Tutorial: `data4health`

Aggregating health data

Before beginning to code, we always need to set up the R environment. This time, we will only load one R-package: *data4health*. Data4health is a package that we are developing at the moment. This means that the package is not yet published, you are among the first to use it!

But this also means, that there still be errors. Please let us know if you encounter any and we will fix them asap. Likewise, if you can think of anything missing or anything you would like to add, let us know too!

Thank you!

In [1]:
```
setwd("dependencies/ghr_libraries/harmonize.data4health")
source('./R/data4health_filter.r')
source('./R/data4health_load.R')
source('./R/data4health_aggregate.r')
```

Now, let's get started! First of all, you need to load the health data in. In this case it is a 'csv' file. You can either choose to use a csv specific function like read.csv or use the umbrella function from the data4health package. The data4health package currently loads in csv, ecxel, rds, and many more formats!

The object data, now contains ALL the clean data. But often, you don't want to use all the data. You can use the data4health_filter function for this.

In [2]:
```
data <- data.frame(
  ID = c(1, 2, 3, 4, 5),
  Name = c("Alice", "Bob", "Charlie", "David", "Emily"),
  Age = c(25, 30, 22, 28, 35),
  Date = as.Date(c("2023-01-15", "2023-02-20", "2023-03-05", "2023-04-10", "2023-05
  City = c("New York", "London", "Paris", "Tokyo", "Sydney"),
  Gender = c('Female', 'Male', 'Male', 'Male', 'Female')
```

```
)

print(head(data))
```

```
  ID    Name Age       Date     City Gender
1  1   Alice  25 2023-01-15 New York Female
2  2     Bob  30 2023-02-20   London   Male
3  3 Charlie  22 2023-03-05    Paris   Male
4  4   David  28 2023-04-10    Tokyo   Male
5  5   Emily  35 2023-05-18   Sydney Female
```

In [3]:
```
?data4health_filter
```

No documentation for 'data4health_filter' in specified packages and libraries: you could try '?? data4health_filter'

As you can see in the description for every column, you need to know whether the data is numeric, dates or character.

- Numeric: "over","under", "between"
- Date: "after", "before", "between"
- character: "include", "exclude"

You can filter as many or as little columns as you want. Here an example:

In [4]:
```
filtered_data <- data4health_filter(
  data,
  Age = list(over = 25),
  Date = list(between = c("2023-02-10", "2023-04-15"))
)

print(head(filtered_data))
```

```
  ID  Name Age       Date   City Gender
2  2   Bob  30 2023-02-20 London   Male
4  4 David  28 2023-04-10  Tokyo   Male
```

Afterwards we aggregate with the data4health_aggregate function.

In [5]:
```
data4health_aggregate(data, time_col = "Date", space_col = "Gender")
```

A data.frame: 10 × 3

Date	Gender	freq
<fct>	<fct>	<int>
2023-01-15	Female	1
2023-02-20	Female	0
2023-03-05	Female	0
2023-04-10	Female	0
2023-05-18	Female	1
2023-01-15	Male	0
2023-02-20	Male	1
2023-03-05	Male	1
2023-04-10	Male	1
2023-05-18	Male	0