



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)



Have a clear
purpose

Show the
data
clearly



Make the
message
obvious

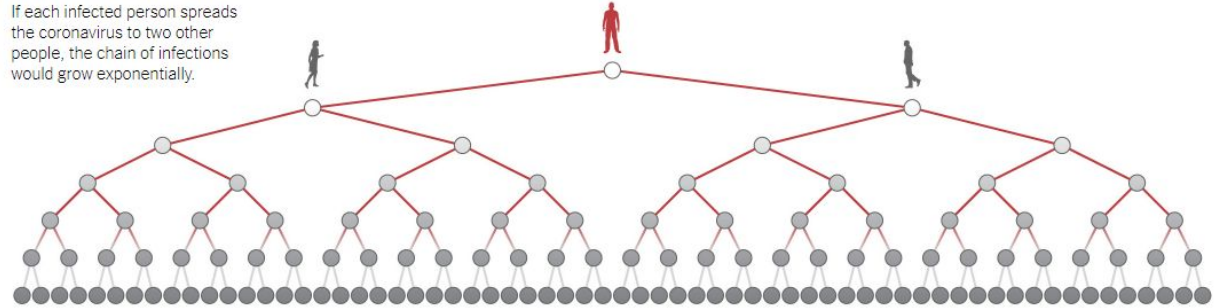
Effective Visual Communication for Quantitative Scientists

Effective visualisation is important

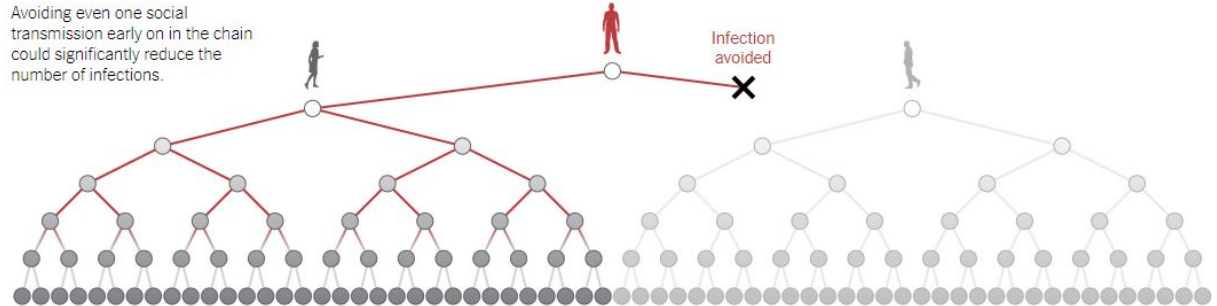
Cutting a Link in the Chain of Transmission

A simple tree diagram shows how limiting contacts early might prevent many infections.

If each infected person spreads the coronavirus to two other people, the chain of infections would grow exponentially.



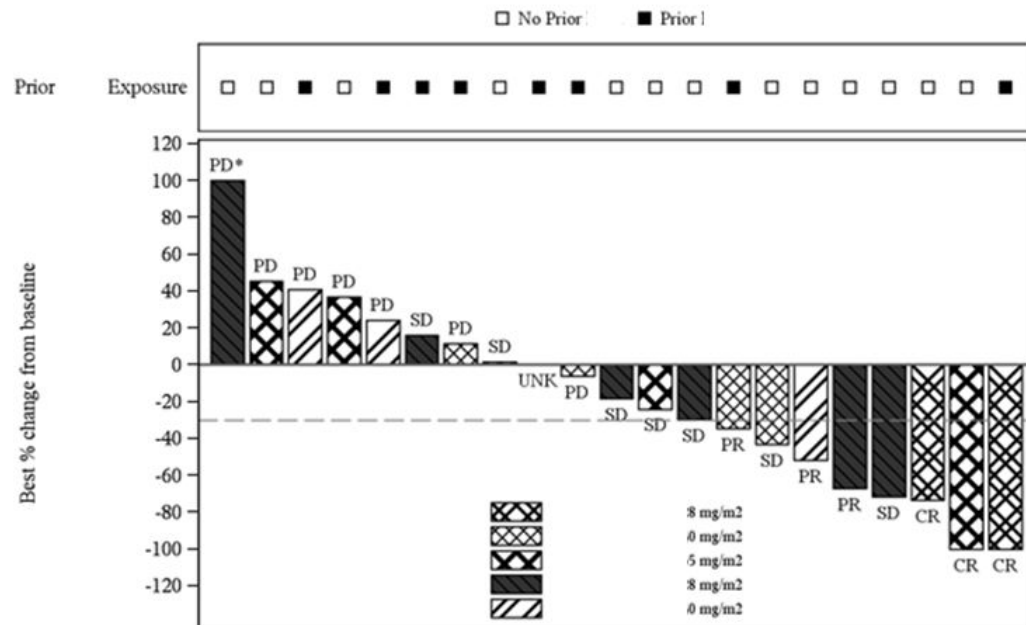
Avoiding even one social transmission early on in the chain could significantly reduce the number of infections.



By Jonathan Corum

We are not always good at it

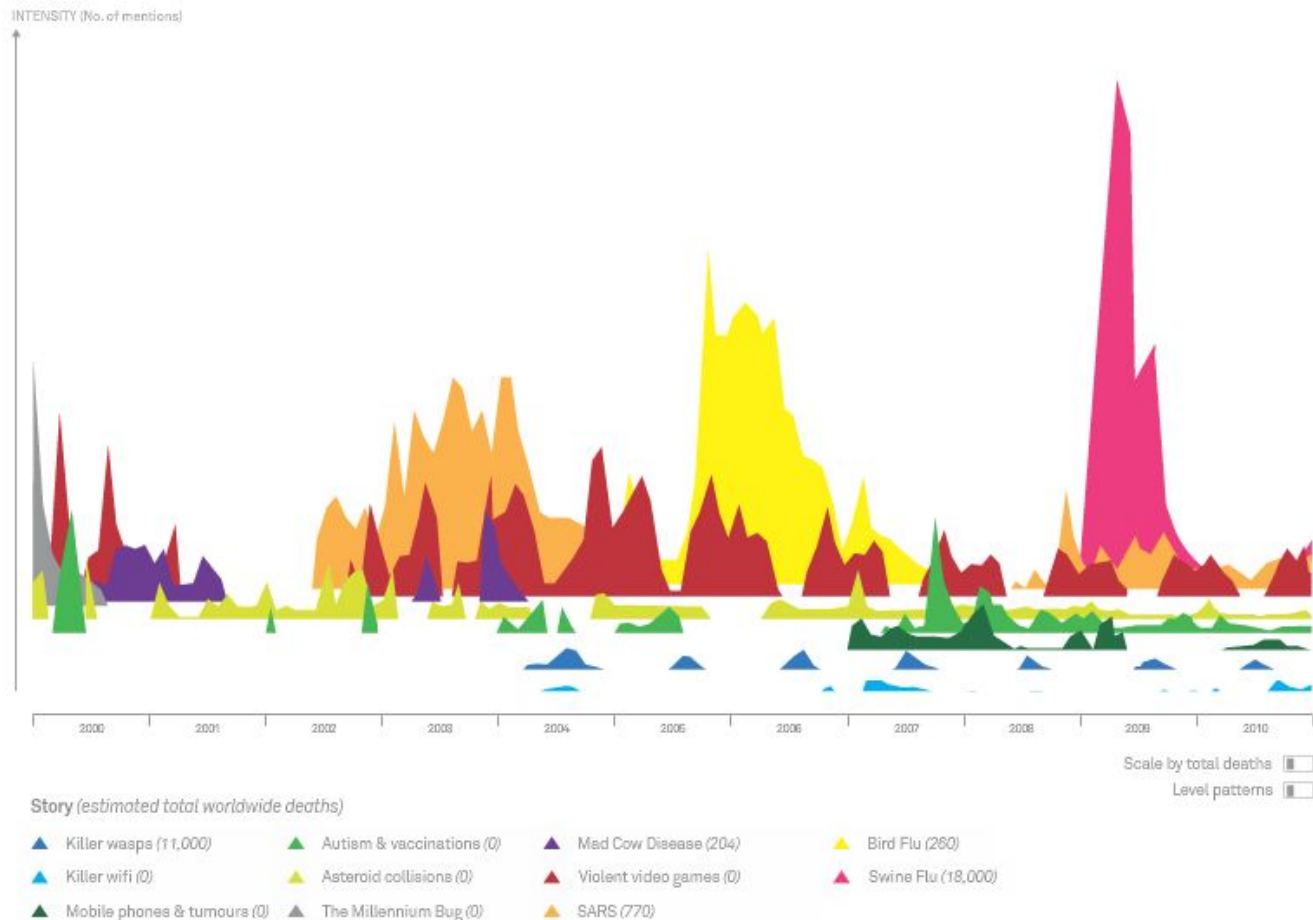
Figure 11-1 (Page 1 of 1)
Best percentage change from baseline in sum of longest diameters and best overall response as per investigator by prior treatment (Full analysis set)



- * 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?



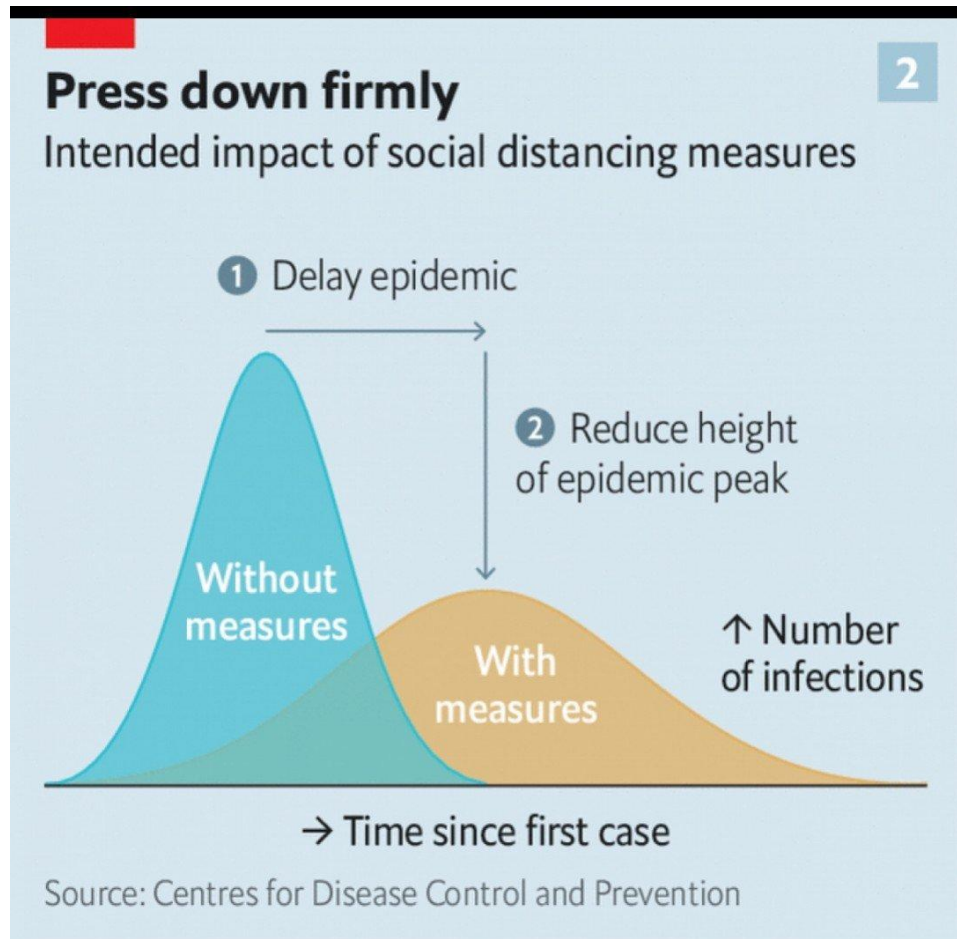
<https://informationisbeautiful.net/>

David McCandless & Joshua Lee // informationisbeautiful.net // @infobeautiful // v1.0
source: Google Insights & News Timeline // note: Seasonal flu has killed around 3.75m people since 2001

Mountains Out of Molehills
A timeline of global media scare stories

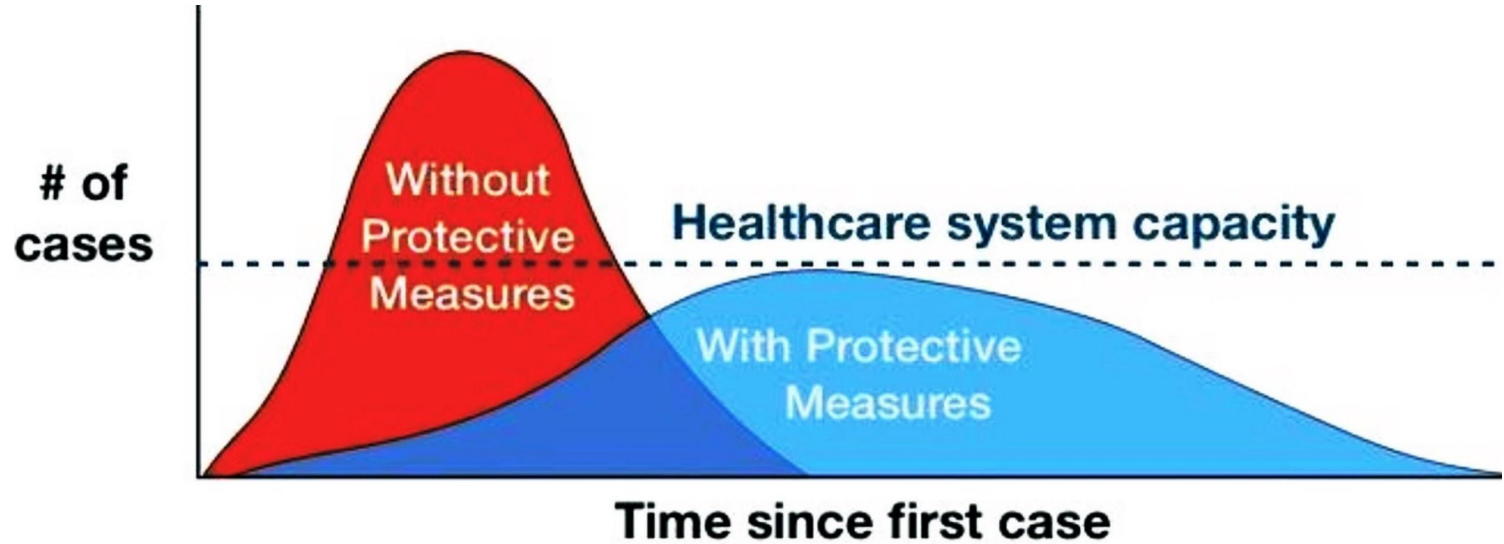
Beautiful and effective?

Need to [#flattenthecurve](#)



The Economist

Even more effective?



Adapted from CDC / The Economist

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

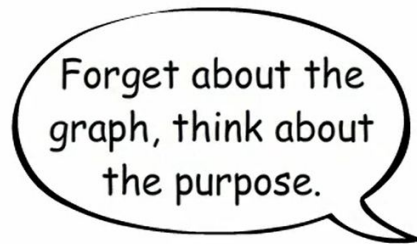
- Avoid misrepresentation (use appropriate scales)
- Choose the appropriate graph type to display your data
- Maximize data to ink ratio (reduce distraction, less is more)

Make the message obvious

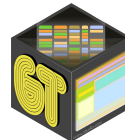
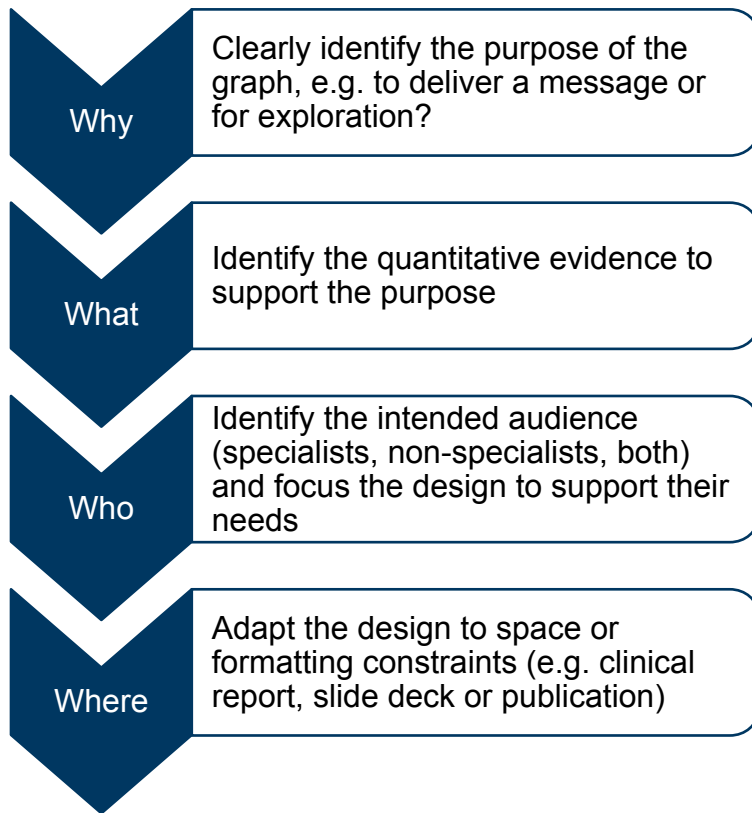
- Use proximity and alignment to aid in comparisons
- Minimize mental arithmetic (e.g. plot the difference)
- Use colors and annotations to highlight important details



Law 1
Have a clear purpose



Have a clear purpose



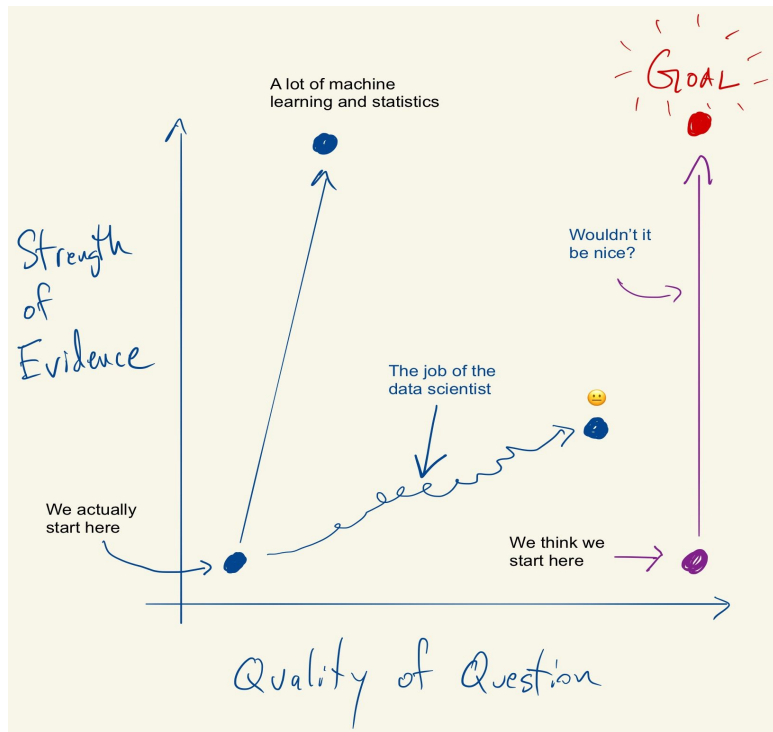
Tukey, Design Thinking, and Better Questions

👤 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



What type of graph do I want to create?

EXPLORATORY

“I want to dig into the data”

“I want to get familiar with the data”

“I want to find the story in my data”

The audience is:
YOU

EXPLANATORY

“I want to communicate the results”

“I want to tell the story behind the data”

The audience is:
SOMEONE ELSE

Do you want your audience to play 'Where's Wally?'



Credit Andrew Wright, Novartis

Law 2

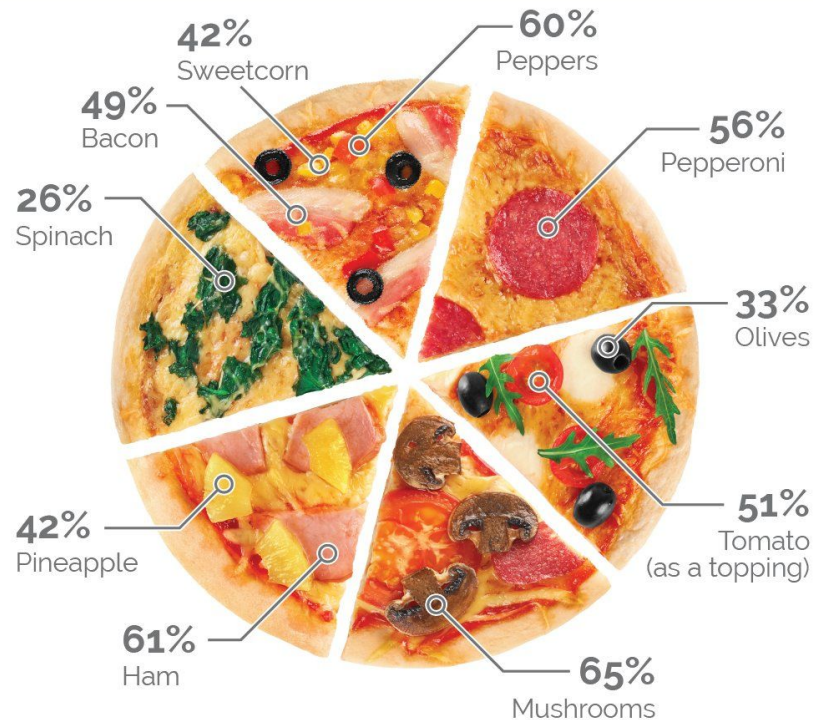
Show the data clearly



Show the data **clearly**

Mushroom is the UK's most liked pizza topping

Generally speaking, which of the following toppings do you like on a pizza? Select as many as you like



Other items not depicted include: onions (62%), chicken (56%), beef (36%), chillies (31%), jalapeños (30%), pork (25%), tuna (22%), anchovies (18%). 2% of people say they only like Margherita pizzas

Show the data **clearly**

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



YouGov ✓
@YouGov



We're very sorry for the confusion, but this is NOT a pie chart - it is just a top-down photo of a pizza with some topping stats pointed out



YouGov ✓ @YouGov · Mar 6, 2017

Forget pepperoni - mushroom is Britain's most liked pizza topping (65%), followed by onion (62%) and then ham (61%) yougov.co.uk/news/2017/03/0...

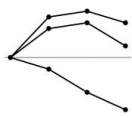


2:56 PM · Mar 6, 2017 · Twitter Web Client

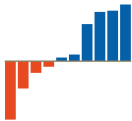
Choosing the Correct Graph Type Aids in interpretation

Deviation

Chg. from baseline

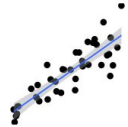


Waterfall

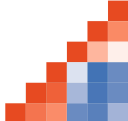


Correlation

Scatter plot

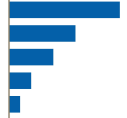


Heat map

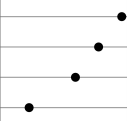


Ranking

Horizontal bar chart

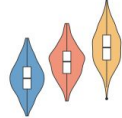


Dotplot



Distribution

Boxplot

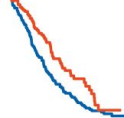


Histogram



Evolution

Kaplan Meier



Line plot



Part-to-whole

Stacked bar chart

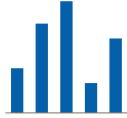


Tree map

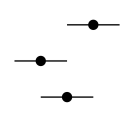


Magnitude

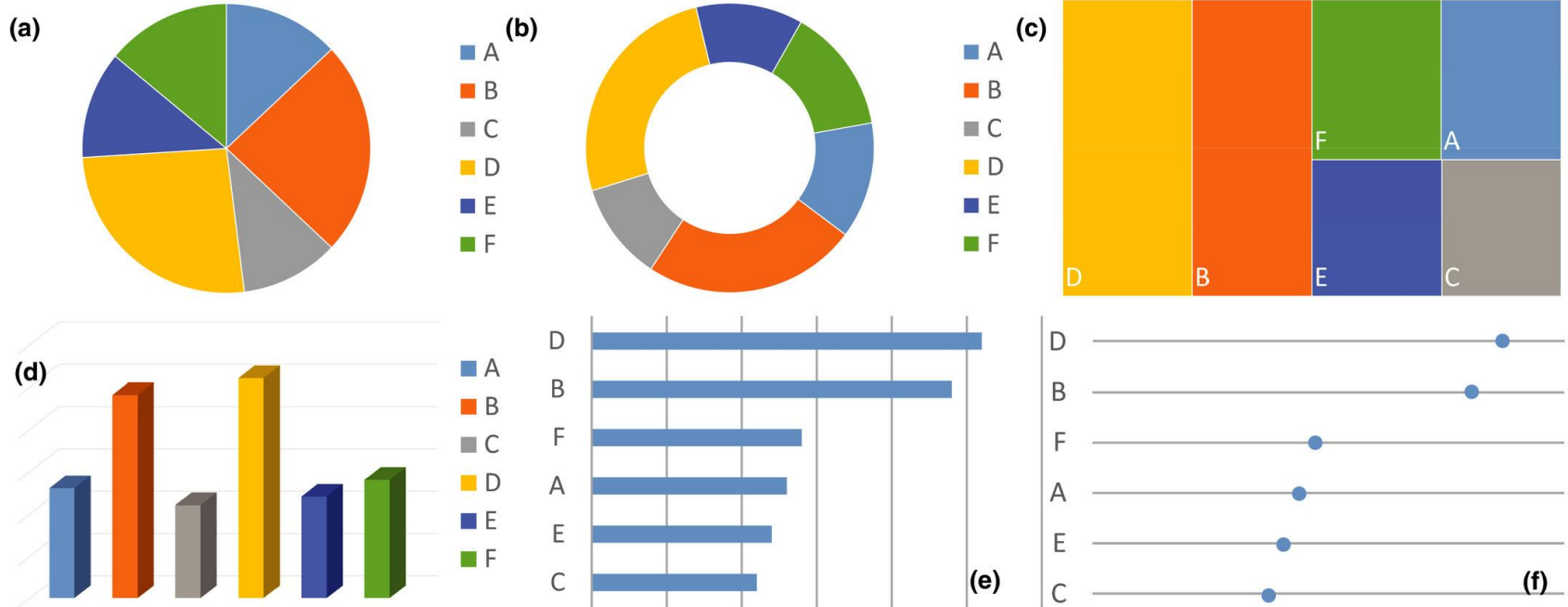
Vertical bar chart



Forest plot

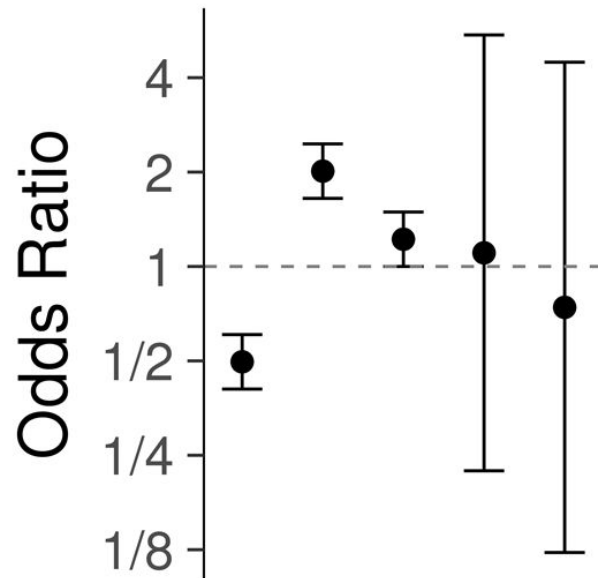
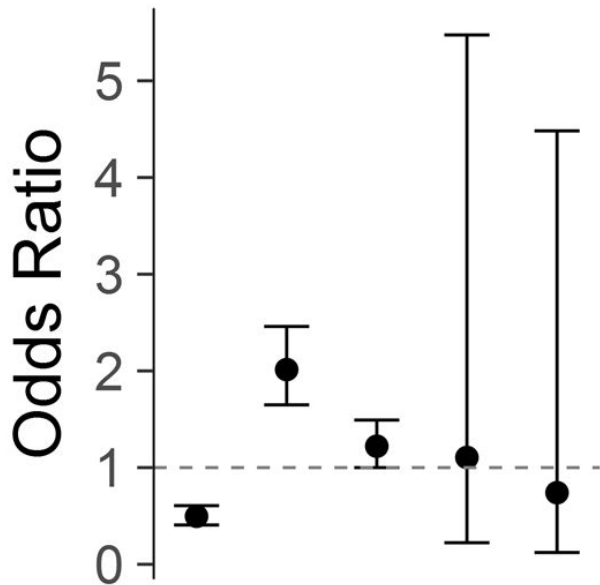


Choosing the Correct Graph Type Aids in interpretation



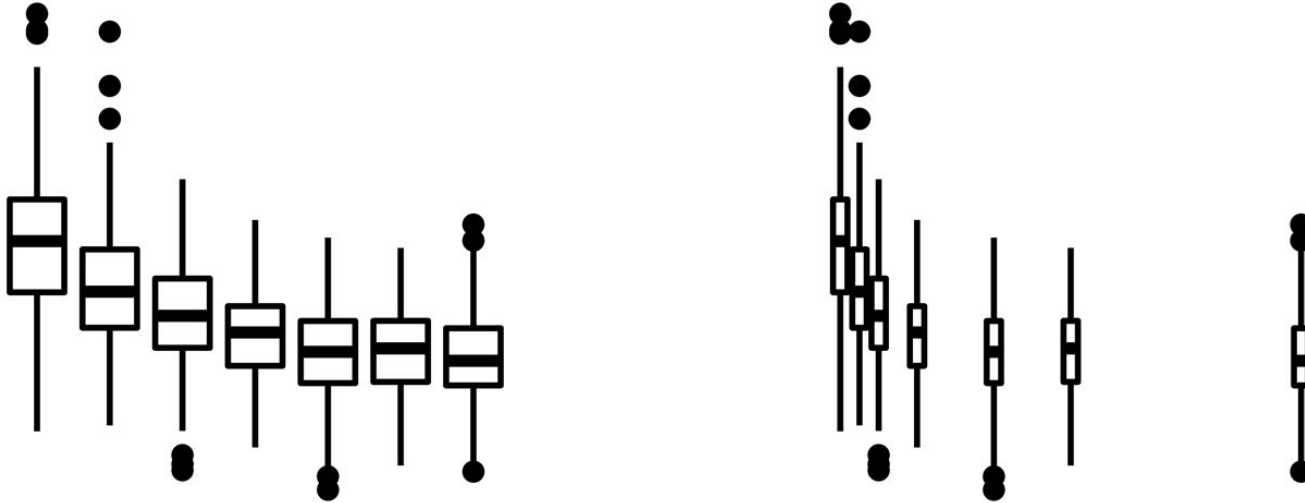
Choose the right scale for your data

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



Space measurements proportional to the time between each

Measurements displayed close together are perceived to be closer in time

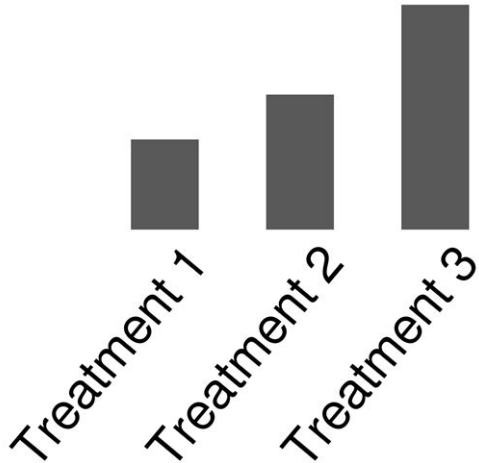


Law 3
Make the message obvious



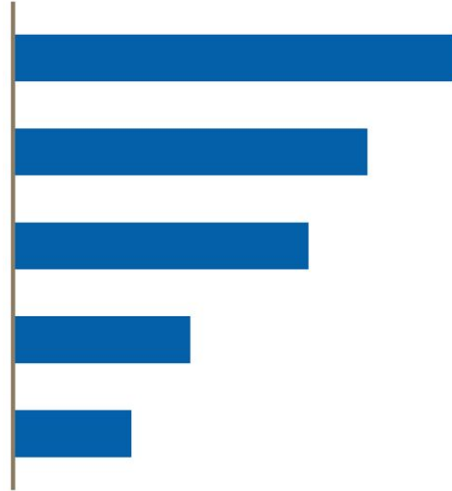
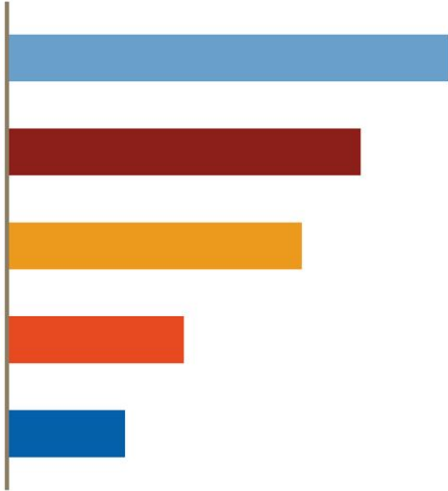
Try not to set text at an angle

Think of alternatives such as transposing the graph



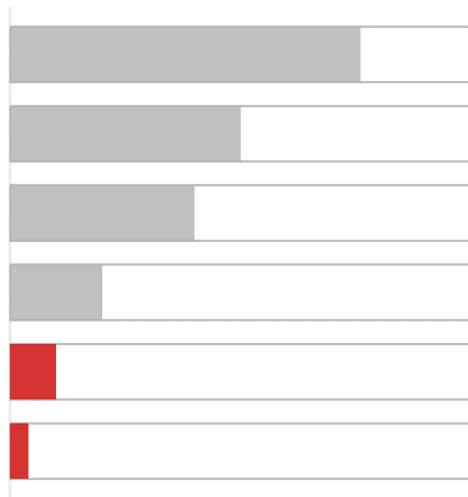
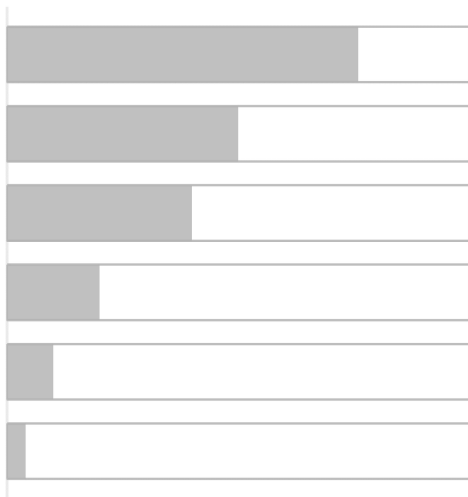
Avoid unnecessary color...

Avoid using color to differentiate between categories of the same variable

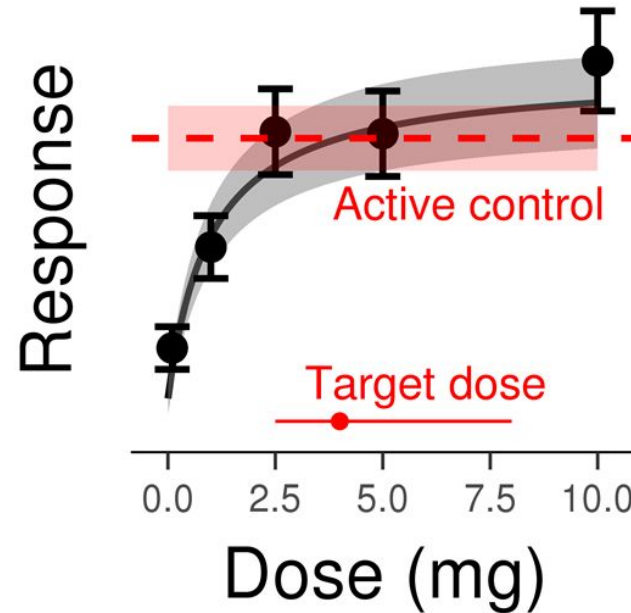
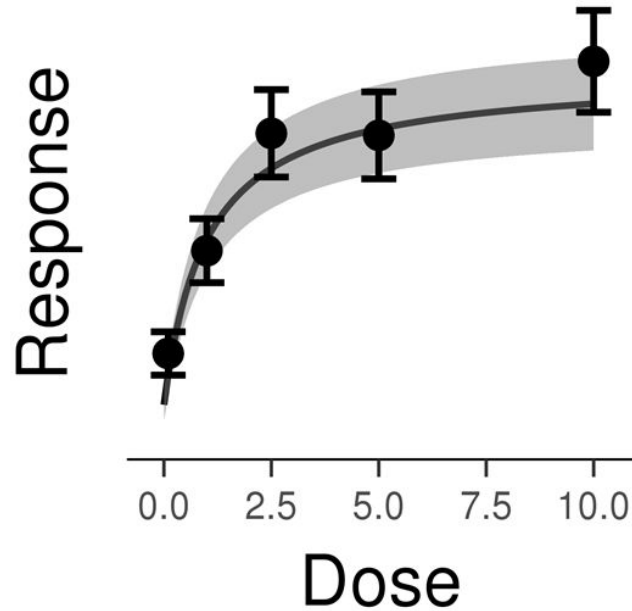


Only use color when it adds value

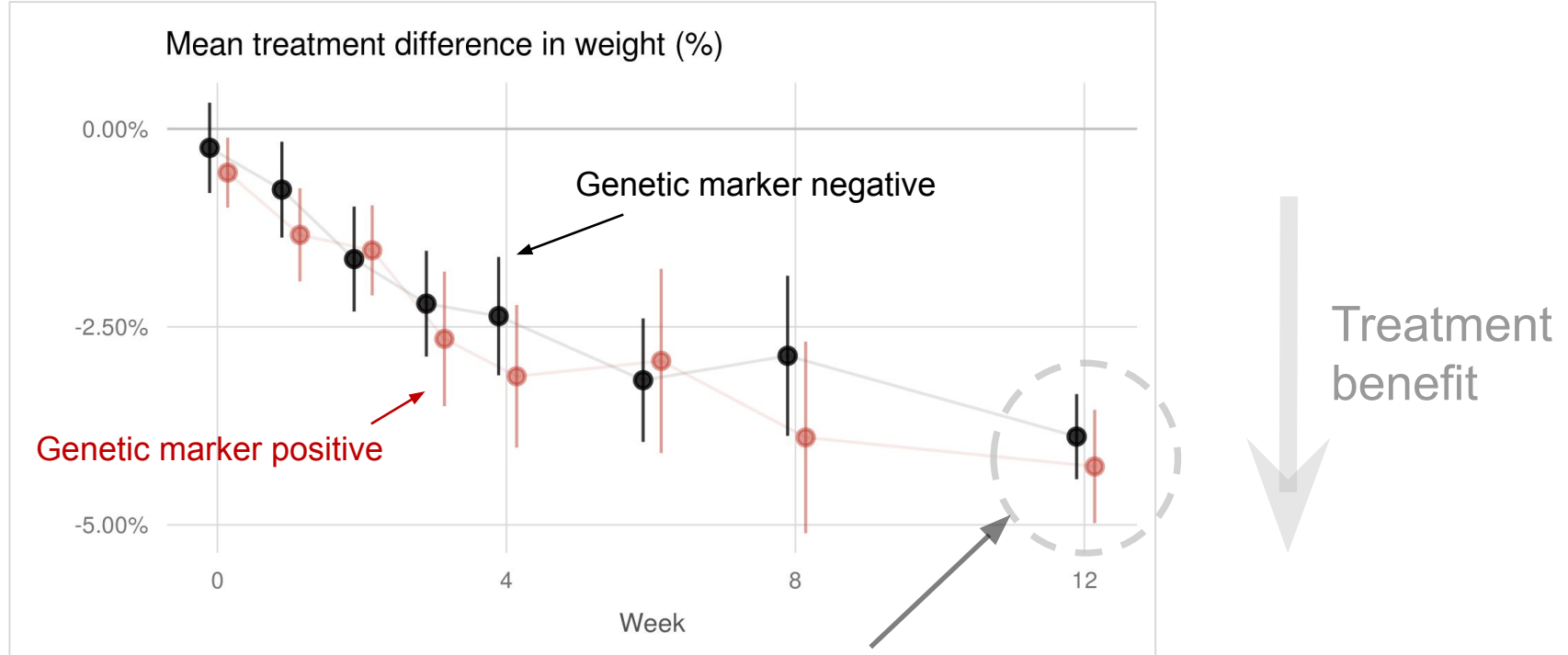
Use a bold, saturated or contrasting color to emphasize important details



Use informative labels and annotations to support the message



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

Graphics Principles Cheat Sheet v1.1

Communication

Effective visualizations communicate complex statistical and quantitative information facilitating insight, understanding, and decision making.

But what is an effective graph?

This cheat sheet provides general guidance and points to consider.

Planning

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

Identify the quantitative evidence to support the purpose

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

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

Principles of Effective Graphic Design

Proximity – group related elements together
Alignment – elements on the same vertical or horizontal plane are perceived as having similar properties
Simplicity – cut anything superfluous, only include elements that add value, limit to 2-3 colors or fonts

White space (empty space) – use white space to minimize distraction & provide clarity
Legibility – sans serif fonts are easier to read, use color for emphasis instead of a new typeface

Color – select colors that present enough contrast to make the graph legible. Choose monochromatic color schemes to prevent clashing. Use dark colors and accent colors to emphasize important information

Visual Hierarchy – use color, font, image size, typeface, alignment & placement to create a viewing order

Focal Points – primary area of interest that immediately attracts the eye, emphasize the most important content and make it your focal point. Use contrasting colors to draw attention

Repetition – repeating elements can be visually appealing, repeated shapes, labels, colors

Familiarity – using familiar styles, icons, navigation structure makes viewers feel confident

Consistency – be consistent with heading sizes, font choices, color scheme, and spacing. Use images with similar style

Effectiveness Ranking

A graph is a representation of data that visually encodes numerical values into attributes such as lines, symbols and colors. The Cleveland-McBili scale can be used to select the most effective attributes for your purpose.

Volume Color hue Depth, 3d position Color intensity Area Slope or Angle Length Position on unanalysed scale Position on common scale



Least accurate

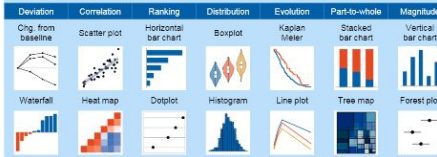
volume charts
poorly designed heat maps
multivariate density plots
bubble charts, mosaic charts
line graphs, pie charts, waterfall chart
stacked bar charts
small multiple plots
dot plots, bar charts, parallel coordinate plots

Most accurate

Selecting the right base graph

Consider if a standard graph can be used by identifying suitable designs based on the: (1) purpose (i.e. message to be conveyed or question to answer) and (2) data (i.e. variables to display).

Example plots categorized by purpose



Facilitating Comparisons

Proximity improves association

Place labels next to data instead of data legends

Group together elements to be compared directly

Ease visual inspection

Order values to help compare across many categories

Judgments are easier to make on a common vertical scale

Reduce mental arithmetic

Plot the final comparison e.g. mean difference not two means

Use reference lines and other visual anchors

Color for emphasis or distinction

Restrained use of color is highly effective in organizing a narrative and calling attention to certain elements.

Think carefully before introducing additional color.

Do not use color to differentiate between categories of the same variable

Use colors or shades to represent meaningful differences such as positive/negative values, treatments or doses

Be consistent, use the same color to mean the same thing in a series of graphs (e.g. treatment, dose)

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

Emphasize the data by minimizing unnecessary ink, e.g. soften gridlines with a light color

Utilize existing resources for selection of appropriate palettes such as Color Brewer or Munsell

Implementation Considerations

Plot causes on the x-axis and effect on the y-axis. Use this standard convention in order to avoid misinterpretation.

Aspect ratio can influence interpretation. Aim for a 4:5 diagonal ratio of change to avoid over-interpretation of slope.

Use position for comparisons rather than length (i.e. dots instead of bars, especially for non-linear scales (e.g. log scale or % change)).

Do not plot log-normally distributed variables on a linear scale (e.g. hazard ratio, AUC, CI).

When displaying data measured on the same scale, also plot them on the same scale for easy comparison.

Connected data imply continuity. Do not connect data across a disconnected or uneven time scale.

Visits displayed close together are perceived to be closer in time. Space the visits proportional to the time between each in order to avoid confusion. Location denotes up or down

Plot data and inferences to support stories about models.

Putting it all together – Remove the clutter & emphasize the message

Creating a graph is an iterative process: produce, review and refine.

Colors, backgrounds, and borders can be removed and try using other methods to distinguish different curves.

It is easier to see differences in position over a difference in length, i.e. a dot over a bar.

Using too many colors can be distracting. Use white background and try using other methods to distinguish different curves.

One solution could be repeating the data in different panels, highlighting individual curves in a darker color.

Legibility and Clarity

Effective graphs stand alone. They use titles, annotations, labels, shapes, colors, and textures to deliver important information.

Label axes with clear measurement units and provide annotations that support the message.

Use font size to create hierarchy (e.g. sat titles 2x larger than all other labels to make them more prominent).

Do not type too small or too condensed. Break long titles into two lines. Shift or adjust size of labels that overlap.

Keep the font style simple – sans serif is easier to read.

Display text with enough contrast to be visible. Favor the use of dark on light instead of light on dark whenever possible.

Bold or italics should only be used for layering or emphasis. Emphasizing everything means nothing gets emphasized.

Try not to test text at an angle, as this decreases readability. Think of alternative solutions such as transposing the graph.

Good graph checklist

- Clear Communication**
 - Is the message of the graph as clear as possible?
 - Is it easy for someone unfamiliar with the data to interpret the graph?
 - Are the patterns/relationships easily identified?
 - Is the graph tailored to its primary purpose and audience?
 - Is the correct graph type used?
- Facilitating Comparisons**
 - Are elements to be compared grouped together?
 - Are labels placed next to data instead of in legends?
 - Have categories been ordered for easy comparison?
 - Can the plot be read without doing mental calculations?
 - Are the estimates of interest plotted (e.g. mean differences with confidence intervals)?
- Color for emphasis or distinction**
 - Are graphical elements displayed in a dark color on a light background?
 - Are grid lines drawn with a thin line and a light color such as grey?
 - Are colors used sparingly (e.g. max 3)?
 - Do all elements in the graph have a purpose (e.g. points, textures, grid lines)?
 - Are the same colors used to mean the same thing in a series of graphs?
- Implementation Considerations**
 - Are multiple panels plotted on the same scale?
 - Are log-normally distributed variables plotted on a log scale?
 - Are common baselines used wherever possible?
 - Does the orientation of the axes aid interpretation?
 - Does the aspect ratio allow the reader to see variations in the data?
 - Are data across a disconnected time scale kept disconnected?
 - Are data spaced proportionally to the actual time interval (instead of according to visit number)?
 - Are data and inferences plotted to support stories about models?
 - Are number of patients by group reported if this adds context?
- Legibility and Clarity**
 - Can all graphical elements be seen?
 - Does the graph have a clear title, axis labels, annotations and data units?
 - Can the font be read without eye strain or effort?
 - Are sans-serif fonts used?
 - Do text sizes have correct hierarchy (big to small, main text to subtext)?
 - Are the elements of the graph clearly labeled (e.g. points, error bars, shaded regions)?
 - Are labels oriented horizontally where possible?

Resources

Books
C. R. Tufte. The visual display of quantitative information. Cheshire, CT: Graphics Press, 2001.
Goodman, W.S. and D.C. Robertson. Graphical perception: theory, experimentation and application to the development of graphical methods. JASA, Vol. 70, No. 267, pp. 521 – 554, 1964.
A. Wong, Shou-Wei Ho, Hui-Hsuan, J. Cheng, L. Cheng, and D. Cheng. The Design of Effective Graphs. Springer, 2012.
D. M. Kelley. The Visual Display of Quantitative Information. The Data and Design of Effective Graphs. Springer, 2012.
J. H. Friedman, J. J. Friedman, and J. J. Friedman. Effective communication for rational decisions. PRINCIPAL: J. H. Friedman, J. J. Friedman, and J. J. Friedman. Effective communication for rational decisions. PRINCIPAL: J. H. Friedman, J. J. Friedman, and J. J. Friedman.

Authors

Alison Magnusson, Mark Rafter, Ralf Magnusson, Julie Jones, Brian Vandeweyer, Koenraad



© 2017 Novartis

Where to find to out more?

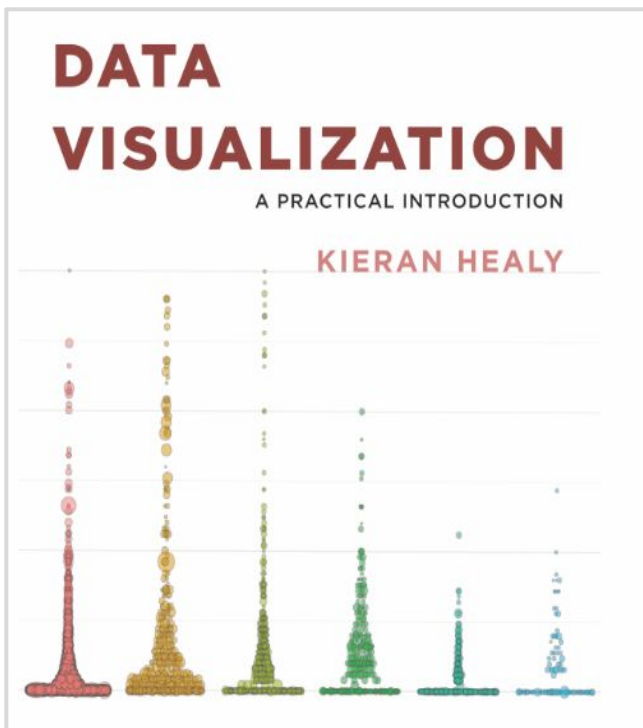
Trees, maps, and theorems

Effective communication for rational minds

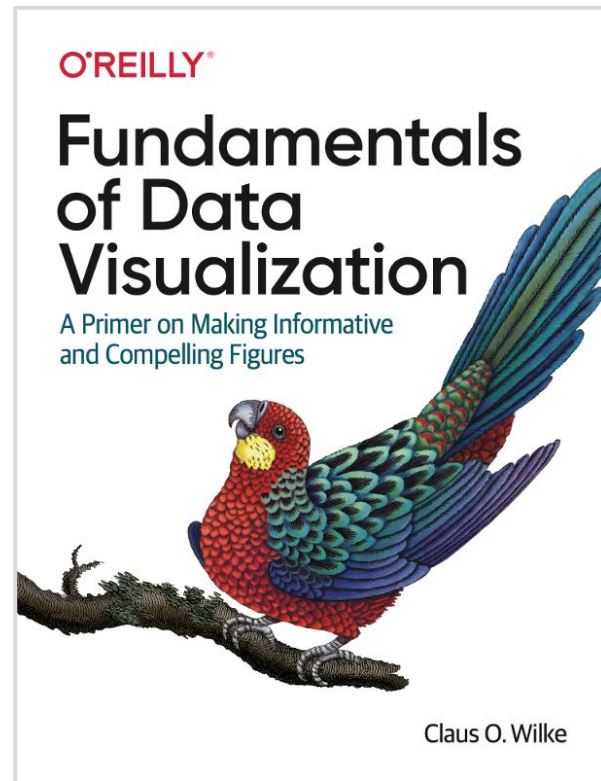
Jean-luc Doumont



<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.

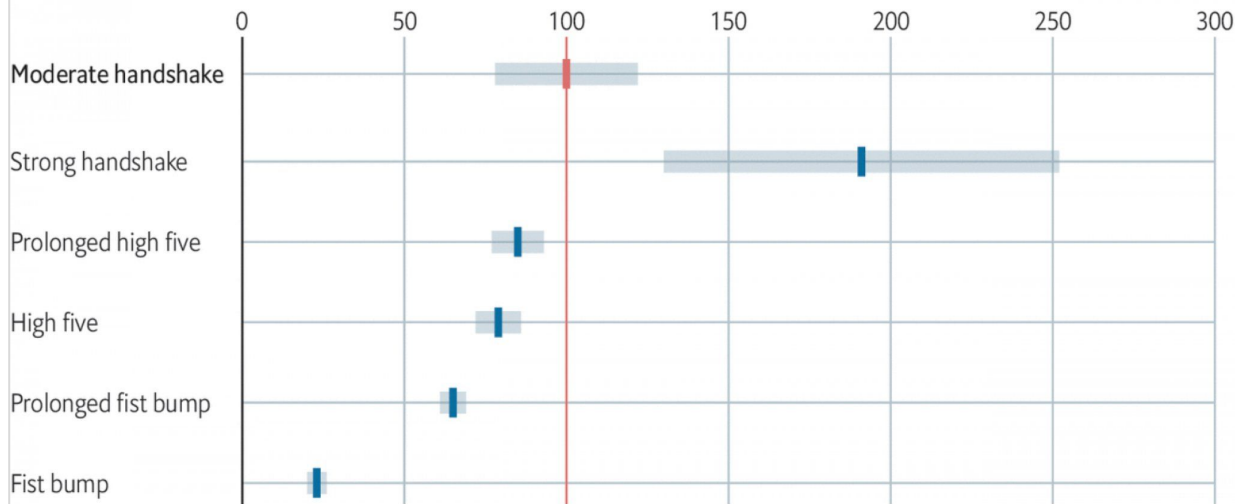
To help design effective visualisations, remember the three laws: purpose, clarity and message



Handover

Pressing the flesh

Transfer of bacteria relative to a moderate-strength handshake, %



Source: "The fist bump: a more hygienic alternative to the handshake" by S. Mela and D. Whitworth, *American Journal of Infection Control* 2014

The Economist

Daily chart

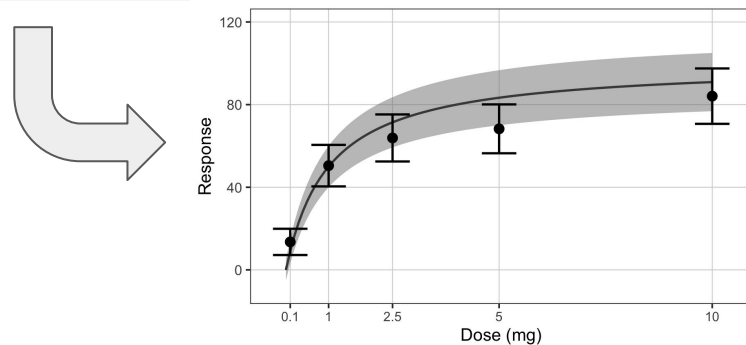
Handshakes are potential harbingers of covid-19

Changing greetings could slow its spread. But handwashing is essential

Graphic detail | Mar 6th 2020

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>

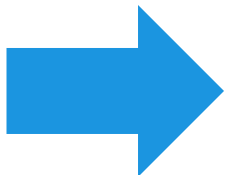
		<Overall>		<exposure 1> N = XXX		<exposure 2> N = XXX			
Characteristics	Parameters or Categories	Value or n	%	Value or n	%	Value or n	%	SMD	p-value
<Quantitative var> (<unit>) ^b	N	XXX	.	XXX	.	XXX	.	X.XX	X.XXX ^a
	Mean	XXX	.	XXX	.	XXX	.		
	Median	XXX	.	XXX	.	XXX	.		
	SD	XXX	.	XXX	.	XXX	.		
	Range	XX-XX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

Seamless integration
into analytics &
reporting workflows



Suitable for analytics
use cases in clinical/
medical development

Combination of ease of
use with flexibility for
complex analyses

**Integrate
graphical
principles in
your analytics
projects**

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

Explore different
visualisations of
analytic data set

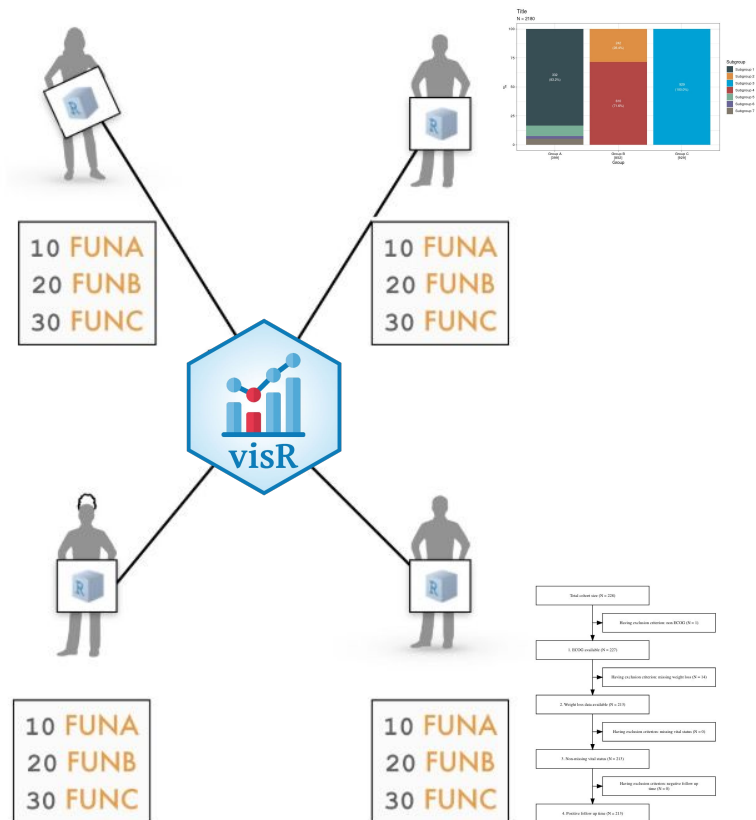
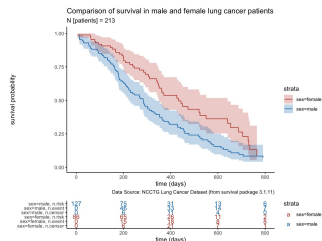
Adaptable to target
audiences without repeating
core analysis

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

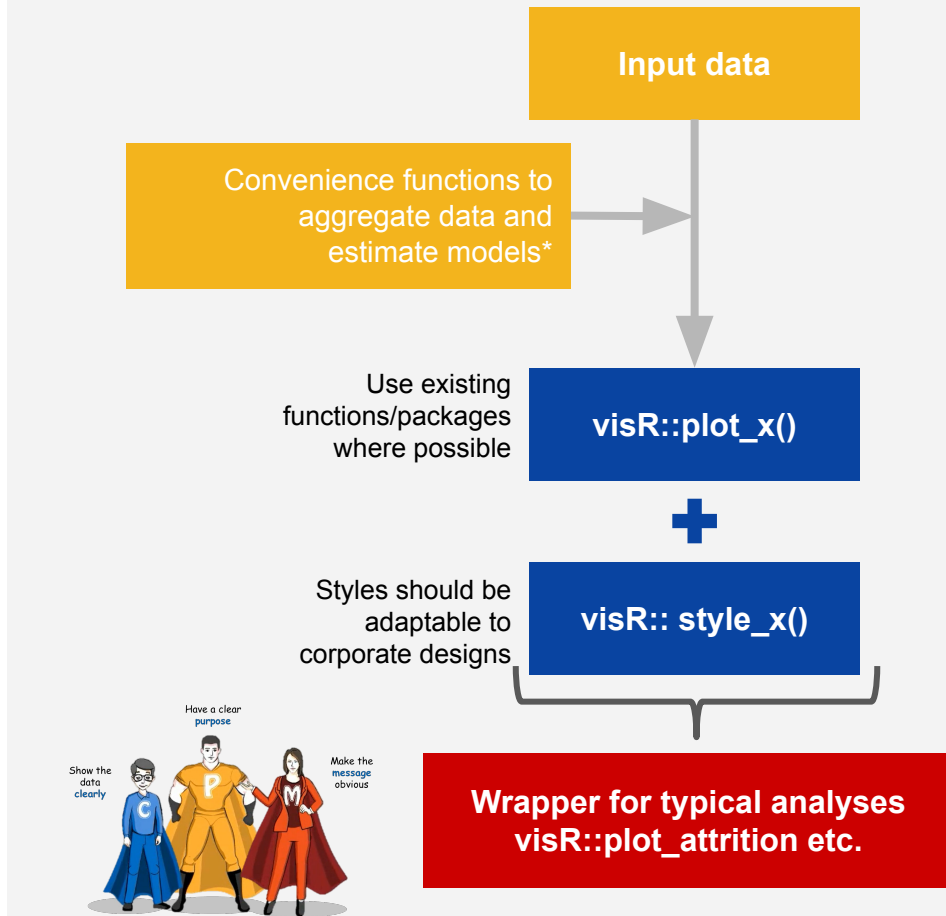
Attrition Table					
Summary of samples fulfilling inclusion/exclusion criteria					
Criteria	Condition	Remaining N	Remaining %	Excluded N	Excluded %
Total cohort size	none	228.00	100.00	0.00	0.00
1. ECOG available	is.na(ECOG)	227.00	99.56	1.00	0.44
2. Weight loss data available	is.na(wt.loss)	213.00	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)



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



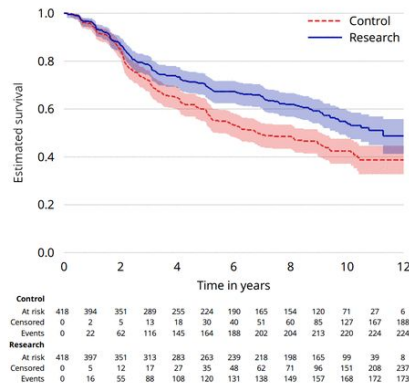
Interim (estimate) data model (e.g. broom)

time	trt	estimate	lower	upper

Visualization
Function
+
Style
Function



Visualization



Typical Time To Event Analysis Workflow

1

Build Analysis Cohort

How many patients are kept after applying inclusion/exclusion criteria?

2

Baseline Characteristics

What are the general characteristics of the population we are analyzing?

3

Estimate Survival Function

What is the probability of a patient having survived at time t given his/her stratum?

Typical Output

Attrition

- Table
- Flow chart

Summary

- “Table One”

Kaplan-Meier

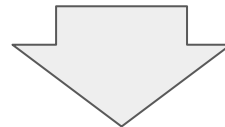
- Kaplan-Meier Plot
- Median Survival

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

```
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

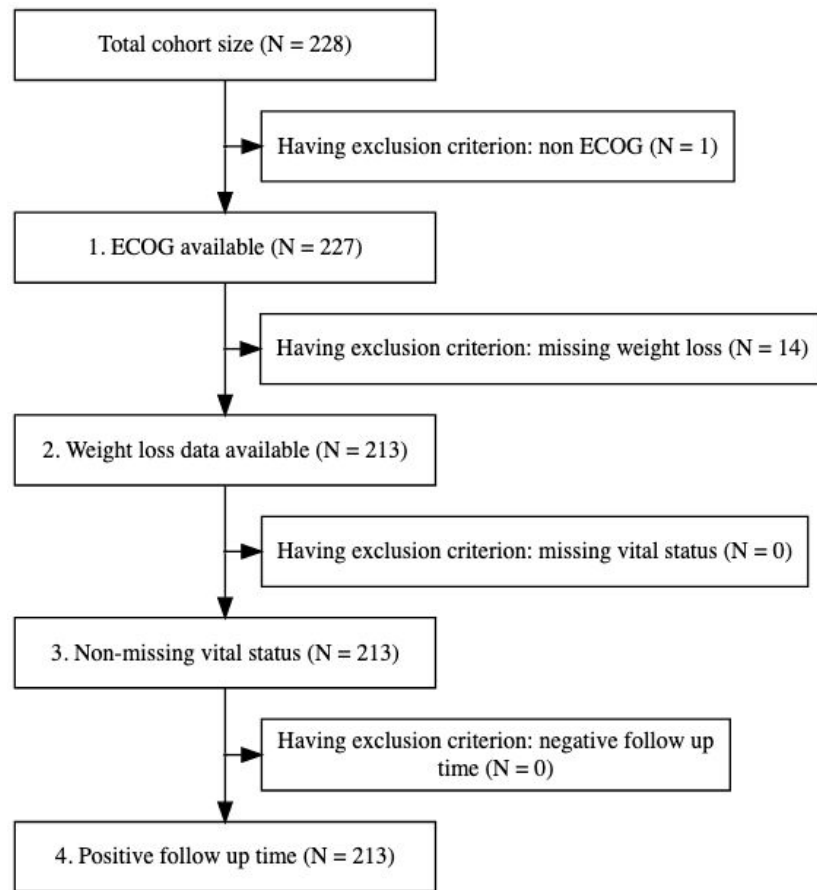
```
complement_descriptions <- c(  
  "Having exclusion criterion: non ECOG",  
  "Having exclusion criterion: missing weight loss ",  
  "Having exclusion criterion: missing vital status",  
  "Having exclusion criterion: negative follow up time"  
)
```

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)	62.5 (9.21)
Median (IQR)	61 (55-68.8)	64 (57-70)	63 (56-70)
Min-max	41-77	39-82	39-82
Missing	0 (0%)	0 (0%)	0 (0%)
ECOG			
0 asymptomatic	26 (30.2%)	35 (27.559%)	61 (28.638%)
1 ambulatory	41 (47.7%)	65 (51.181%)	106 (49.765%)
2 in bed less than 50% of day	19 (22.1%)	26 (20.472%)	45 (21.27%)
3 in bed more than 50% of day	NA	1 (0.787%)	1 (0.469%)
Karnofsky			
Mean (SD)	82.6 (12)	81.9 (12.5)	82.2 (12.2)
Median (IQR)	80 (80-90)	80 (70-90)	80 (80-90)
Min-max	50-100	50-100	50-100
Missing	0 (0%)	0 (0%)	0 (0%)
Age Group			
30-50y	12 (14.0%)	13 (10.2%)	25 (11.7%)
51-70y	59 (68.6%)	85 (66.9%)	144 (67.6%)
> 70y	15 (17.4%)	29 (22.8%)	44 (20.7%)

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

Rendering
as gt table

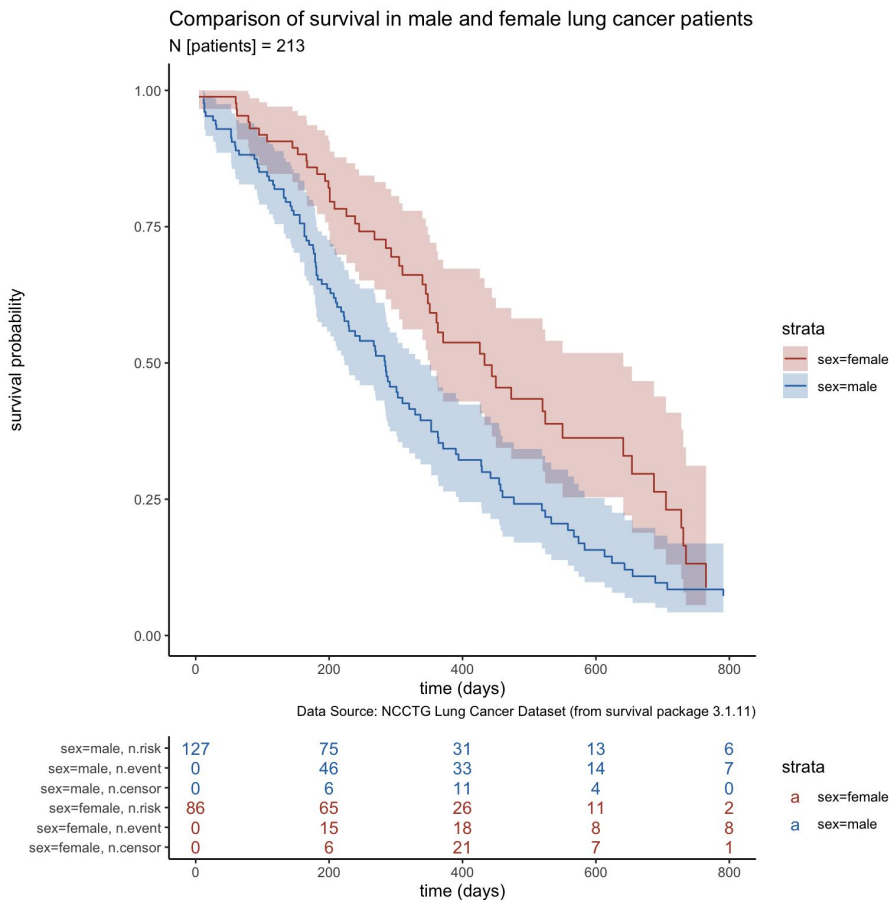
Copy	CSV	Excel	Filter: <input type="text"/>			
Overview over Lung Cancer patients - Baseline characteristics of study cohort stratified by treatment type						
	variable	statistic	female	male	Overall	
1	Sample	N	86	127	213	
2	time	Mean (SD)	345 (205)	295 (216)	315 (213)	
3	time	Median (IQR)	294 (201-467)	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)						

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

```
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

```
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

```
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)

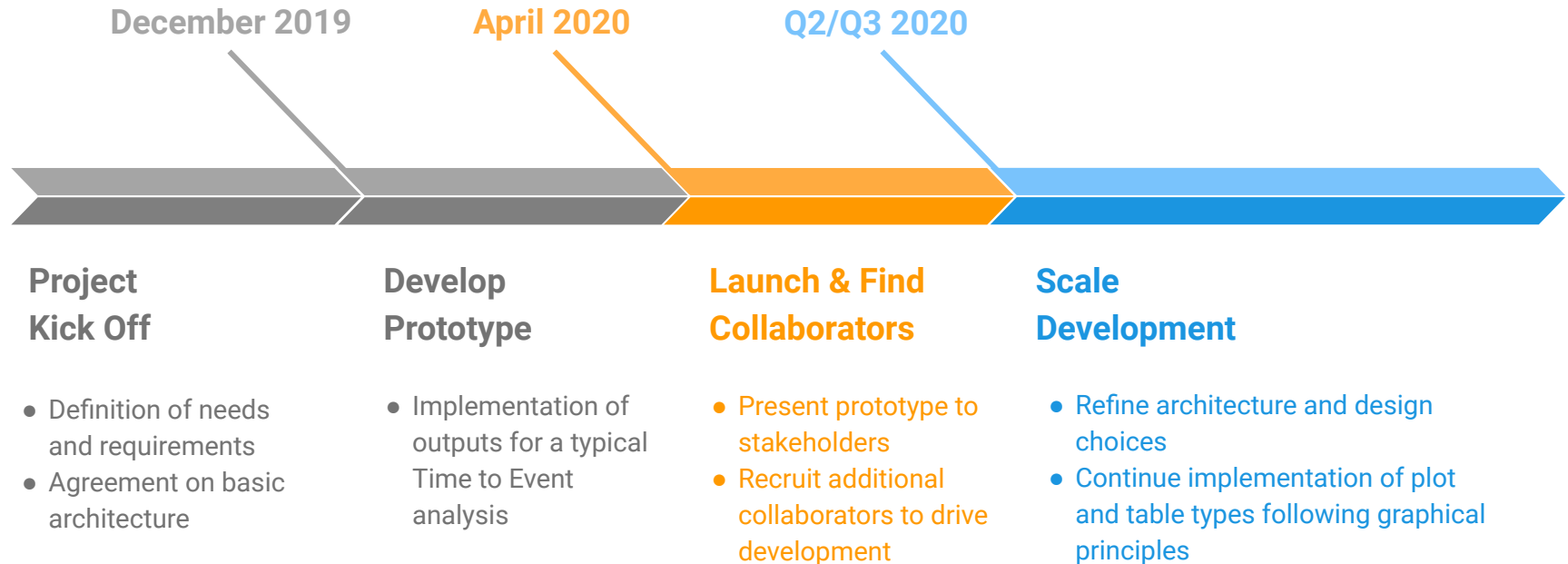
- We have also implemented other convenient estimation functions.
- Median survival times by stratum.
- Multiple methods for testing equality between strata.

```
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



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

Acknowledgements



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