() for valid inputs.
max_year	last year of interest. Results will be generated from 1918:max_year.
min_specimens	if fewer than min_specimens (default 50) were tested in the country and year of interest, the function will substitute data from the corresponding WHO region.

Details

For 1918-1996, we use annual intensities from Gostic et al., Science, (2016). For 1997-present, we calculate country or region-specific intensities using surveillance data from WHO Flu Mart. Intensity is calculated as: [fraction of processed samples positive for flu A]/[mean fraction of processed samples positive for flu A]. Country-specific data are used by default. Regional or global data are substituted when country-specific data contain too few observations or fail quality checks. Global data are only used in years when regional data are insufficient.

Value

A tibble showing the year and intensity score.

`get_imprinting_probabilities` *Calculate imprinting probabilities*

Description

For each country and year of observation, calculate the probability that cohorts born in each year from 1918 through the year of observation imprinted to a specific influenza A virus subtype (H1N1, H2N2, or H3N2), or group (group 1 contains H1N1 and H2N2; group 2 contains H3N2).

Usage

```
get_imprinting_probabilities(
  observation_years,
  countries,
  annual_frequencies = NULL,
  df_format = "long"
)
```

Arguments

`observation_years`
: year(s) of observation in which to output imprinting probabilities. The observation year, together with the birth year, determines the birth cohort's age when calculating imprinting probabilities. Cohorts <=12 years old at the time of observation have some probability of being naive to influenza.

`countries`
: a vector of countries for which to calculate imprinting probabilities. Run `show_available_countries()` for a list of valid inputs, and proper spellings.

`annual_frequencies`
: an optional input allowing users to specify custom circulation frequencies for arbitrary types of imprinting in order to study, e.g. imprinting to specific strains, clades, or imprinting by vaccination. If nothing is input, the default is to calculate subtype-specific probabilities (possible imprinting types are A/H1N1, A/H2N2, A/H3N2, or naive). See Details.

`df_format`
: must be either 'long' (default) or 'wide'. Controls whether the output data frame is in long format (with a single column for calculated probabilities and a second column for imprinting subtype), or wide format (with four columns, H1N1, H2N2, H3N2, and naive) showing the probability of each imprinting status.

Details

Imprinting probabilities are calculated following doi:10.1126/science.aag1322Gostic et al. Science, (2016). Briefly, the model first calculates the probability that an individual's first influenza infection occurs 0, 1, 2, ... 12 years after birth using a modified geometric waiting time model. The annual circulation intensities output by `get_country_intensity_data()` scale the probability of primary infection in each calendar year.

Then, after calculating the probability of imprinting 0, 1, 2, ... calendar years after birth, the model uses data on which subtypes circulated in each calendar year (from `get_country_cocirculation_data()`) to estimate that probability that a first infection was caused by each subtype. See `get_country_cocirculation_data()` for details about the underlying data sources.

To calculate other kinds of imprinting probabilities (e.g. for specific clades, strains, or to include pediatric vaccination), users can specify custom circulation frequencies as a list, `annual_frequencies`. This list must contain one named element for each country in the `countries` input vector. Each list element must be a data frame or tibble whose first column is named "year" and contains numeric years from 1918:max(`observation_years`). Columns 2:N of the data frame must contain circulation frequencies that sum to 1 across each row, and each column must have a unique name indicating the exposure kind. E.g. column names could be "year", "H1N1", "H2N2", "H3N2", "vaccinated" to include probabilities of imprinting by vaccine, or "year", "3C.3A", "not_3C.3A" to calculate clade-specific probabilities. Do not include a naive column. Any number of imprinting types is allowed, but the code is not optimized to run efficiently when the number of categories is very large. Frequencies within the column must be supplied by the user. See Vieira et al. 2021 for methods to estimate circulation frequencies from sequence databases like GISAID or the NCBI Sequence Database.

See `vignette("custom-imprinting-types")` for use of a custom `annual_frequencies` input.

Value

- If `format=long` (the default), a long tibble with columns showing the imprinting subtype (H1N1, H2N2, H3N2, or naive), the year of observation, the country, the birth year, and the imprinting probability.
- If `format=wide`, a wide tibble with each row representing a country, observation year, and birth year, and with a column for each influenza A subtype (H1N1, H2N2, and H3N2), or the probability that someone born in that year remains naive to influenza and has not yet imprinted. For cohorts >12 years old in the year of observation, the probability of remaining naive is 0, and the subtype-specific probabilities are normalized to sum to 1. For cohorts <=12 years old in the year of observation, the probability of remaining naive is non-zero. For cohorts not yet born at the time of observation, all output probabilities are 0.

Examples

```
# ============================================================
# Get imprinting probabilities for one country and year
get_imprinting_probabilities(2022, "United States")
# ============================================================
# Return the same outputs in wide format
get_imprinting_probabilities(2022,
  "United States",
  df_format = "wide"
```

```
)
```

`get_p_infection_year` *Calculate the probability imprinting occurs n years after birth*

Description

Given an individual's birth year, the year of observation, and pre-calculated influenza circulation intensities, calculate the probability that the first influenza infection occurs exactly 0, 1, 2, ... 12 years after birth.

Usage

```
get_p_infection_year(
  birth_year,
  observation_year,
  intensity_df,
  max_year,
  baseline_annual_p_infection = 0.28
)
```

Arguments

`birth_year`	year of birth (numeric). Must be between 1918 and the current calendar year.
`observation_year`	year of observation, which affects the birth cohort's age.
`intensity_df`	data frame of annual intensities, output by `get_country_intensity_data()`.
`max_year`	maximum year for which to output probabilities. Must be greater than or equal to observation_year. (If in doubt, set equal to observation year.)
`baseline_annual_p_infection`	average annual probability of primary infection. The default, 0.28, was estimated using age-seroprevalence data in doi:10.1126/science.aag1322Gostic et al. Science, (2016).

Details

The probability of primary influenza infection n years after birth is calculated based on a modified geometric distribution: let p be the average annual probability of a primary influenza infection. Then the probability that primary infection occurs n=0,1,2,... years after birth is $p * (1-p)^n$.

This function modifies the geometric model above to account for changes in annual circulation intensity, so that annual probabilities of primary infection p_i are scaled by the intensity in calendar year i. Details are given in doi:10.1126/science.aag1322Gostic et al. Science, (2016).

Value

a vector whose entries show the probability that a person born in year 0 was first infected by influenza in year 0, 1, 2, 3, ...12 We only consider the first 13 probabilities (i.e. we assume everyone imprints before age 13. These outputs are not normalized, so the vector sum asymptotically approaches one, but is not exactly equal to one. For cohorts born <13 years prior to the year of observation, the output vector will have less than 13 entries.

Examples

```
# For a cohort under 12 years old and born in 2000, return the
# probabilities of primary infection in 2000, 2001, ... 2012:
get_p_infection_year(
  birth_year = 2000,
  observation_year = 2022,
  intensity_df = get_country_intensity_data("Canada", 2022),
  max_year = 2022
)

# If the cohort is still under age 12 at the time of observation, return
# a truncated vector of probabilities:
get_p_infection_year(
  birth_year = 2020,
  observation_year = 2022,
  intensity_df = get_country_intensity_data("Mexico", 2022),
  max_year = 2022
)
```

---

get_regional_inputs_1997_to_present

Return raw flu surveillance data for a specific WHO region

Description

Load and return a tibble containing raw influenza surveillance data, aggregated across all countries in the WHO region of interest. Data are from WHO Flu Mart.

Usage

```
get_regional_inputs_1997_to_present(region, max_year)
```

Arguments

region	name of WHO region. Run show_available_regions() for a list of options.
max_year	results will be output for all available years up to max_year.

Value

A tibble with the following columns:

- `WHOREGION`: name of WHO region.
- `Year`: calendar year .
- `n_H1N1`, `n_H2N2`, `n_H3N2`: number of influenza specimens that tested positive for each influenza A subtype.
- `n_A`: total specimens positive for influenza A (= `n_H1N1` + `n_H2N2` + `n_H3N2`).
- `n_BYam`, `n_BVic`: number of specimens positive for each lineage of influenza B: Victoria or Yamagata.
- `n_B`: total specimens positive for influenza B.
- `n_processed`: total specimens processed.

Examples

```
get_regional_inputs_1997_to_present("americas", 2017)
```

`get_template_data` *Get country-independent flu circulation data for 1918-1996*

Description

`get_template_data()` returns a tibble showing the fraction of influenza A cases caused by subtype H1N1, H2N2, or H3N2 in each year from 1918-1996. These data are country-independent. Country-specific data are only available from 1997 on.

- For years 1918-1976 only one influenza A subtype circulated, so all fractions are 0 or 1.
- From 1977-1996 H1N1 and H3N2 both circulated. `get_template_data()` reports the fraction of influenza A-positive specimens of each subtype observed in US flu surveillance. See Thompson et al. JAMA, 2003, Table 1.
- Country-specific data from WHO Flu Mart will be appended to this template in later steps.

Usage

```
get_template_data()
```

Value

A tibble with the following columns:

- year
- `A/H1N1`, `A/H2N2`, and `A/H3N2` show the fraction of influenza cases caused by each subtype.
- `A` = `A/H1N1` + `A/H2N2` + `A/H3N2`

- B is a placeholder for future calculate of influenza B imprinting probabilities, which currently contains NA.
- group1 and group2 show the fraction of cases caused by group 1 subtypes (H1N1 and H2N2), or group 2 (H3N2).
- data_from notes the data source.

See Also

doi:10.1126/science.aag1322Gostic et al. Science, (2016) for detailed methods.

`get_WHO_region` *Look up a country's WHO region*

Description

Look up a country's WHO region

Usage

```
get_WHO_region(this.country)
```

Arguments

this.country	name of the input country (a string)

Value

Name of the corresponding WHO region

Examples

```
get_WHO_region("Germany")
get_WHO_region("China")
```

plot_many_country_years *Plot imprinting probabilities for up to five country-years*

Description

For each country and year, generate two plots:

- A stacked barplot, where each bar represents a birth cohort, and the colors within the bar show the probabilities that someone born in that cohort has a particular imprinting status, for the first observation year.
- A lineplot showing the age-specific probability of imprinting to H3N2 in the first and last observation year. When the data contain more than one observation year, this plot shows how cohorts age over time.

Usage

```
plot_many_country_years(imprinting_df)
```

Arguments

imprinting_df	A long data frame of imprinted probabilities output by get_imprinting_probabilities(). Up to five countries and an arbitrary span of years can be plotted.

Value

No return value. Opens a plot of the data frame.

Examples

```
imprinting_df <- get_imprinting_probabilities(
  observation_years = c(1920, 1921),
  countries = c("Oman", "Indonesia")
)
plot_many_country_years(imprinting_df)
```

plot_one_country_year *Plot imprinting probabilities for a single country and year*

Description

Generate a stacked barplot, where each bar represents a birth cohort, and the colors within the bar show the probabilities that someone born in that cohort has a particular imprinting status. If the data frame contains more than one country or observation year, the first country-year is plotted by default. Specify other countries and years using the country and year options.

Usage

```
plot_one_country_year(imprinting_df, country = NULL, year = NULL)
```

Arguments

imprinting_df	A long data frame of imprinted probabilities output by get_imprinting_probabilities(). If the data frame contains more than one country and year, on the first will be plotted.
country	An optional country name to plot. The input country name must exist in the imprinting_df.
year	Similar to country, and optional input specifying the year for which to plot.

Value

No return value. Opens a plot of the data frame.

Examples

```
# Generate imprinting probabilities for one country and year
imprinting_df <- get_imprinting_probabilities(
  observation_years = 1920,
  countries = "Aruba"
)
plot_one_country_year(imprinting_df)

# If we generate probabilities for more than one country and year,
imprinting_df <- get_imprinting_probabilities(
  observation_years = c(1922, 1925),
  countries = c(
    "Algeria",
    "South Africa"
  )
)

# The default is to plot the first country year in the outputs
plot_one_country_year(imprinting_df)

# Or, specify a country and year of interest (both must exist in the
# imprinting_df).
plot_one_country_year(imprinting_df,
  country = "South Africa",
  year = 1925
)
```

show_available_countries

Show a list of all available countries

Description

Lists all available countries, with valid spelling and formatting. Each country in the list matches or can be mapped to a country with data in WHO Flu Mart. (Note: for convenience, this package sometimes uses different country names or spellings than Flu Mart.)

Usage

```
show_available_countries()
```

Value

A data frame of valid country names.

Examples

```
show_available_countries()
```

show_available_regions

Show all WHO regions

Description

Lists all WHO regions, with valid spelling and formatting.

Usage

```
show_available_regions()
```

Value

A data frame of valid region names.

Examples

```
show_available_regions()
```

Index