World Health Organization
REGIONAL OFFICE FOR THE Eastern Mediterranean

منظمة الصحة العالمية
المكتب الإقليمي لشرق المتوسط

REGIONAL TRAINING ON CAPACITY DEVELOPMENT OF DATA ANALYTICS AND DISSEMINATION USING "R" SOFTWARE

AMMAN, JORDAN, 3 - 7 DECEMBER, 2023

Day #3

HEALTH FOR ALL BY ALL
a call for solidarity and action

World Health Organization
REGIONAL OFFICE FOR THE Eastern Mediterranean

Outline

- Wrap-up
- Data processing – part II
  - Working with dates
  - Factors
- Exporting cleaned data
- Data summarization
- Q&A

Session 3 Agenda

- 9:00 – 9: 30 (30 min): **Wrap-up**
- 9:30 – 9:50 (20 min): **Presentation "Data processing – part II"**
- 9:50 – 10:20 (30 min): **Demonstration**
- 10:20 – 10:40 (20 min): **Stretching / coffee break**
- 10:40 – 12:30 (1.8 hr): **Practice/Exercise**
- 12:30 – 13:00 (30 min): **Quick debrief/ Q&A**
- 13:00 – 14:00 (60 min): **Lunch**
- 14:00 – 14:20 (20 min): presentation "Data summarization"
- 14:20 – 14:50 (30 min): Demonstration
- 14:50 – 15:10 (20 min): Stretching / coffee break
- 15:10 – 16:30 (80 min): Practice/Exercise
- 16:30 – 17:00 (30 min): Quick debrief/ Q&A

Working with dates in R

- The end goal is class () of the column is "**date**"
- The standard format of date in R is

  "**YYYY – MM – DD**"

- **Transform character to date class**
  - base R function: as.Date()
  - lubridate::ymd(), mdy(), dmy()
- **Transform numeric to date class**
  - as.Date()

```
> date <- "3-30-2023"
> date
[1] "3-30-2023"
> class(date)
[1] "character"
> date <- as.Date(date, format = "%m-%d-%Y")
> date
[1] "2023-03-30"
> class(date)
[1] "Date"
> date <- lubridate::mdy(date)
> date
[1] "2023-03-30"
> class(date)
[1] "Date"
>
```

```
> date <- 43215
> class(date)
[1] "numeric"
> date <-  as.Date(date, origin = "1899-12-30")
> date
[1] "2018-04-25"
> class(date)
[1] "Date"
```

HEALTH FOR ALL BY ALL a call for solidarity and action

World Health Organization
REGIONAL OFFICE FOR THE Eastern Mediterranean

Working with dates in R

- **Extracting date components**
  - We can extract any component of the date e.g year, month
  - lubridate::year(), month(), day()

```
> date
[1] "2023-03-30"
> month(date)
[1] 3
> year(date)
[1] 2023
> day(date)
[1] 30
```

- **Creating epi week of specific date**
  - aweek::date2week()
  - lubridate::epiweek()

```
mutate(cholera, epiweek= lubridate::epiweek(adm_date),
       epiweek2= aweek::date2week(adm_date , week_start = "sun", floor_day = T))
```

epiweek	epiweek2
31	2023-W31
27	2023-W27
29	2023-W29
30	2023-W30
30	2023-W30
29	2023-W29
27	2023-W27
24	2023-W24
25	2023-W25
22	2023-W22

Working with dates in R

- **strptime nomenclature for date display**

  The commonly used ones are:

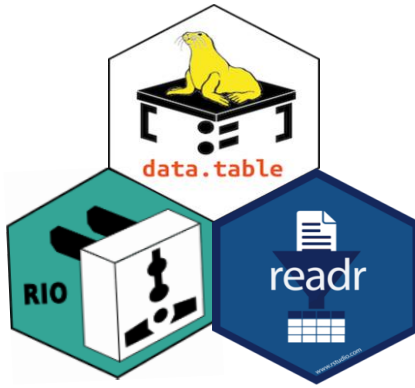
Symbol	Meaning
%d	Day number
%a	Abbreviated weekday
%m	Month number
%b	Abbreviated month name
%y	Year number (2 digits)
%Y	Year number (YYYY)
%U	Week number (start day: Sunday)
%W	Week number (start day: Monday)

Working with factors in R

- **Class factor (ordinal) of a column:**
  - Assign a specific order for the variable's values
  - very useful in plots and statistical tests
  - Base R function: factor()

```
> sex <- c(1,2,3,1,1,2,1,1,3)
> sex_f <- factor(sex, levels = c(1,2,3), labels = c("Female", "Male", "Unkown"))
> sex_f
[1] Female Male   Unkown Female Female Male   Female Female Unkown
Levels: Female Male Unkown
```

Exporting data



1. What is the name of the object?!!

2. What is the file format?!!

base R function

- write.csv ()

Installed packages

- readr::write_csv()
- rio::export()
- data.table::fwrite()

3. Where do you want to put that file?!!

Absolute path

Relative path

Demonstration

Exercise: Managing dates in R

- Open your training R project
- Open the R script “cholera.R”
- Add a new section “data processing 2” and do the following data steps:
    1. Convert “adm_date” column into class date
    2. Convert “outcome_date” column into class date
    3. Create a new column “epiweek” for admission date “adm_date” in format YYYY-W#
    4. Create a new column “epiweek_date” for the start day of the “epiweek”
    5. Create a new column “los” denoting length of stay [**HINT:** dates are stored as numbers!]
    6. Export the cleaned dataset “cholera_cleaned.csv” to the data folder

    + Bonus!!

    7. Create a new column “sym_to_adm” denoting symptom onset to admission
    8. Create a new column “sex_f” as factor

Creating tables to summarize data
&
Data Visualization

HEALTH FOR ALL BY ALL
a call for solidarity and action

World Health Organization
REGIONAL OFFICE FOR THE Eastern Mediterranean

Session 3 Agenda

- 9:00 – 9: 30 (30 min): Wrap-up
- 9:30 – 9:50 (20 min): Presentation “Data processing – part II”
- 9:50 – 10:20 (30 min): Demonstration
- 10:20 – 10:40 (20 min): Stretching / coffee break
- 10:40 – 12:30 (1.8 hr): Practice/Exercise
- 12:30 – 13:00 (30 min): Quick debrief/ Q&A
- 13:00 – 14:00 (60 min): Lunch
- 14:00 – 14:20 (20 min): **presentation “Data summarization”**
- 14:20 – 14:50 (30 min): **Demonstration**
- 14:50 – 15:10 (20 min): **Stretching / coffee break**
- 15:10 – 16:30 (80 min): **Practice/Exercise**
- 16:30 – 17:00 (30 min): **Quick debrief/ Q&A**

Elements of descriptive epidemiology

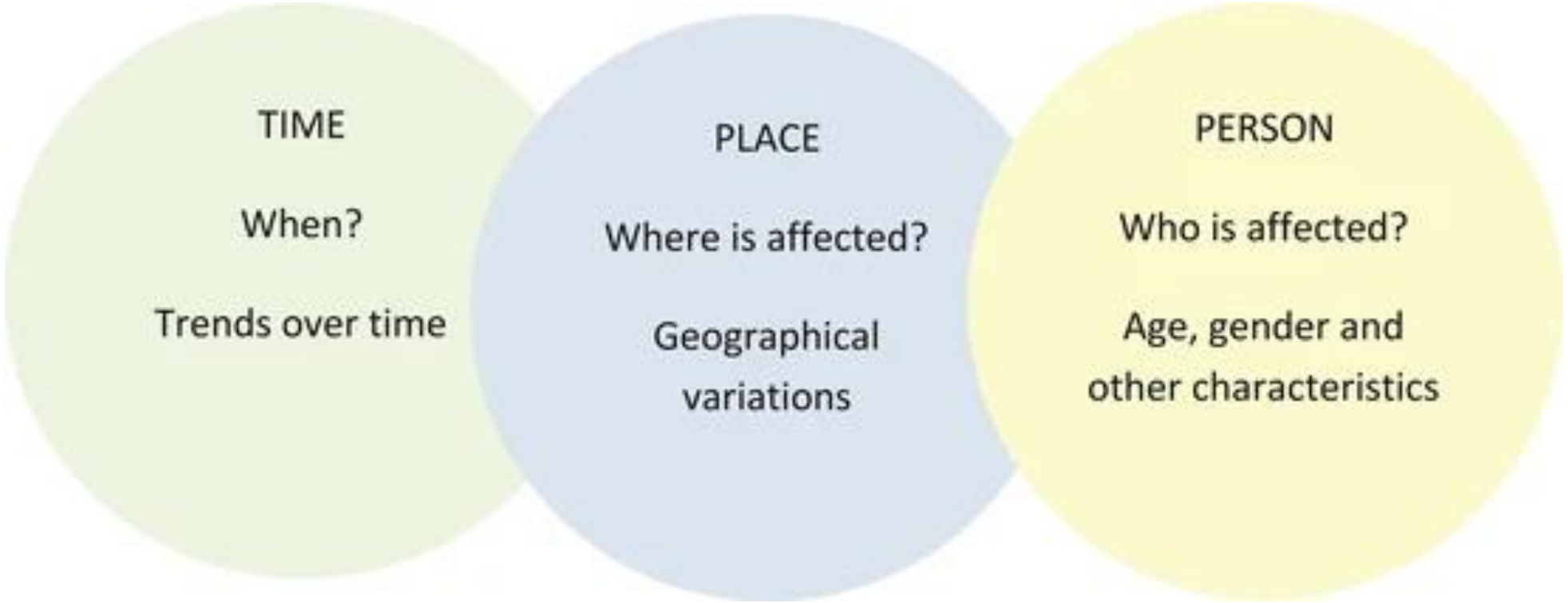


Image source: Epidemiology and public health intelligence

Creating tables

- useful format for comparative data analysis (explore the data)
- It can be many ways

label	variable	Severity - 4 cat					Total
		Mild	Moderate	Severe	Critical	Undefined	
death	0	34791 (c% 84.95% r% 24.31%)	7609 (c% 88.54% r% 5.32%)	2817 (c% 75.87% r% 1.97%)	41049 (c% 81.40% r% 28.68%)	56856 (c% 84.18% r% 39.73%)	143122 (83.58%)
	1	6166 (c% 15.05% r% 21.93%)	985 (c% 11.46% r% 3.50%)	896 (c% 24.13% r% 3.19%)	9381 (c% 18.60% r% 33.37%)	10686 (c% 15.82% r% 38.01%)	28114 (16.42%)
	Total	40957 (23.92%)	8594 (5.02%)	3713 (2.17%)	50430 (29.45%)	67542 (39.44%)	171236 (100.00%)

Which one, % by row or by col?

```
> crosstable(adm_out, c(death), by=c(severity_4cat_f), total="both", percent_pattern="{n}\n (c% {p_col} \n r% {p_row})") %>% flextable::as_flextable(keep_id=FALSE)
```

A tibble: 4 × 4

.id <chr>	label <chr>	variable <chr>	value <chr>
adm14	Admission date	Min / Max	2020-01-06 - 2023-01-11
adm14	Admission date	Med [IQR]	2020-11-15 [2020-04-15;2021-06-15]
adm14	Admission date	Mean (std)	2020-12-10 (7.3 months)
adm14	Admission date	N (NA)	103694 (0)

4 rows

```
> crosstable(adm_out, c(adm14))
```

label	variable	value
Discharge date	Min / Max	2020-01-02 - 2023-01-30
	Med [IQR]	2020-12-19 [2020-05-29;2021-07-19]
	Mean (std)	2021-01-09 (7.4 months)
	N (NA)	186729 (7475)

Length of hospital stay?

```
> crosstable(adm_out, c(dis47)) %>% flextable::as_flextable(keep_id=FALSE)
```

Creating tables

fio2	spo2	pao2	resp_rate	count
character	character	character	character	integer
No	No	No	No	640,311
No	No	No	Yes	105,655
No	No	Yes	No	102
No	No	Yes	Yes	9
No	Yes	No	No	2,636
No	Yes	No	Yes	57,850
No	Yes	Yes	No	21
No	Yes	Yes	Yes	57
Yes	No	No	No	23,413
Yes	No	No	Yes	52,652

n: 16

```
> tab <- adm_out %>%
group_by(fio2, spo2, pao2, resp_rate) %>%
summarise(count = n(), .groups = "drop")
```

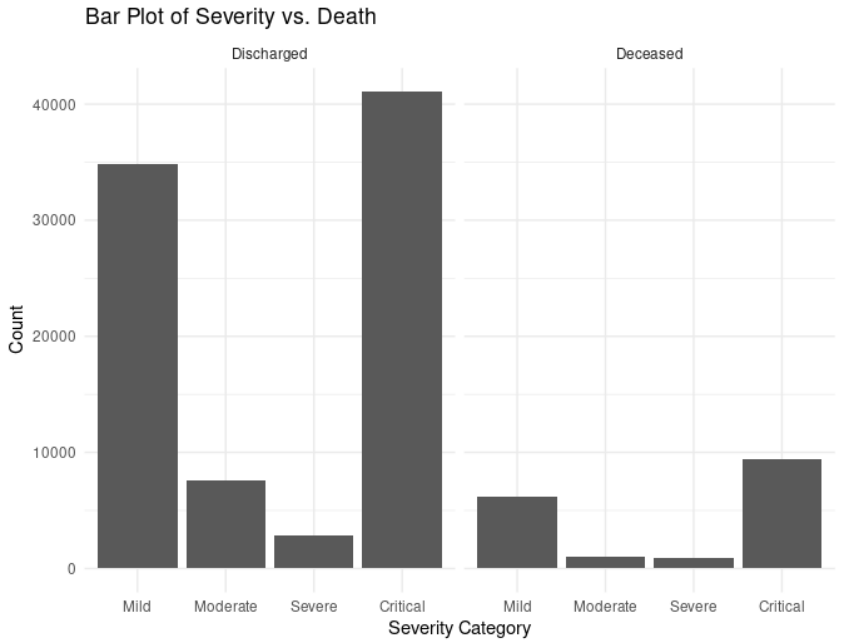
```
> write.csv2(tab, "oxyge.csv")
```

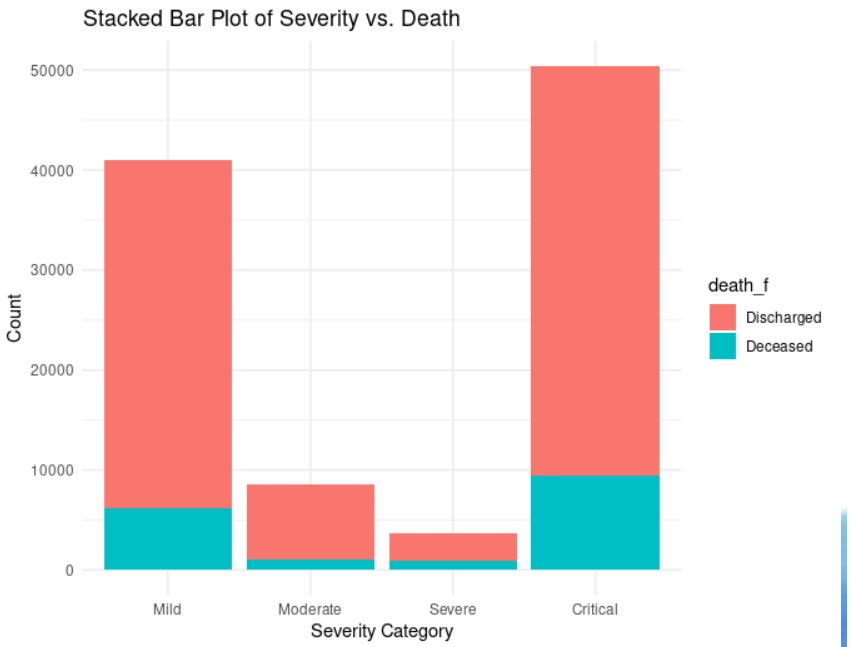
Export table as output for a report or as data for analysis in another software

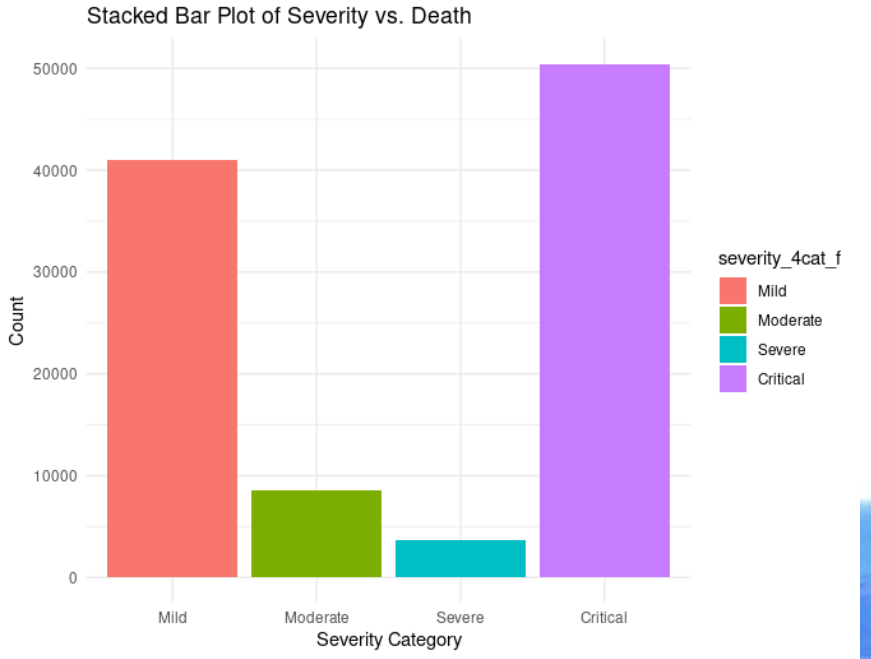
Tables & Graphs

Which graph? Why?

label	variable	Severity - 4 cat				Total	test
		Mild	Moderate	Severe	Critical		
Mortality	Discharged	34791 (c% 84.95% r% 40.33%)	7609 (c% 88.54% r% 8.82%)	2817 (c% 75.87% r% 3.27%)	41049 (c% 81.40% r% 47.58%)	86266 (83.19%)	
	Deceased	6166 (c% 15.05% r% 35.38%)	985 (c% 11.46% r% 5.65%)	896 (c% 24.13% r% 5.14%)	9381 (c% 18.60% r% 53.83%)	17428 (16.81%)	p value: <0.0001 (Pearson's Chi-squared test)
	Total	40957 (39.50%)	8594 (8.29%)	3713 (3.58%)	50430 (48.63%)	103694 (100.00%)	



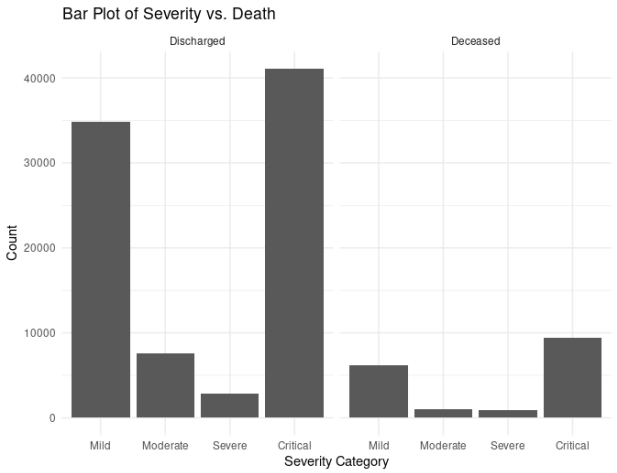


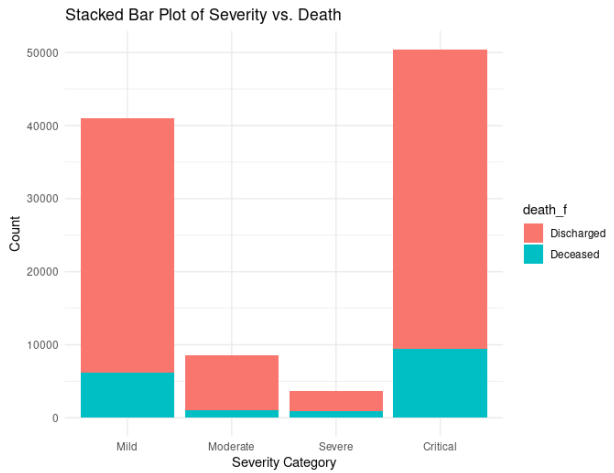


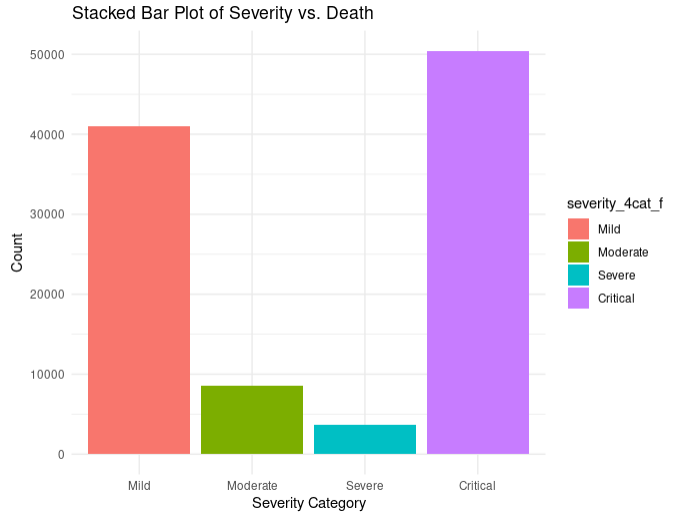
Tables & Graphs

Which
graph? Why?

label	variable	Severity - 4 cat				Total	test
		Mild	Moderate	Severe	Critical		
Mortality	Discharged	34791 (c% 84.95% r% 40.33%)	7609 (c% 88.54% r% 8.82%)	2817 (c% 75.87% r% 3.27%)	41049 (c% 81.40% r% 47.58%)	86266 (83.19%)	p value: <0.0001 (Pearson's Chi-squared test)
	Deceased	6166 (c% 15.05% r% 35.38%)	985 (c% 11.46% r% 5.65%)	896 (c% 24.13% r% 5.14%)	9381 (c% 18.60% r% 53.83%)	17428 (16.81%)	
	Total	40957 (39.50%)	8594 (8.29%)	3713 (3.58%)	50430 (48.63%)	103694 (100.00%)	








```
> ggplot(adm_out, aes(x = severity_4cat_f)) +
geom_bar() +
facet_grid(. ~ death_f) +
 labs(title = "Severity vs. Death", x =
"Severity", y = "Count") +
theme_minimal()
```

```
> ggplot(adm_out, aes(x = severity_4cat_f, fill
= death_f)) +
geom_bar() +
labs(title = "Severity vs. Death", x =
"Severity", y = "Count") +
theme_minimal()
```

```
> ggplot(adm_out, aes(x = severity_4cat_f, fill
= severity_4cat_f)) +
geom_bar() +
labs(title = "Severity vs. Death", x =
"Severity", y = "Count") +
theme_minimal()
```

What is different in the ggplot() code?

Tables & Graphs

label	variable	sex_f=Female		sex_f=Male		Total
		severity_f=Mild/Moderate	severity_f=Severe/Critical	severity_f=Mild/Moderate	severity_f=Severe/Critical	
Mortality	Discharged	31052 (c% 89.19% r% 36.04%)	9864 (c% 76.35% r% 11.45%)	32279 (c% 85.21% r% 37.46%)	12974 (c% 72.27% r% 15.06%)	86169 (83.20%)
	Deceased	3765 (c% 10.81% r% 21.64%)	3056 (c% 23.65% r% 17.56%)	5604 (c% 14.79% r% 32.20%)	4977 (c% 27.73% r% 28.60%)	17402 (16.80%)
	Total	34817 (33.62%)	12920 (12.47%)	37883 (36.58%)	17951 (17.33%)	103571 (100.00%)

Which one?
Why?

label	variable	severity_f=Mild/Moderate		severity_f=Severe/Critical		Total
		sex_f=Female	sex_f=Male	sex_f=Female	sex_f=Male	
Mortality	Discharged	31052 (c% 89.19% r% 36.04%)	32279 (c% 85.21% r% 37.46%)	9864 (c% 76.35% r% 11.45%)	12974 (c% 72.27% r% 15.06%)	86169 (83.20%)
	Deceased	3765 (c% 10.81% r% 21.64%)	5604 (c% 14.79% r% 32.20%)	3056 (c% 23.65% r% 17.56%)	4977 (c% 27.73% r% 28.60%)	17402 (16.80%)
	Total	34817 (33.62%)	37883 (36.58%)	12920 (12.47%)	17951 (17.33%)	103571 (100.00%)

HEALTH FOR ALL BY ALL
a call for solidarity and action

World Health Organization
REGIONAL OFFICE FOR THE Eastern Mediterranean

Tables

```
> table(adm_out$death)

    0     1
86266 17428
```

```
> crosstable(adm_out, c(severity_4cat))
```

```
# A tibble: 4 x 4
  .id           label         variable   value
  <chr>         <chr>         <chr>      <chr>
1 severity_4cat severity_4cat Min / Max  0 / 3.0
2 severity_4cat severity_4cat Med [IQR]  2.0 [0;3.0]
3 severity_4cat severity_4cat Mean (std) 1.6 (1.4)
4 severity_4cat severity_4cat N (NA)     103694 (0)
```

```
> crosstable(adm_out, c(severity_4cat)) %>%
flextable::as_flextable(keep_id=FALSE)
```

label	variable	value
severity_4cat	Min / Max	0 / 3.0
	Med [IQR]	2.0 [0;3.0]
	Mean (std)	1.6 (1.4)
	N (NA)	103694 (0)

```
> data %>%
dplyr::select(severity_4cat) %>%
tbl_summary()
```

Characteristic	N = 103,694[1]
severity_4cat	
0	40,957 (39%)
1	8,594 (8.3%)
2	3,713 (3.6%)
3	50,430 (49%)

[1] n (%)

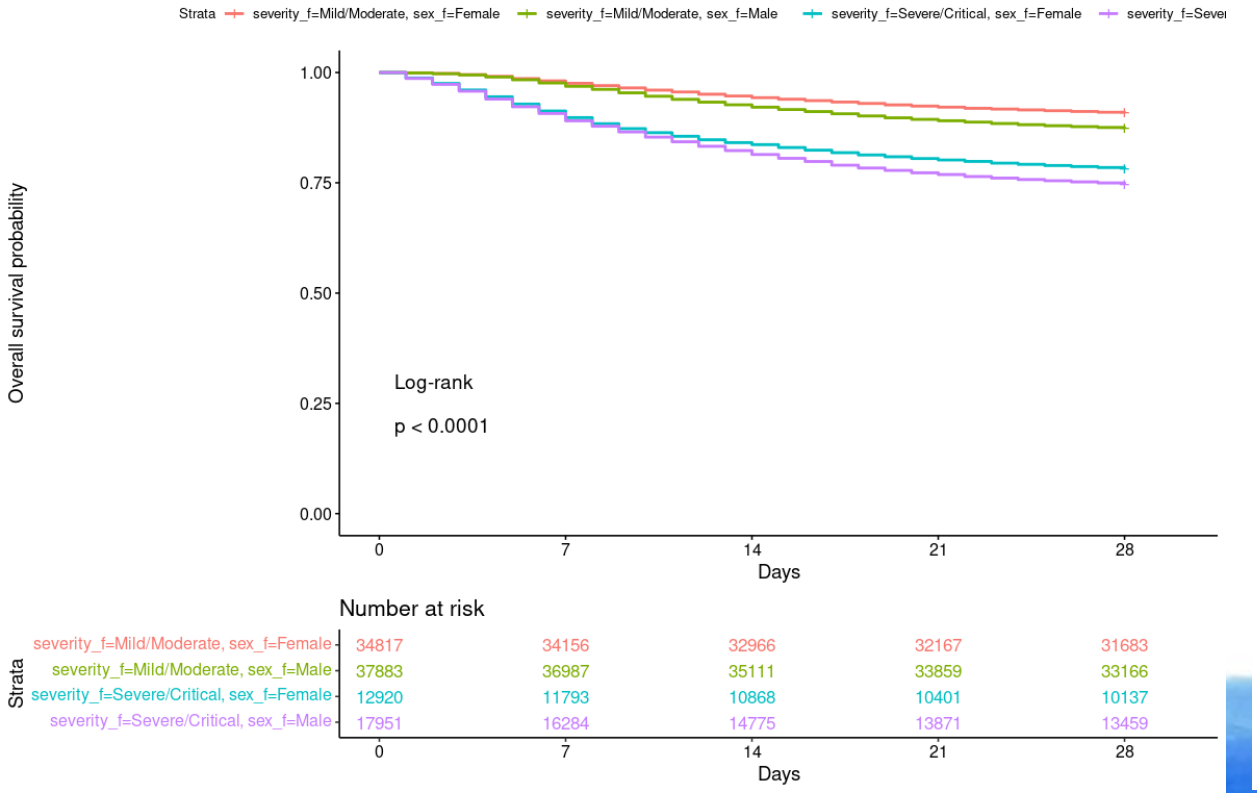
```
> adm_out %>%
dplyr::select(severity_4cat) %>%
tbl_summary(type = list(severity_4cat ~ "categorical"))
```

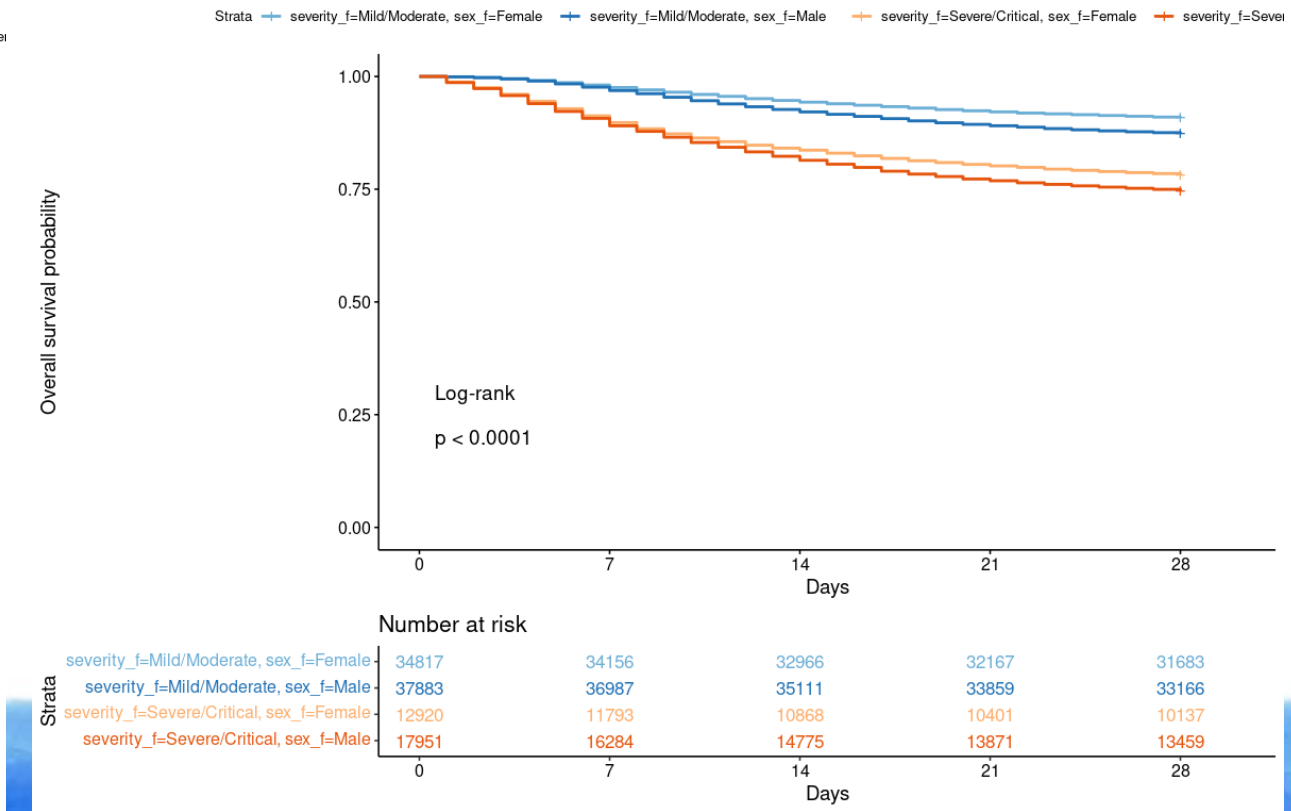
Needs to be a factor, otherwise, it is considered a continuous variable unless it has only 2 categories

Tables & Graphs

Which one?
Why?
Other
options?

label	variable	severity_f=Mild/Moderate		severity_f=Severe/Critical		Total
		sex_f=Female	sex_f=Male	sex_f=Female	sex_f=Male	
Mortality	Discharged	31052 (c% 89.19% r% 36.04%)	32279 (c% 85.21% r% 37.46%)	9864 (c% 76.35% r% 11.45%)	12974 (c% 72.27% r% 15.06%)	86169 (83.20%)
	Deceased	3765 (c% 10.81% r% 21.64%)	5604 (c% 14.79% r% 32.20%)	3056 (c% 23.65% r% 17.56%)	4977 (c% 27.73% r% 28.60%)	17402 (16.80%)
	Total	34817 (33.62%)	37883 (36.58%)	12920 (12.47%)	17951 (17.33%)	103571 (100.00%)





Demonstration

Exercise

Your team deployed you to support the response to a Cholera outbreak in Egypt, Jordan, and Tunisia. Are Lebanon and Afghanistan affected?

Cholera is an extremely virulent disease that can cause severe acute watery diarrhoea. It takes between 12 hours and 5 days for a person to show symptoms after ingesting contaminated food or water. Cholera affects both children and adults and can kill within hours if untreated.

Most people infected with *V. cholerae* do not develop any symptoms, although the bacteria are present in their faeces for 1-10 days after infection and are shed back into the environment, potentially infecting other people.

Among people who develop symptoms, the majority have mild or moderate symptoms, while a minority develop acute watery diarrhoea with severe dehydration. This can lead to death if left untreated.[source https://www.who.int/news-room/fact-sheets/detail/cholera]

The Ministry of Health and specialized agencies such as WHO are working on surveillance and response strategies. You were asked to analyze the shared data and evaluate the current epidemiological situation.

How do you analyze this data?

Exercise

1. Person

- 1.a. Demographics – provide tables and graphs for the team to evaluate the affected population and strategize the response;
  - is it needed to strengthen paediatric units, or adult population is the most affected?
  - What about elderly people?
  - Is the Female population more affected?

- 1.b. Clinical Characteristics – provide tables and graphs for the team to evaluate the symptoms, underlying conditions, and outcomes to evaluate the clinical characteristics of the Cholera disease.
  - What are the most common symptoms?
  - What are the 3 top underlying conditions?

2. Place

- 2.a. To find out what is the burden on the health system, it is necessary to map what are the affected zones and countries
  - What are the affected countries and zones?

- How would you share the output of your analysis?

i) Open the script you have been working on previous days – we need those data processing steps for this section

ii) For the tables and crosstable, you have several options:

```
> table(cholera$var1)
```

```
> crosstable(cholera, c(vari1)) %>%
flextable::as_flextable(keep_id=FALSE)
```

```
> data %>%
dplyr::select(var1) %>%
tbl_summary()
```

iii) For the graphs, according to your variable type (continuous or categorical) you may want to build on the following:

```
> ggplot(cholera, aes(x = var1, fill = var2)) +
geom_bar() +  labs(title = "Severity vs. Death", x =
"Severity", y = "Count") +  theme_minimal()
```

```
> ggplot(cholera, aes(y = var1)) + geom_boxplot() +
geom_jitter(aes(x = 0), width = 0.1, size = 1, color =
"black") + labs(title = "Boxplot of Age", y = "Age")
```

iv) Save your outputs:

```
> write.csv2(tab, "tab.csv")
```

```
> write.csv(tab, "tab_2.csv")
```

```
> ggsave("plot_age_sex.png",width = 12, height = 7)
```

HEALTH FOR ALL BY ALL
a call for solidarity and action

World Health Organization
REGIONAL OFFICE FOR THE Eastern Mediterranean