

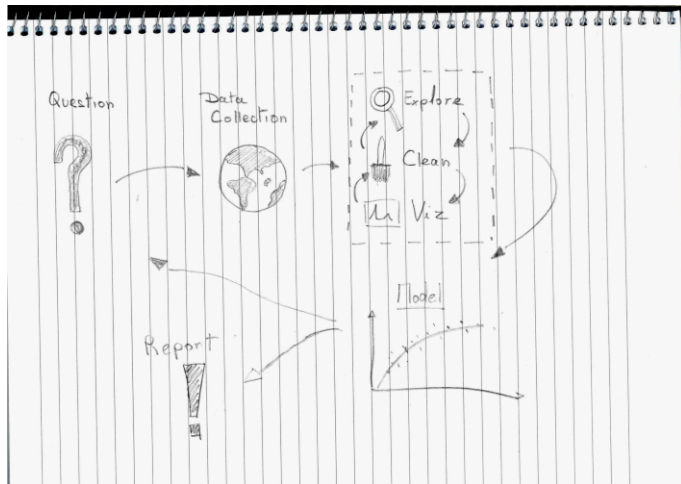
# Data manipulation with tidyverse

Damien Georges

International Agency for Research on Cancer

June 2023 - Tartu

# Epidemiological study workflow



# Data manipulation tools in R



- ▶ R core function
- ▶ tidyverse / dplyr
- ▶ data.table
- ▶ ...

=> The best tool is the one you feel the most comfortable with

# Tidyverse (from [www.tidyverse.org](http://www.tidyverse.org))

## R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.



pipe functions %>% or |>

```
chill(fold(add(melt(add(chocolate, butter)),  
               beat(add(eggs.white, cream))))
```

## pipe functions %>%

```
chill(fold(add(melt(add(chocolate, butter)),  
                beat(add(eggs.white, cream)))))
```

```
chocolate %>%  
  add(butter) %>%  
  melt() %>%  
  add(  
    eggs.white %>%  
      add(cream) %>%  
        beat()  
  ) %>%  
  fold() %>%  
  chill()
```

## Non-standard evaluation rules for function calls

- ▶ used in different R packages
- ▶ provide flexibility and ease of use
- ▶ more concise and expressive programming in R

```
dat <- data.frame(x = 1:10)
## subset supports NSE
subset(dat, x < 5)
## and SE
subset(dat, data$x < 5)
```

## data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)



## data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)
- ▶ select columns (`filter`)

## data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)
- ▶ select columns (`filter`)
- ▶ arrange rows (`arrange`)

# data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)
- ▶ select columns (`filter`)
- ▶ arrange rows (`arrange`)
- ▶ columns selection (`select`)

# data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)
- ▶ select columns (`filter`)
- ▶ arrange rows (`arrange`)
- ▶ columns selection (`select`)
- ▶ create/modify columns (`mutate`)

# data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)
- ▶ select columns (`filter`)
- ▶ arrange rows (`arrange`)
- ▶ columns selection (`select`)
- ▶ create/modify columns (`mutate`)
- ▶ group and summarize data (`group_by` and `summarise`)

# data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)
- ▶ select columns (`filter`)
- ▶ arrange rows (`arrange`)
- ▶ columns selection (`select`)
- ▶ create/modify columns (`mutate`)
- ▶ group and summarize data (`group_by` and `summarise`)
- ▶ bind different tables (`bind_rows`, `bind_cols`)

# data manipulation with



**Code as you speak:** Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.



- ▶ select rows (`slice`)
- ▶ select columns (`filter`)
- ▶ arrange rows (`arrange`)
- ▶ columns selection (`select`)
- ▶ create/modify columns (`mutate`)
- ▶ group and summarize data (`group_by` and `summarise`)
- ▶ bind different tables (`bind_rows`, `bind_cols`)
- ▶ merge different tables (`left_join`, `right_join`, `inner_join`, `full_join`)

## discovering other tidyverse packages features




- ▶ data visualization with  (ggplot, geom\_bars, ...)







## discovering other tidyverse packages features

- ▶ data visualization with  (ggplot, geom\_bars, ...)
- ▶ pivoting data with  (pivot\_wider, pivo\_longer)






## discovering other tidyverse packages features

- ▶ data visualization with  (ggplot, geom\_bars, ...)
- ▶ pivoting data with  (pivot\_wider, pivo\_longer)
- ▶ reading data with  (read\_table, read\_csv)

## discovering other tidyverse packages features

- ▶ data visualization with  (ggplot, geom\_bars, ...)
- ▶ pivoting data with  (pivot\_wider, pivo\_longer)
- ▶ reading data with  (read\_table, read\_csv)
- ▶ manipulating lists with  (map, map\_chr, reduce, ...)

## discovering other tidyverse packages features

- ▶ data visualization with  (ggplot, geom\_bars, ...)
- ▶ pivoting data with  (pivot\_wider, pivo\_longer)
- ▶ reading data with  (read\_table, read\_csv)
- ▶ manipulating lists with  (map, map\_chr, reduce, ...)
- ▶ manipulating strings with  (str\_length, str\_remove, ...)