Effective Visualizations for Credible, Data-Driven Decision Making

Marc Vandemeulebroecke, Mark Baillie, Charlotta Früchtenicht and Diego Saldana
On behalf of the visR collaboration team
http://openpharma.github.com/visR

Agenda

Intro and Motivation (Marc)

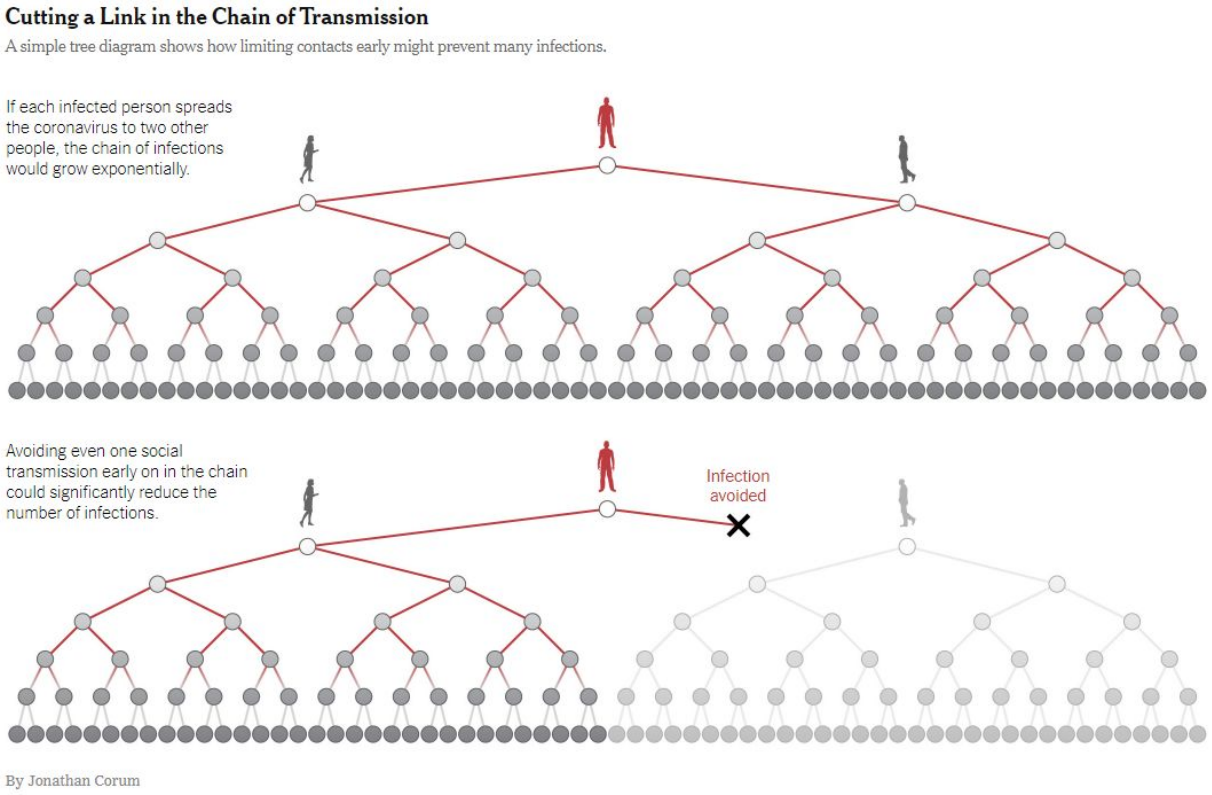
Effective Visual Communication (Mark)

visR - Motivation (Charlotta)

visR - Prototype Example and Call for Contributors (Diego)

Effective Visual Communication for Quantitative Scientists

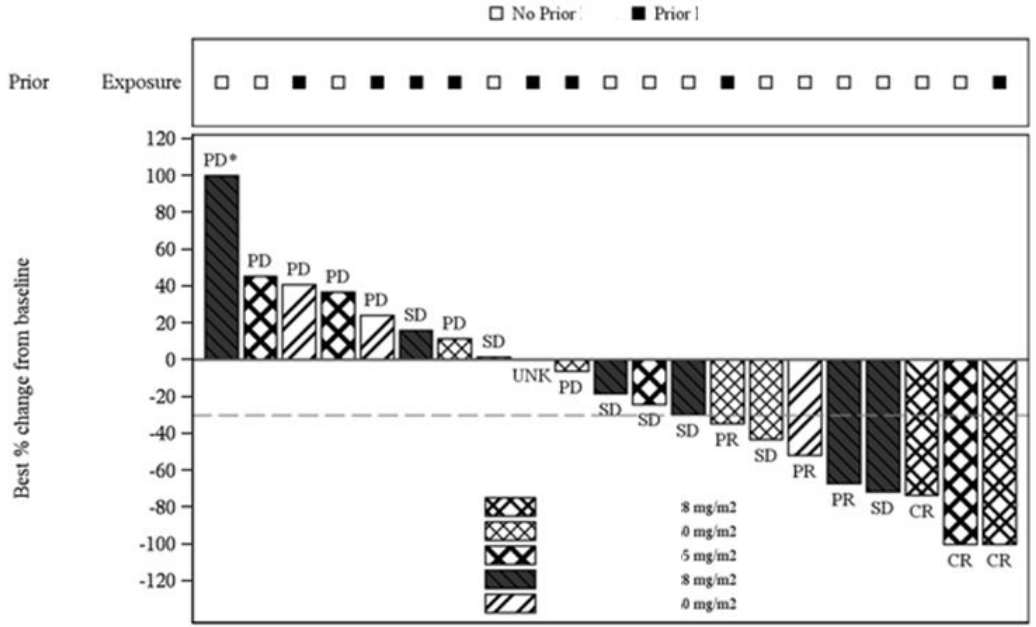
Effective visualisation is important



https://www.nytimes.com/2020/03/19/health/coronavirus-distancing-transmission.html

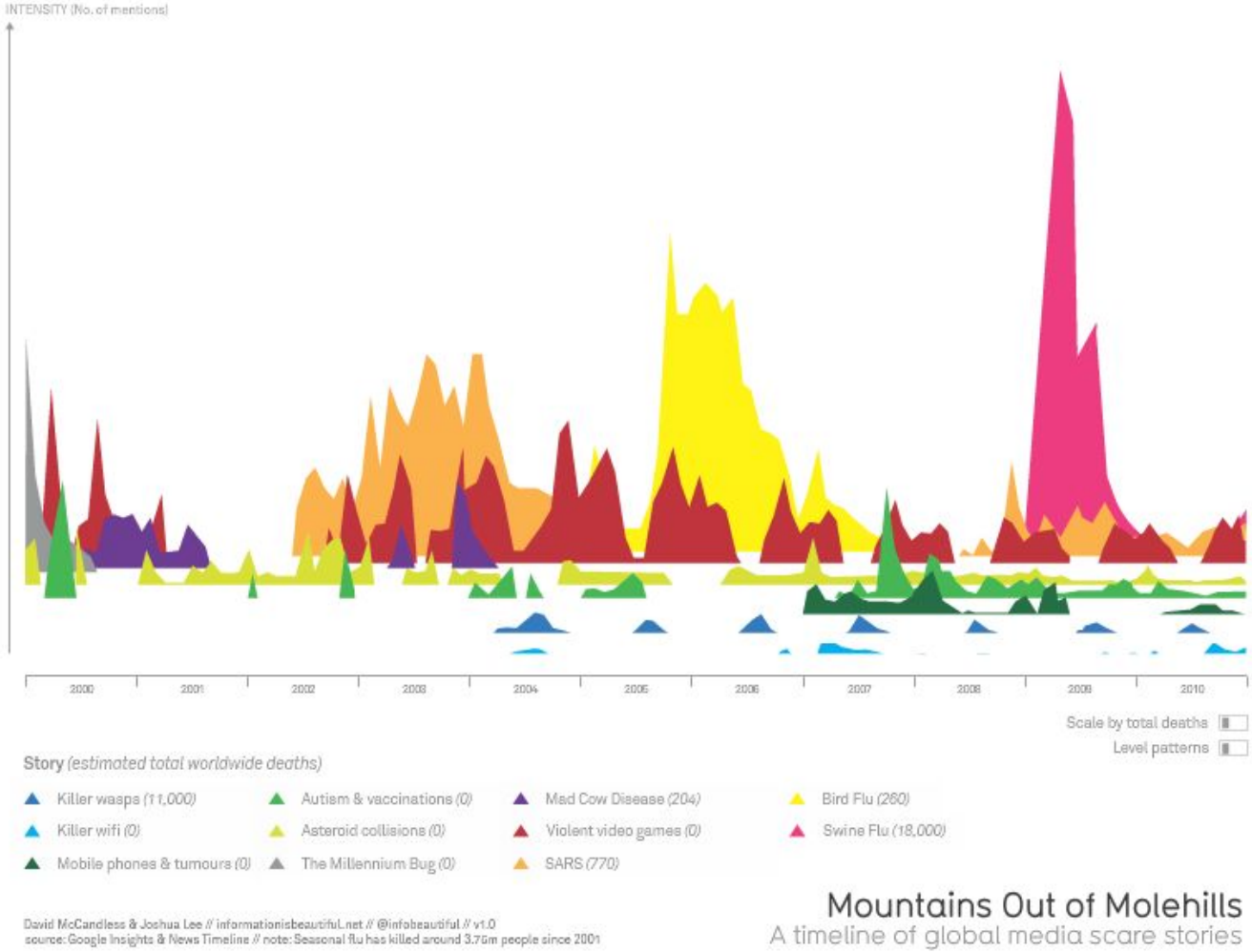
We are not always good at it

Figure 11-1 (Page 1 of 1)
Best percentage change from baseline in sum of longest diameters and best overall response as per investigator by prior treatment
(Full analysis set)



- * Denotes the percentage change from baseline greater than 100.
Source: Table 11-4, Listing 14.2-1.2 and Listing 16.2.4-1.5

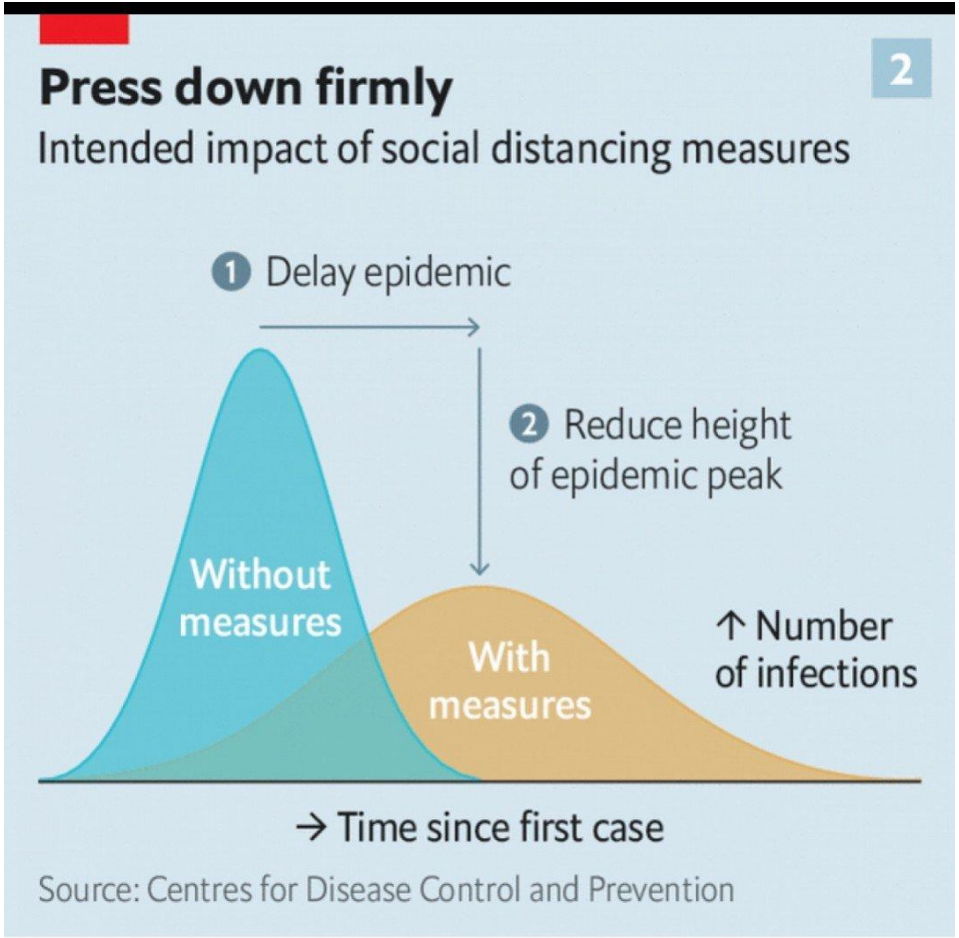
Beautiful, but effective?



https://informationisbeautiful.net/

Beautiful and effective?

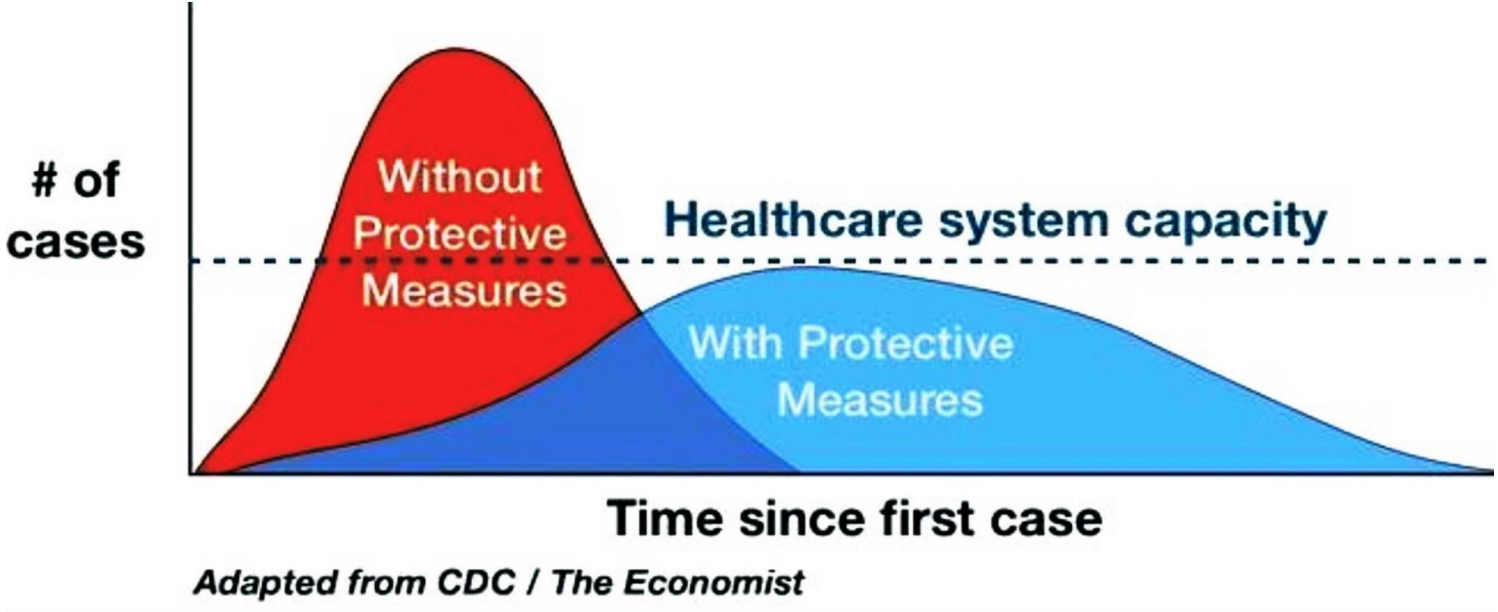
Need to #flattenthecurve



The Economist

https://www.economist.com/briefing/2020/02/29/covid-19-is-now-in-50-countries-and-things-will-get-worse

Even more effective?



https://twitter.com/CT_Bergstrom/status/1235865328074153986

Effective data visualisation is effective **visual** communication

Effective graphs...

- are visually appealing, intuitive, legible
- use the correct graph type and axis scales
- use proximity & alignment to facilitate comparison
- use labels and annotations to add clarity to the message

Most importantly, effective use of visualisations

- Enables clear and impactful communication
- Elevates our influence with our stakeholders
- Facilitates informed decision making

Three laws for improving **visual** communication

Have a clear purpose

- Know the purpose of creating the graph
- Identify the quantitative evidence to support the purpose
- Identify the audience and focus the design to support their needs

Show the data clearly

- Avoid misrepresentation (use appropriate scales)
- Choose the appropriate graph type to display your data
- Maximize data to ink ratio (reduce distraction, less is more)

Make the message obvious

- Use proximity and alignment to aid in comparisons
- Minimize mental arithmetic (e.g. plot the difference)
- Use colors and annotations to highlight important details

https://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/psp4.12455

Law 1
Have a clear purpose

Have a clear purpose

Why
Clearly identify the purpose of the graph, e.g. to deliver a message or for exploration?

What
Identify the quantitative evidence to support the purpose

Who
Identify the intended audience (specialists, non-specialists, both) and focus the design to support their needs

Where
Adapt the design to space or formatting constraints (e.g. clinical report, slide deck or publication)

https://graphicsprinciples.github.io/

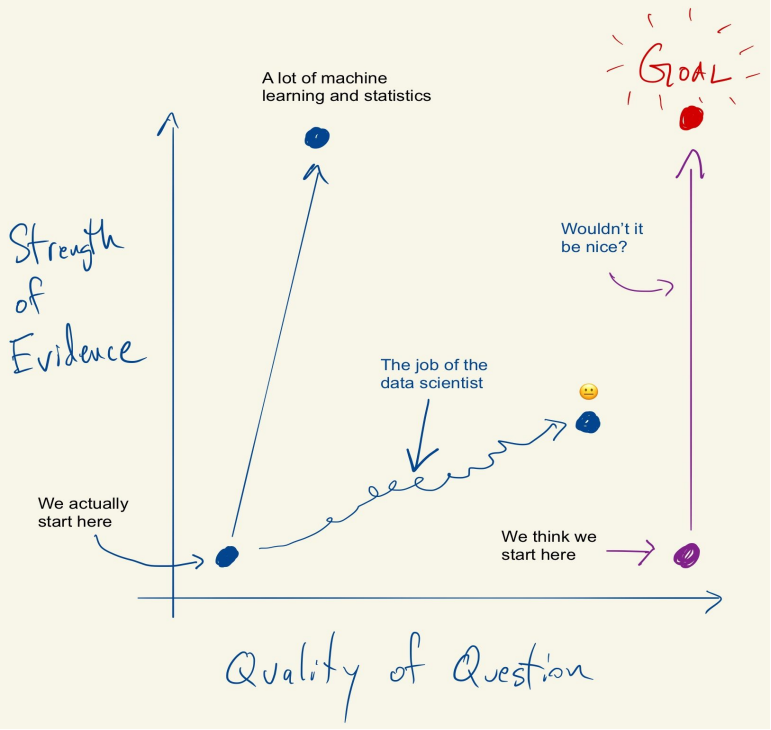
Tukey, Design Thinking, and Better Questions

Roger Peng 2019/04/17

“More Questions, Better Questions

Most of the time in data analysis, we are trying to answer a question with data. I don’t think it’s controversial to say that, but maybe that’s the wrong approach? Or maybe, that’s what we’re *not* trying to do at first. Maybe what we spend most of our time doing is figuring out a better question.”

“An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.” - John Tukey



https://simplystatistics.org/2019/04/17/tukey-design-thinking-and-better-questions/

What type of graph do I want to create?

EXPLORATORY	EXPLANATORY
"I want to dig into the data"	"I want to communicate the results"
"I want to get familiar with the data"	"I want to tell the story behind the data"
"I want to find the story in my data"	
The audience is: **YOU**	The audience is: **SOMEONE ELSE**

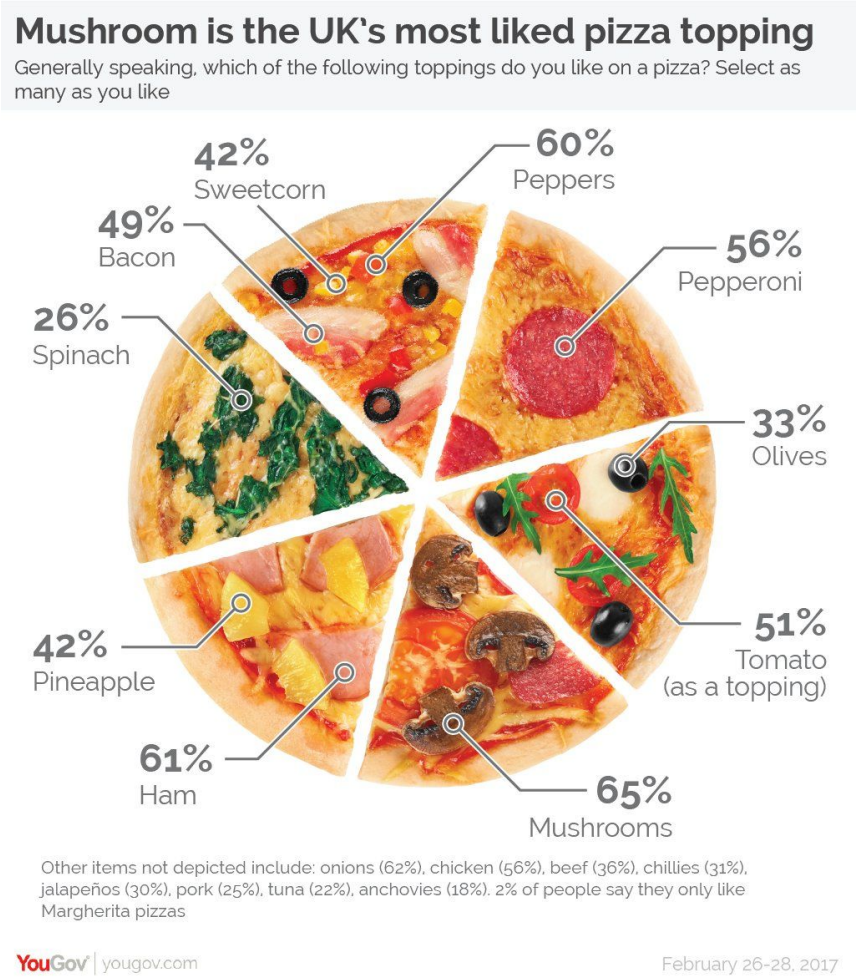
Credit Andrew Wright, Novartis

Do you want your audience to play ‘Where’s Wally?’

Credit Andrew Wright, Novartis

Law 2
Show the data clearly

Show the data clearly

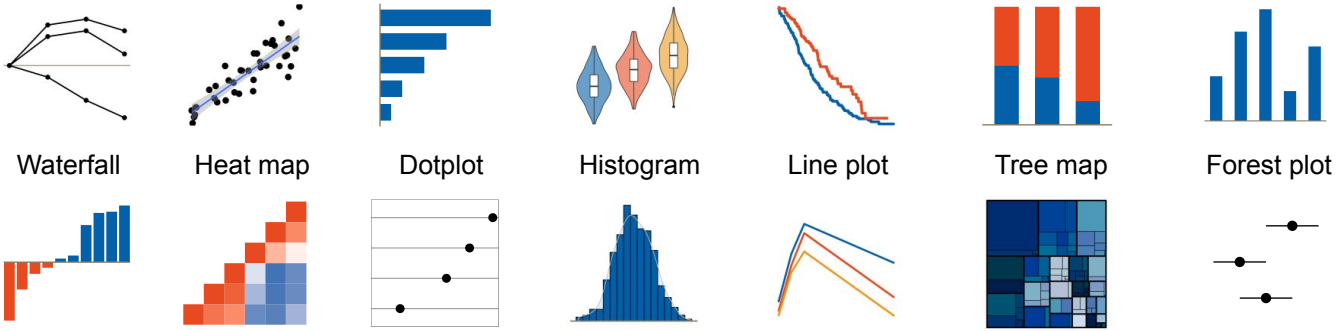


Show the data clearly

https://twitter.com/YouGov/status/838750115796041728

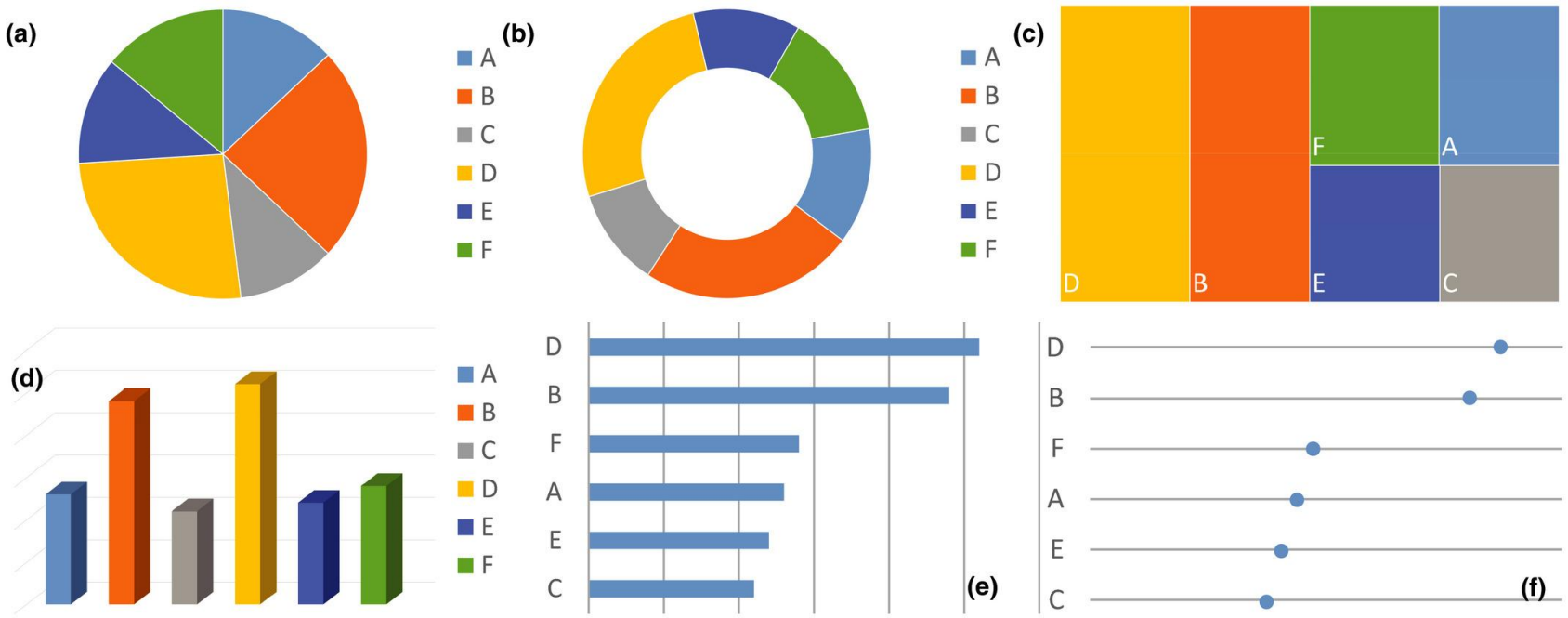
Choosing the Correct Graph Type Aids in interpretation

Deviation	Correlation	Ranking	Distribution	Evolution	Part-to-whole	Magnitude
Chg. from baseline	Scatter plot	Horizontal bar chart	Boxplot	Kaplan Meier	Stacked bar chart	Vertical bar chart
Waterfall	Heat map	Dotplot	Histogram	Line plot	Tree map	Forest plot



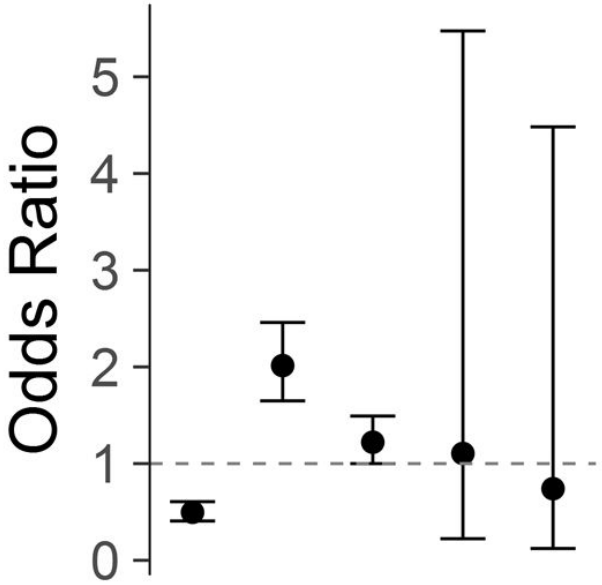
https://graphicsprinciples.github.io/

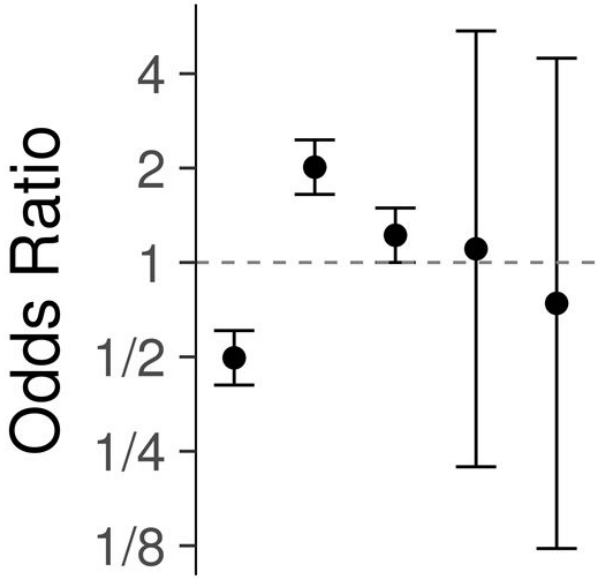
Choosing the Correct Graph Type Aids in interpretation



Choose the right scale for your data

Avoid plotting log-normally distributed variables on a linear scale (e.g. hazard ratio, AUC, CL)

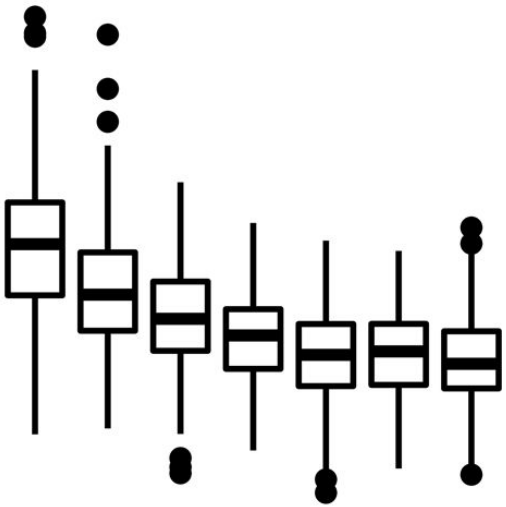


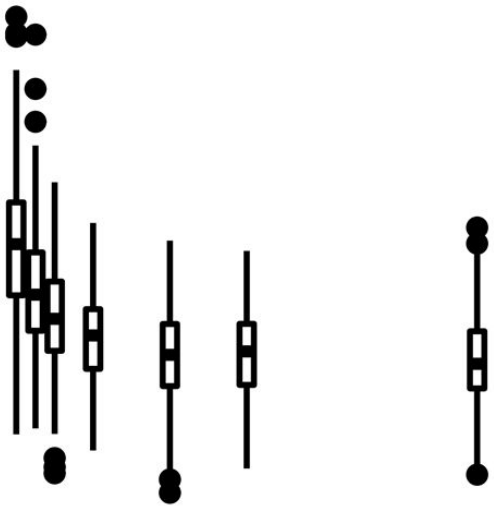


https://graphicsprinciples.github.io/

Space measurements proportional to the time between each

Measurements displayed close together are perceived to be closer in time





https://graphicsprinciples.github.io/

Law 3
Make the message obvious

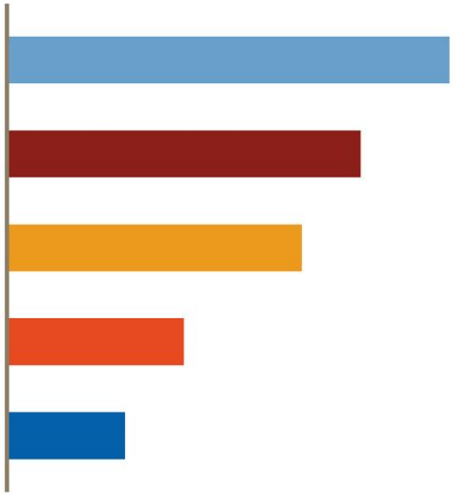
Try not to set text at an angle

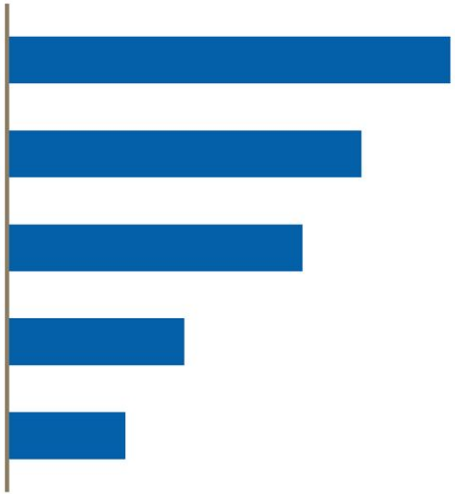
Think of alternatives such as transposing the graph

https://graphicsprinciples.github.io/

Avoid unnecessary color...

Avoid using color to differentiate between categories of the same variable

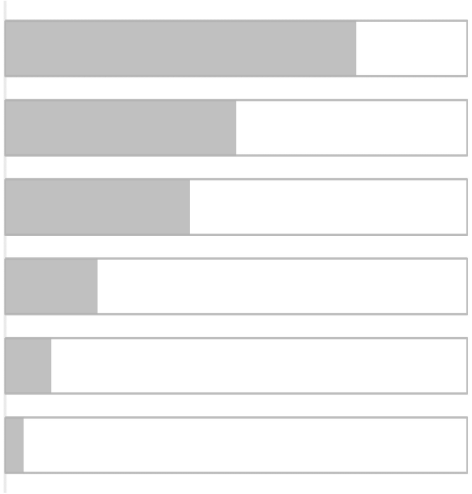


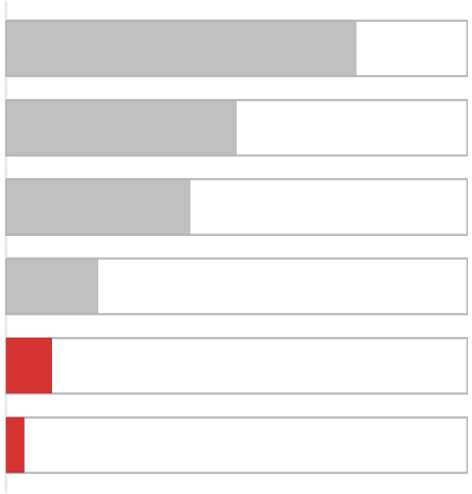


https://graphicsprinciples.github.io/

Only use color when it adds value

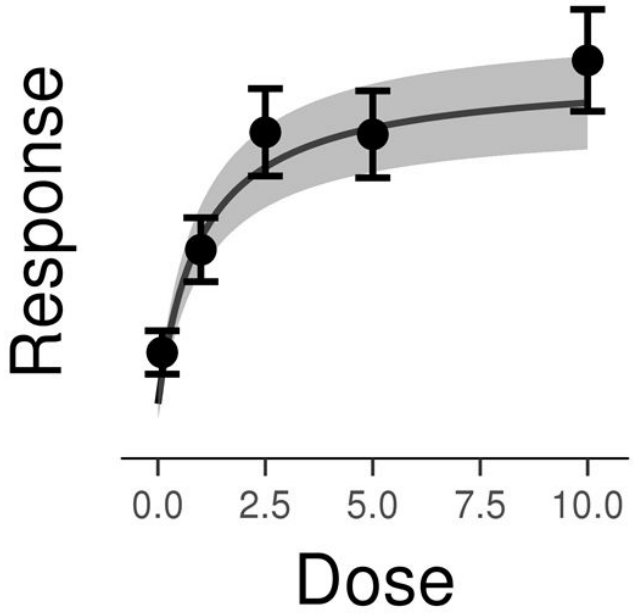
Use a bold, saturated or contrasting color to emphasize important details

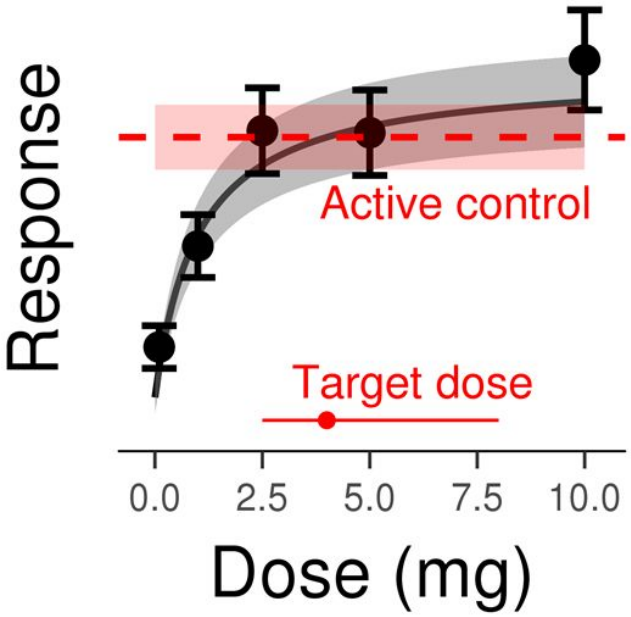




https://graphicsprinciples.github.io/

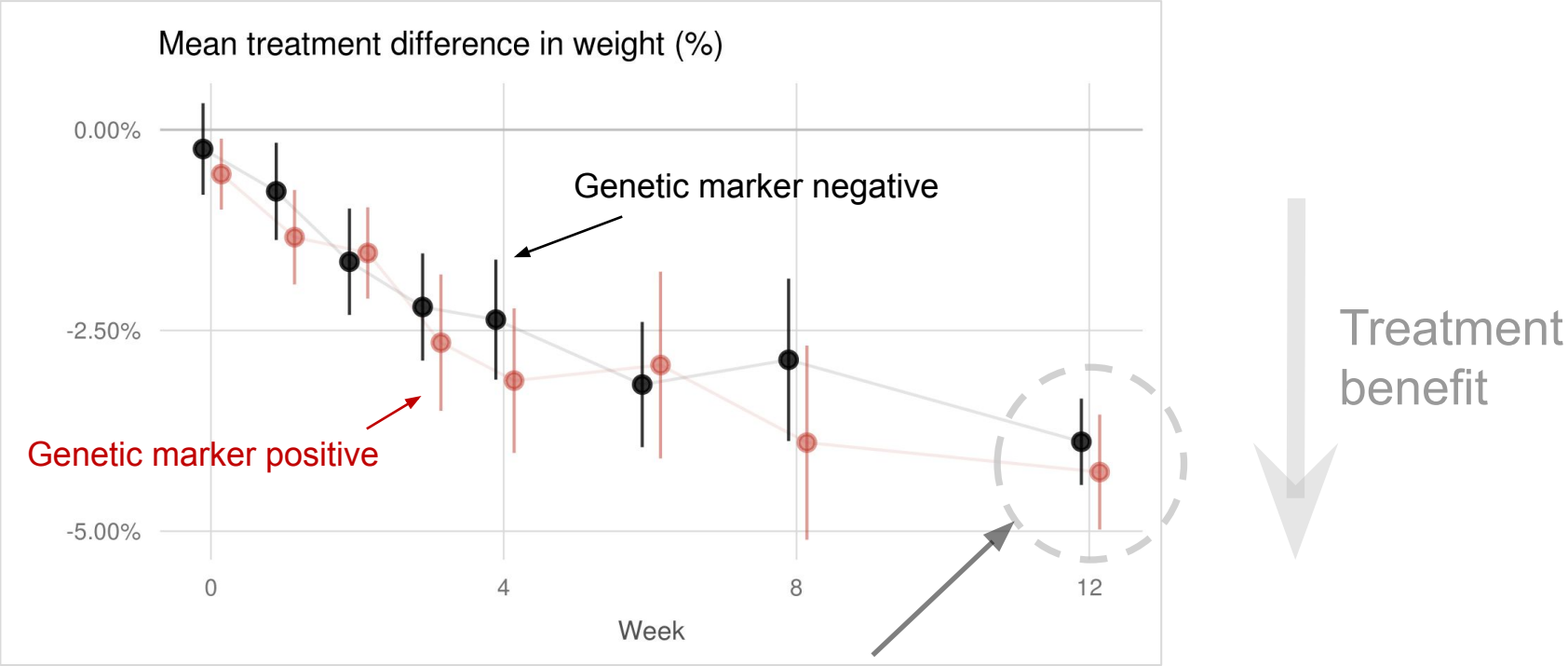
Use informative labels and annotations to support the message





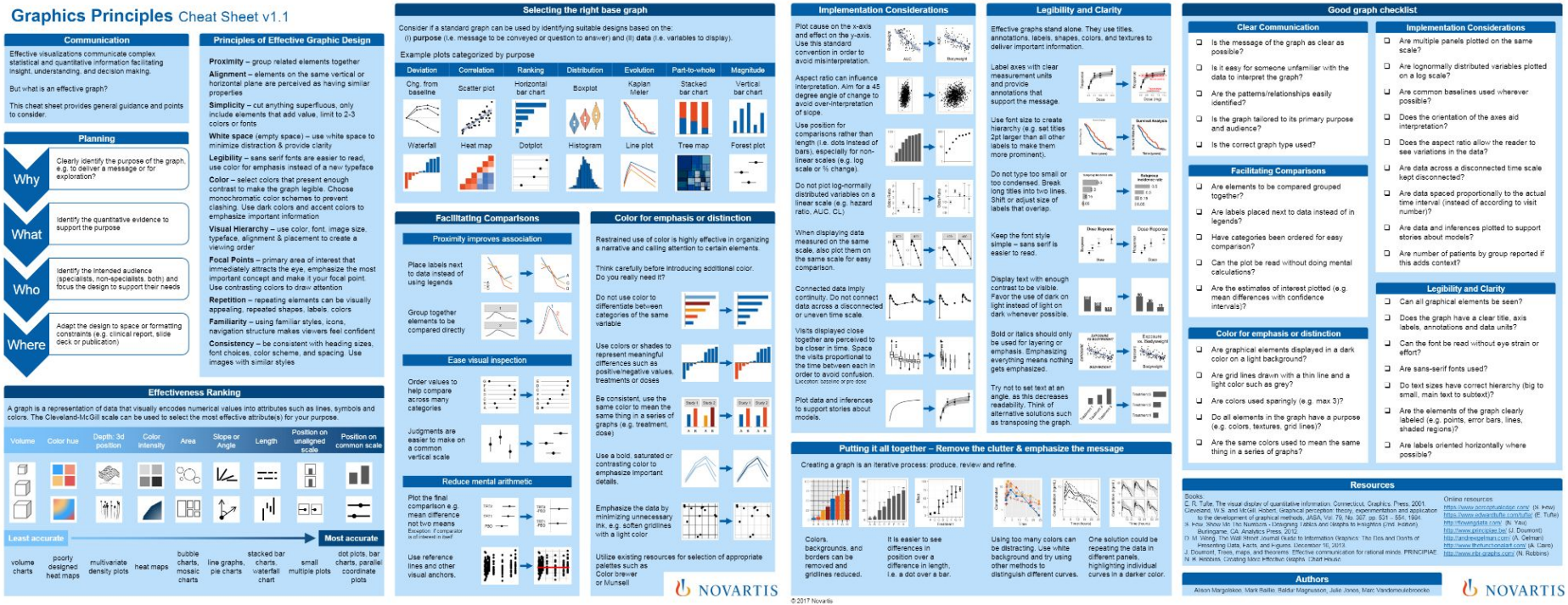
https://graphicsprinciples.github.io/

Genetic marker positive is not predictive of treatment response



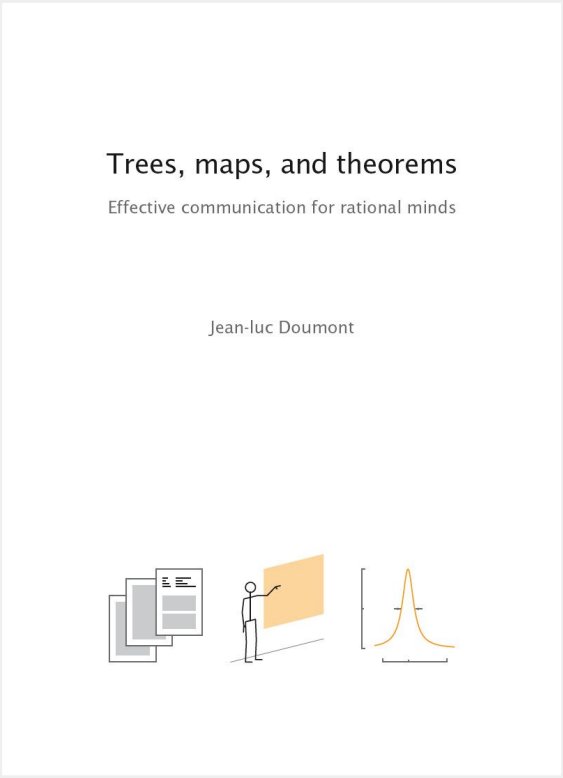
The average treatment effect is **similar** in **both** the genetic marker positive and negative subgroups and does **not** warrant further investigation

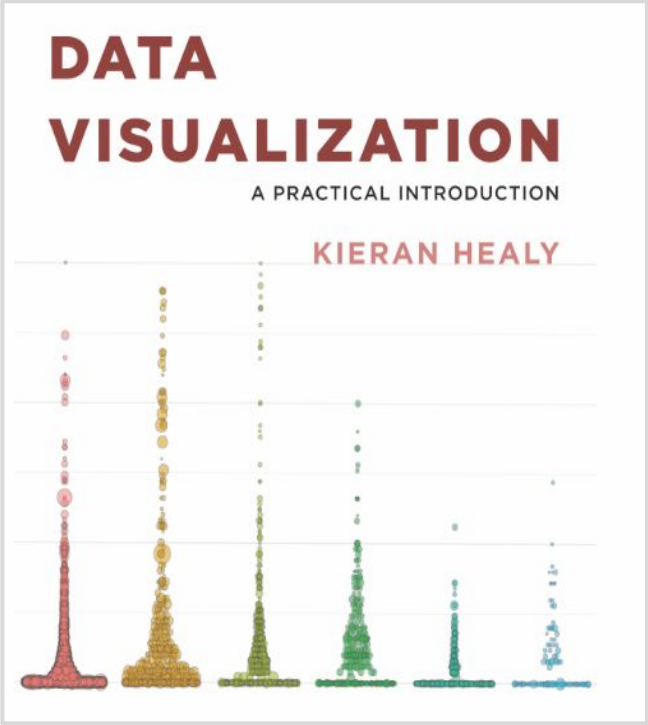
Principles for effective visual communication

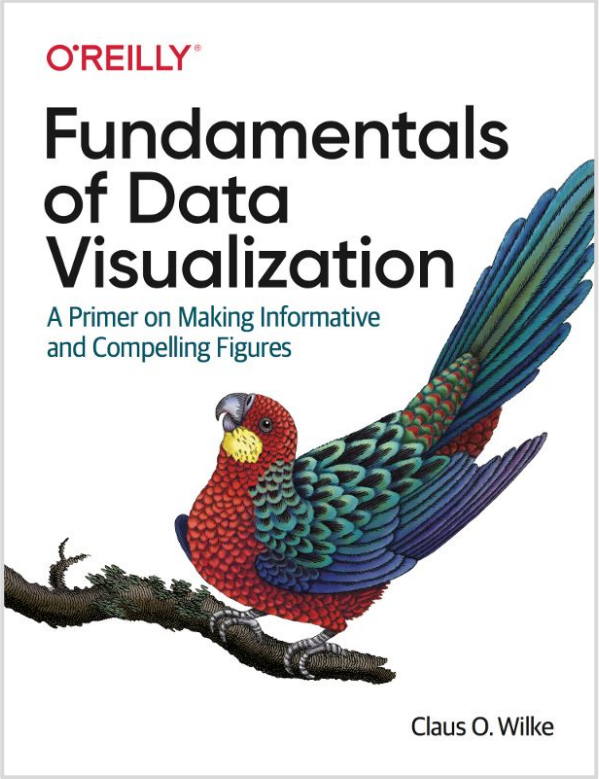


https://graphicsprinciples.github.io/

Where to find to out more?







https://www.principiae.be/book/

https://socviz.co/

https://serialmentor.com/dataviz/

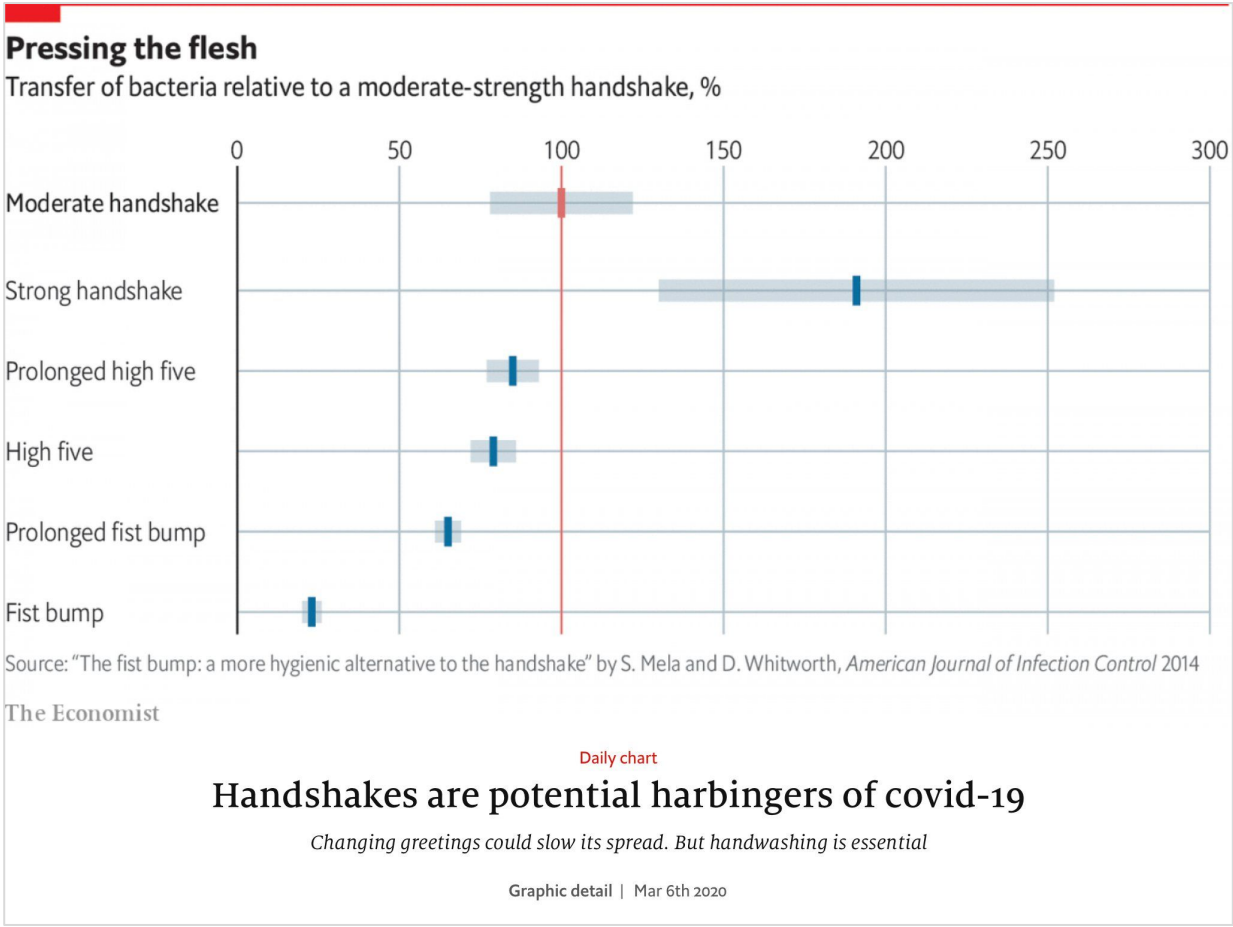
Effective data visualisation is effective **visual** communication

Effective visualisations

- enable clear and impactful communication,
- elevate our influence with our stakeholders,
- facilitate informed decision making.

To help design effective visualisations, remember the three laws: purpose, clarity and message

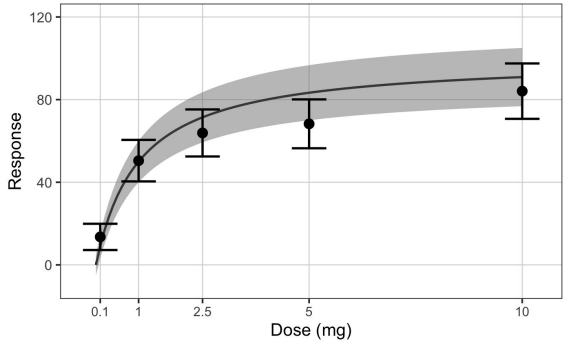
Handover



https://twitter.com/EricTopol/status/1236001710507585536

Implementing visual principles in a **reproducible way** is **tedious, but essential** at any step of a clinical development project – starting with the first exploratory analyses

```
ggplot(data = my.data, aes(x = Dose, y = Response)) +
  geom_line(size = 1) +
  geom_ribbon(aes(ymin = ymin, ymax = ymax), fill = rgb(0.5,0.5,0.5), alpha = 0.5) +
  geom_point(data = my.data[Dose %in% c(0.1,1,2.5,5,10),],
             aes(x=Dose, y=obs), size = 4) +
  geom_errorbar(data = my.data[Dose %in% c(0.1,1,2.5,5,10),],
                aes(x = Dose, ymin = obs-5-0.1*obs, ymax = obs+5+0.1*obs), size = 1) +
  scale_x_continuous(breaks = c(0.1,1,2.5,5,10), labels = c(0.1,1,2.5,5,10)) +
  xlab("Dose (mg)") +
  ylab("Response") +
  coord_cartesian(ylim = c(-10,120)) +
  theme_bw(base_size = 16) +
  theme(panel.grid.minor=element_blank(),
        panel.grid.major=element_line(color = "lightgrey", size = 0.4),
        legend.position="none",
        axis.text.x=element_text(size = 12)
  )
```



Problem

Styling and annotating plots is time consuming, so most exploratory analyses do not adhere to these principles thus creating additional work downstream

Reproducible Reporting

Requirements

Figures and tables in reports should always have:

- Title
- Dataset source & version
- Abbreviations
- Statistical tests
- Sample size
- Harmonized color theme across outputs (e.g., same color by treatment group)

Example: Table Shell for Baseline Demographics

Table <>. <Title>

		<Overall>		<exposure 1> N = XXX		<exposure 2> N = XXX			
Characteristics	**Parameters or Categories**	**Value or n**	**%**	**Value or n**	**%**	**Value or n**	**%**	**SMD**	**p-value**
<Quantitative var> (<unit>) [b]	N	XXX	.	XXX	.	XXX	.	X.XX	X.XXX[a]
	Mean	XXX	.	XXX	.	XXX	.		
	Median	XXX	.	XXX	.	XXX	.		
	SD	XXX	.	XXX	.	XXX	.		
	Range	XX-XX	.	XX-XX	.	XX-XX	.		
	IQR	XX-XX	.	XX-XX	.	XX-XX	.		
<Categorical var> [b]	<group 1>	XXX	XX.X	XX.X	XX.X	XX.X	XX.X	X.XX	X.XXX[a]
	<group 2>	XXX	XX.X	XX.X	XX.X	XX.X	XX.X		

SD = Standard Deviation, IQR = Interquartile range

[a] <Statistical test adopted> calculated on non-missing values,

[b] <Missing values> report row if there are several missing data

[d] <Variable definition if not standard> e.g. Age calculated at advanced diagnosis rather than at Index Date

Essential meta data needs to be part of the rendered object as to not get lost
Additional context can to be provided as a separate numbered caption in the report

Development Considerations

Integrate graphical principles in your analytics projects

Seamless integration into analytics & reporting workflows

Export outputs (plots & tables) to a variety of formats

Explore different visualisations of analytic data set

Adaptable to target audiences without repeating core analysis

Combination of ease of use with flexibility for complex analyses

Suitable for analytics use cases in clinical/ medical development

Why an R Package?

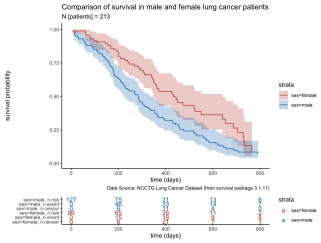
- R increasingly popular as programming language in clinical development
- Excellent existing packages solving parts of the problem that we can build upon
- Flexible towards multiple analyses questions and stages in the workflow
- Allows full documentation and examples
- Functions can be tested and versioned
- Open source so everyone can use the package and contribute to future development

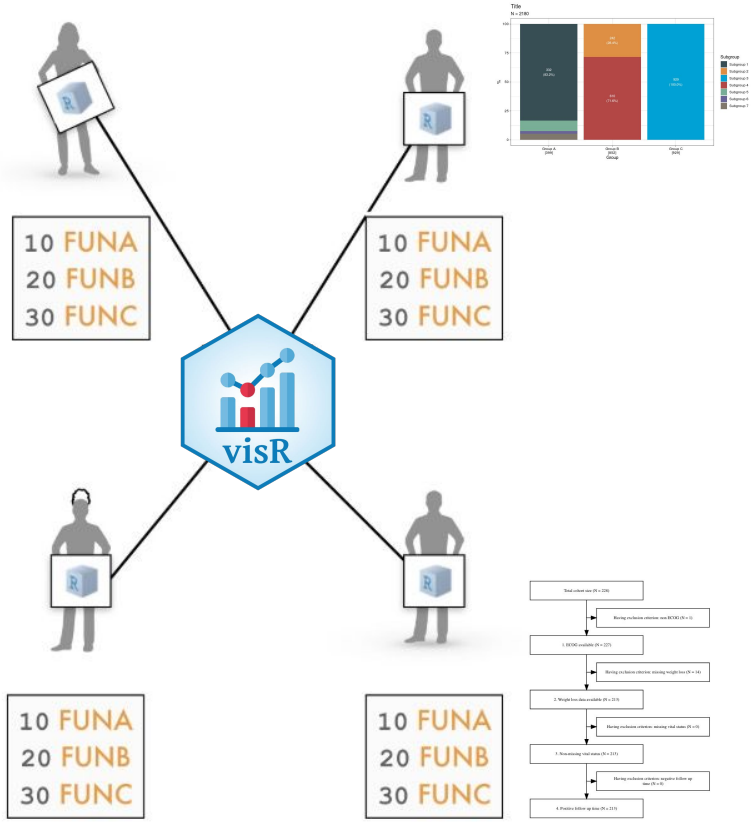
Attrition Table

Summary of samples fulfilling inclusion/exclusion criteria

Criteria	Condition	Remaining N	Remaining %	Excluded N
Total cohort size	none	228.00	100.00	0.00
1. ECOG available	!is.na(ECOG)	227.00	99.56	1.00
2. Weight loss data available	!is.na(wt.loss)	213.00	93.42	14.00
3. Non-missing vital status	!is.na(status)	213.00	93.42	0.00
4. Positive follow up time	time >= 0	213.00	93.42	0.00

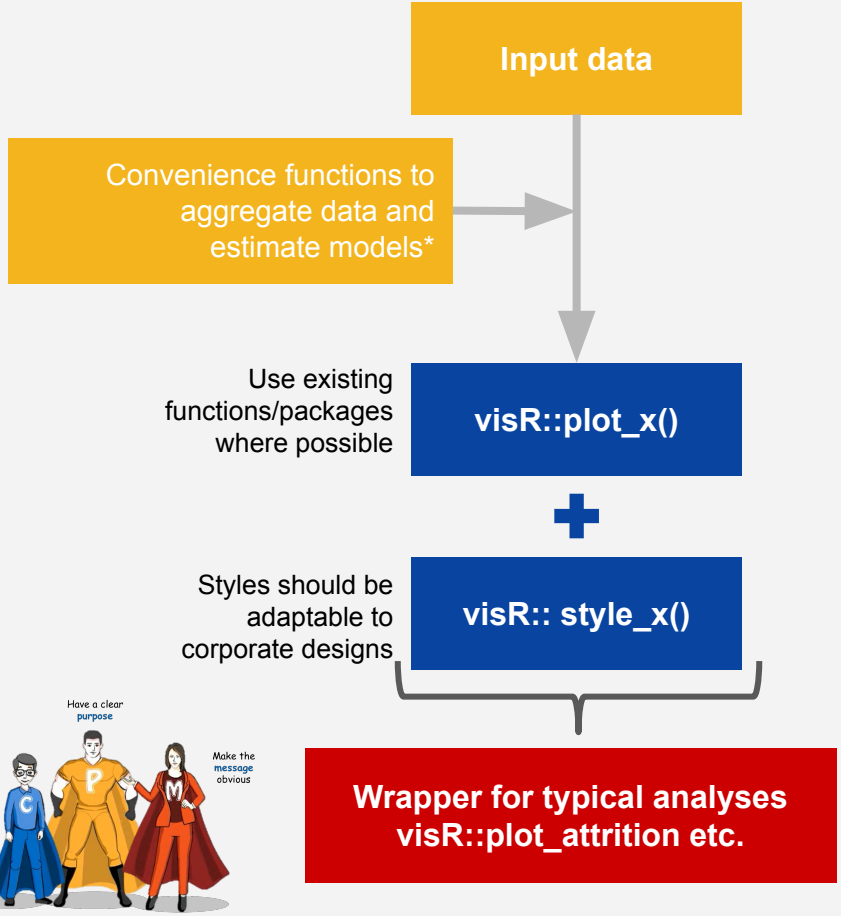
Data Source: NCCTG Lung Cancer Dataset (from survival package 3.1.11)





Package Architecture

- Should integrate seamlessly into tidyverse
  - Re-use established tools where possible
  - Interact with dplyr and modeling packages
  - Plotting should build on ggplot2
- Full transparency on data modification
- Multiple rendering and styling options to allow for various output formats (html, pdf, word, …)



**Functions for survival models, p-values, confidence intervals,...*
… make available separately & allow to call on patient-level data

Basic Architecture

- Fixed input data models for:
  - **Estimator function:** computes estimates, as well as upper and lower bounds, p-values, etc.
  - **Visualization function:** visualizes data as a plot or a table (or something else).
  - **Style function**: applies common theme and color palettes to all outputs
- Broom can handle different variations of survival plot (KM, cumulative incidence, etc).
- We could define custom time windows (e.g. three years).
- We could also add the p-value (with custom hypothesis tests, mentioned in the footnote).

Input data model (a data.frame / tibble)

trt	time	status

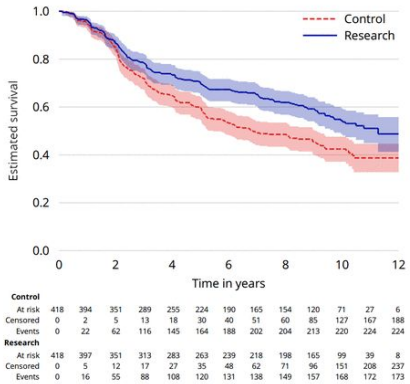
Estimator Function

Interim (estimate) data model (e.g. broom)

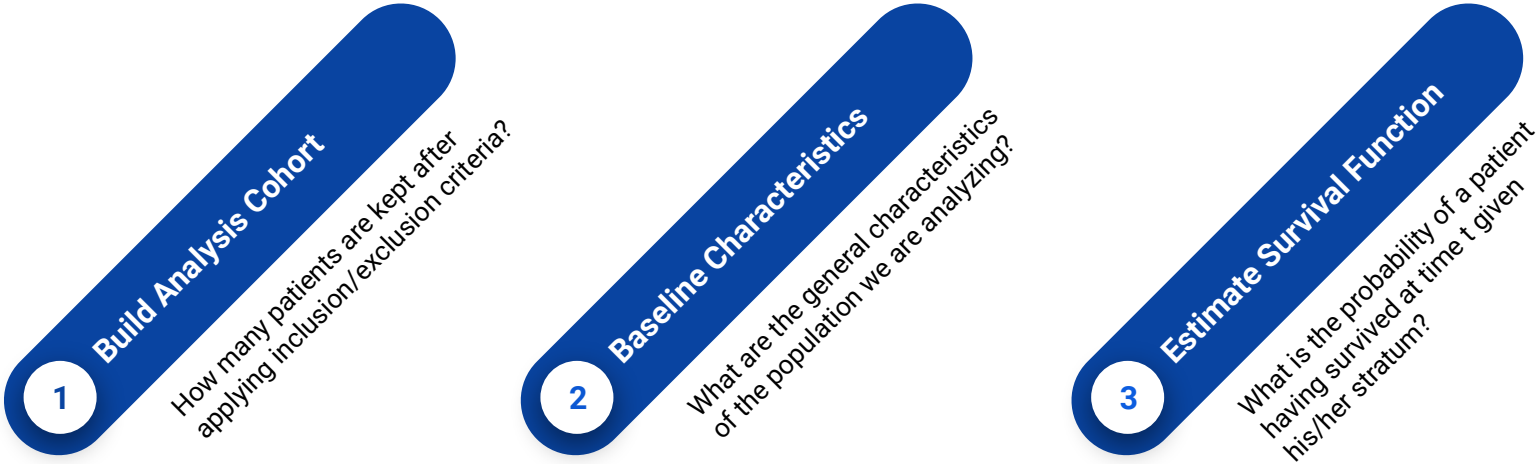
time	trt	estimate	lower	upper

Visualization Function + Style Function

Visualization



Typical Time To Event Analysis Workflow



Typical Output

Attrition

- Table
- Flow chart

Summary

- “Table One”

Kaplan-Meier

- Kaplan-Meier Plot
- Median Survival

1 Build Analysis Cohort

Create list of filters and description to easily evaluate step-wise attrition in the original cohort.

Table is annotated with critical metadata including title and data source

```r
cohort_attrition <- vr_attrition_table(
  data = lung_cohort,
  criteria_descriptions = c("1. ECOG available",
                            "2. Weight loss data available",
                            "3. Non-missing vital status",
                            "4. Positive follow up time"),
  criteria_conditions   = c("!is.na(ECOG)",
                            "!is.na(wt.loss)",
                            "!is.na(status)",
                            "time >= 0"),
  subject_column_name   = 'patid')


vr_render_table(data = cohort_attrition,
                title = "Attrition Table",
                caption = "Summary of samples fulfilling inclusion/exclusion criteria",
                datasource = DATASET,
                engine = "gt")
```

Attrition Table

Summary of samples fulfilling inclusion/exclusion criteria

Criteria	Condition	Remaining N	Remaining %	Excluded N	Excluded %
Total cohort size	none	228.00	100.00	0.00	0.00
1. ECOG available	!is.na(ECOG)	227.00	99.56	1.00	0.44
2. Weight loss data available	!is.na(wt.loss)	213.00	93.42	14.00	6.14
3. Non-missing vital status	!is.na(status)	213.00	93.42	0.00	0.00
4. Positive follow up time	time >= 0	213.00	93.42	0.00	0.00

Data Source: NCCTG Lung Cancer Dataset (from survival package 3.1.11)

1 Build Analysis Cohort (II)

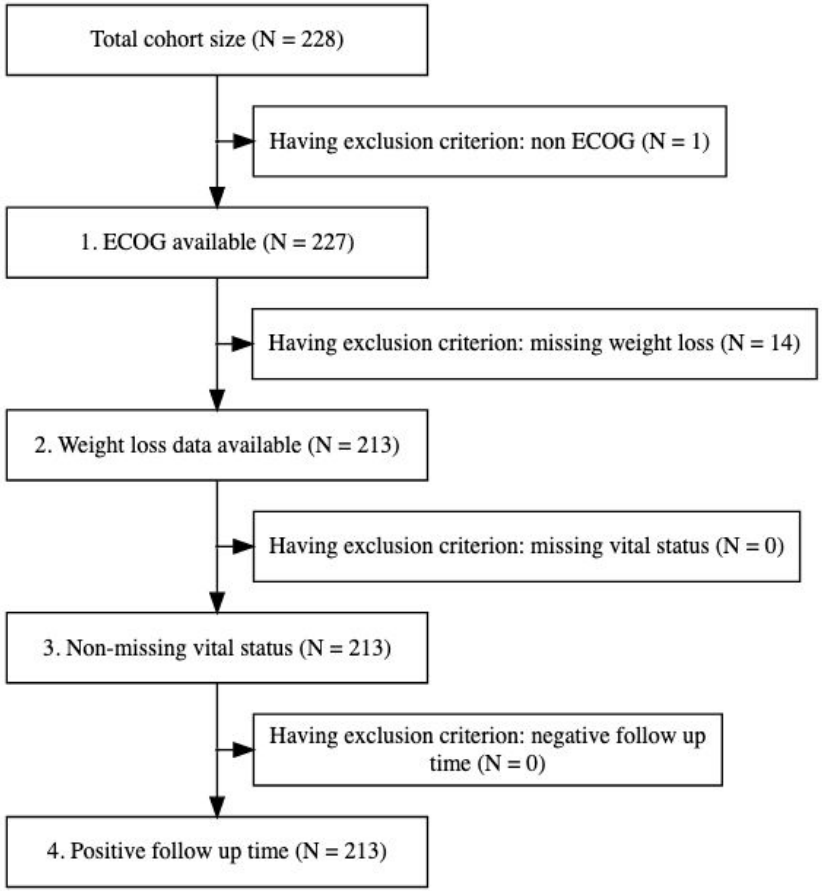
Quickly convert attrition table into flow diagram

```r
complement_descriptions <- c(
    "Having exclusion criterion: non ECOG",
    "Having exclusion criterion: missing weight loss ",
    "Having exclusion criterion: missing vital status",
    "Having exclusion criterion: negative follow up time"
)

# Create attrition flowchart
attrition_flow <- vr_attrition(
  N_array = cohort_attrition$`Remaining N`,
  descriptions = cohort_attrition$Criteria,
  complement_descriptions = complement_descriptions,
  output_path = "./attrition_diagram.svg")
```

created by vr_attrition_table.R

Can help to comply with reporting guidelines like **CONSORT** and **STROBE**.



2 Baseline Characteristics

- Summary statistics can be calculated and displayed in a table
- Level of detail on summary stats can be easily adapted using in-built or custom summary functions
- Output available as kable, Rstudio gt or DT html tables with and without download feature

```
vr_table_one(lung_cohort_tab1,
             title = "Overview over Lung Cancer patients",
             caption = "Baseline characteristics of study cohort",
             datasource = DATASET, groupCols = c("sex"), engine="dt")
```

Overview over Lung Cancer patients

Baseline characteristics of study cohort stratified by treatment type

	female	male	Overall
Sample			
N	86	127	213
time			
Mean (SD)	345 (205)	295 (216)	315 (213)
Median (IQR)	294 (201-467)	229 (163-392)	269 (176-428)
Min-max	5-965	11-1022	5-1022
Missing	0 (0%)	0 (0%)	0 (0%)
status			
Alive/Censored	37 (43%)	25 (19.7%)	62 (29.1%)
Dead	49 (57%)	102 (80.3%)	151 (70.9%)
age			
Mean (SD)	61.1 (9.01)	63.5 (9.25)	6
Median (IQR)	61 (55-68.8)	64 (57-70)	6
Min-max	41-77	39-82	3
Missing	0 (0%)	0 (0%)	0
ECOG			
0 asymptomatic	26 (30.2%)	35 (27.559%)	6
1 ambulatory	41 (47.7%)	65 (51.181%)	1
2 in bed less than 50% of day	19 (22.1%)	26 (20.472%)	4
3 in bed more than 50% of day	NA	1 (0.787%)	1
Karnofsky			
Mean (SD)	82.6 (12)	81.9 (12.5)	8
Median (IQR)	80 (80-90)	80 (70-90)	8
Min-max	50-100	50-100	5
Missing	0 (0%)	0 (0%)	0
Age Group			
30-50y	12 (14.0%)	13 (10.2%)	2
51-70y	59 (68.6%)	85 (66.9%)	1
> 70y	15 (17.4%)	29 (22.8%)	4

Data Source: NCCTG Lung Cancer Dataset (from survival package 3.1.11

Rendering as gt table

Copy CSV Excel Filter:

Overview over Lung Cancer patients - Baseline characteristics of study cohort stratified by treatment type

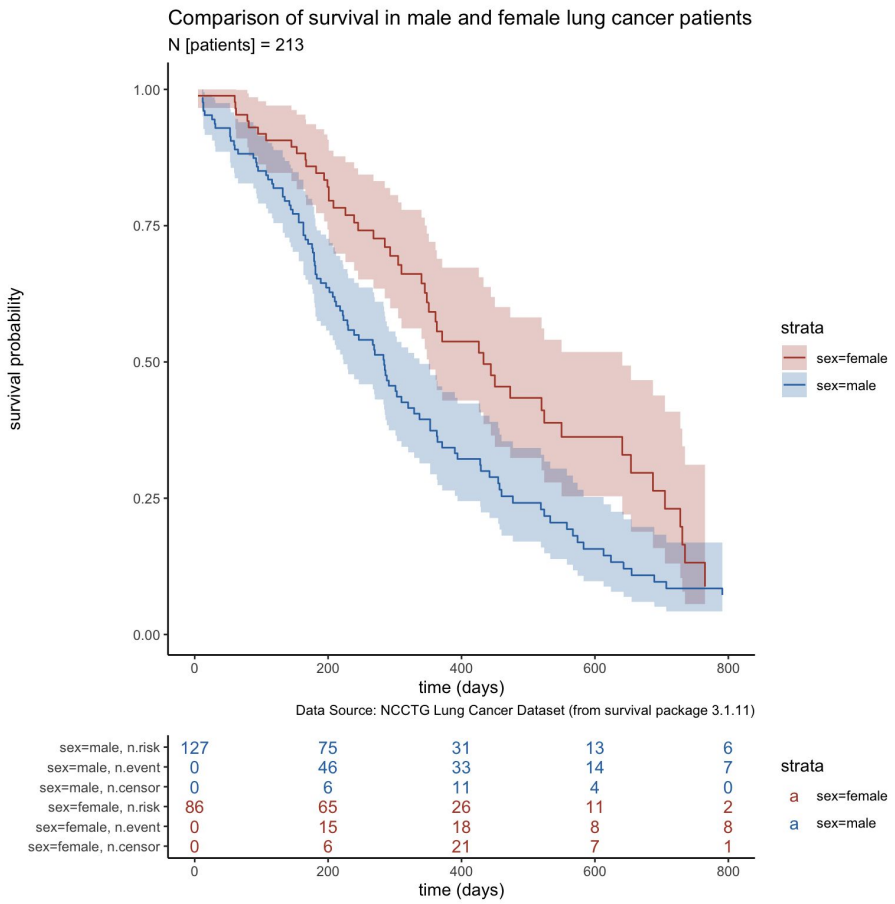
	variable	statistic	female	male	Overall
1	Sample	N	86	127	213
2	time	Mean (SD)	345 (205)	295 (216)	315 (213)
3	time	Median (IQR)	294 (201-467)	229 (163-392)	269 (176-428)
4	time	Min-max	5-965	11-1022	5-1022
5	time	Missing	0 (0%)	0 (0%)	0 (0%)
6	status	Alive/Censored	37 (43%)	25 (19.7%)	62 (29.1%)
7	status	Dead	49 (57%)	102 (80.3%)	151 (70.9%)
8	age	Mean (SD)	61.1 (9.01)	63.5 (9.25)	62.5 (9.21)
9	age	Median (IQR)	61 (55-68.8)	64 (57-70)	63 (56-70)
10	age	Min-max	41-77	39-82	39-82
11	age	Missing	0 (0%)	0 (0%)	0 (0%)
12	ECOG	0 asymptomatic	26 (30.2%)	35 (27.559%)	61 (28.638%)
13	ECOG	1 ambulatory	41 (47.7%)	65 (51.181%)	106 (49.765%)
14	ECOG	2 in bed less than 50% of day	19 (22.1%)	26 (20.472%)	45 (21.127%)
15	Karnofsky	Mean (SD)	82.6 (12)	81.9 (12.5)	82.2 (12.2)
16	Karnofsky	Median (IQR)	80 (80-90)	80 (70-90)	80 (80-90)
17	Karnofsky	Min-max	50-100	50-100	50-100
18	Karnofsky	Missing	0 (0%)	0 (0%)	0 (0%)
19	Age Group	30-50y	12 (14.0%)	13 (10.2%)	25 (11.7%)
20	Age Group	51-70y	59 (68.6%)	85 (66.9%)	144 (67.6%)
21	Age Group	> 70y	15 (17.4%)	29 (22.8%)	44 (20.7%)
22	ECOG	3 in bed more than 50% of day		1 (0.787%)	1 (0.469%)

Data Source: NCCTG Lung Cancer Dataset (from survival package 3.1.11)

Rendering as dt table

3 Survival Analysis

- Based graphical principles on findings from a survey and paper by Morris *et al.* conducted among 1176 researchers about what the perfect Kaplan-Meier plot should look like
- KM plot shows relevant information such as number of patients, axis labels (with units where needed), data source.
- Risk table shows num. at risk, events, censored, by stratum at regular timepoints.



3 Survival Analysis (II)

Kaplan-Meier Curve

```r
os_obj <- vr_est_kaplan_meier(lung_cohort,
    equation = "survival::Surv(time, status) ~ sex")
vr_plt_kaplan_meier(os_obj) +
    theme_minimal()
```

estimation

plotting

Risk Table

```r
risk_obs <- vr_est_km_risk_table(lung_cohort,
    equation = "survival::Surv(time, status) ~ sex")
vr_plt_km_risk_table(risk_obs, time_unit = "days")
```

estimation

plotting

or

Kaplan-Meier Curve + Risk Table

```r
vr_kaplan_meier(
    lung_cohort,
    equation = "survival::Surv(time, status) ~ sex",
    data_source = DATASET,
    time_unit = "days",
    title = "Comparison of survival in male and female lung cancer patients"
)
```

wrapper

3 Survival Analysis (III)

- We have also implemented other convenient estimation functions.
- Median survival times by stratum.
- Multiple methods for testing equality between strata.

```
km_summary <- vr_kaplan_meier_summary(
    data = lung_cohort,
    equation = "survival::Surv(time, status) ~ sex")

km_summary[[1]] %>%
  vr_render_table(title = "Overall Survival",
                  caption = "Median survival times in days for each strata",
                  datasource = DATASET)

km_summary[[2]] %>%
  vr_render_table(title = "Equality between Strata",
                  caption = "Summary table with test of equality over strata",
                  datasource = DATASET)
```

Overall Survival

Median survival times in days for each strata

strata	# persons	# events	median survival time	0.95LCL	0.95UCL
female	86.00	49.00	433.00	351.00	641.00
male	127.00	102.00	284.00	223.00	337.00

Data Source: NCCTG Lung Cancer Dataset (from survival package 3.1.11)

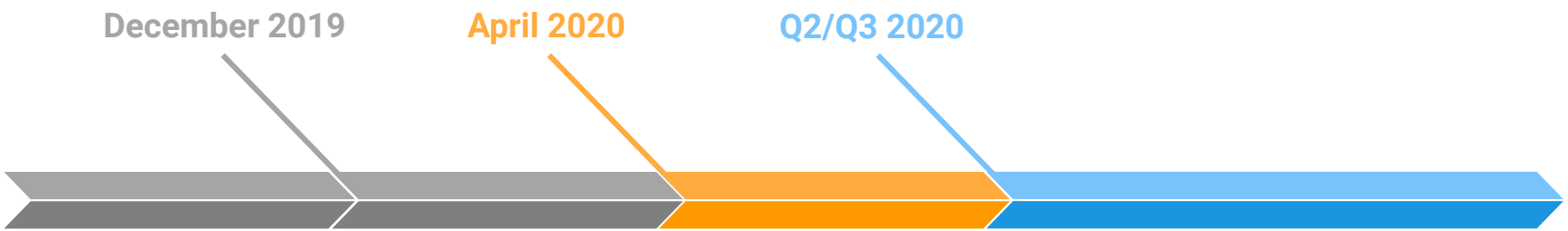
Equality between Strata

Summary table with test of equality over strata

Test	statistic	df	p.value	Description
Log-rank	9.60	1.00	0.00	Log-rank gives more weight on higher values of time
Wilcoxon	11.84	1.00	0.00	Wilcoxon gives more weight on lower values of time
Tarone-Ware	11.63	1.00	0.00	Tarone-Ware is in between

Data Source: NCCTG Lung Cancer Dataset (from survival package 3.1.11)

Package Roadmap



Project Kick Off

- Definition of needs and requirements
- Agreement on basic architecture

Develop Prototype

- Implementation of outputs for a typical Time to Event analysis

Launch & Find Collaborators

- Present prototype to stakeholders
- Recruit additional collaborators to drive development

Scale Development

- Refine architecture and design choices
- Continue implementation of plot and table types following graphical principles

Looking for Contributors: Join the visR Team

visR is still in its experimental phase and we are looking for partners to further develop the package!

- Add feedback/ideas for features using github issues
- Contribute code the open source-way: pick an issue & work on it
- Reach out to us to join core team

What contributions are we looking for?

- Design choices
- Project governance
- Hands on engineering
- Help maintain an actively used package

How to reach out?

Email: mark.baillie@novartis.com & james.black.jb2@roche.com

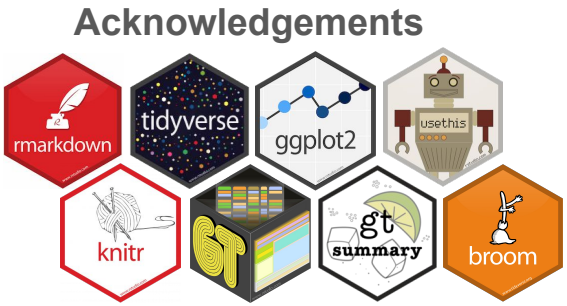
Acknowledgements

visR
http://openpharma.github.com/visR

Graphics Principles
https://graphicsprinciples.github.io/

Acknowledgments
Thanos Siadimas, Pawel Kawski, James Black,
Janine Hoffart, Baldur Magnusson, Alison Margolskee

How to reach out?
Email: mark.baillie@novartis.com &
james.black.jb2@roche.com



Survival, DT Datatables,
kable & kableExtra, ggpubr,
and many more