visR

A Package for Effective Visualizations in Pharma

Charlotta Früchtenicht
On behalf of the visR collaboration team
http://openpharma.github.com/visR

NOVARTIS

Roche

Agenda

Intro and Motivation

Design Considerations

Package Architecture

Example Time to Event Analysis

NOVARTIS

Roche

Effective data visualisation is effective **visual** communication

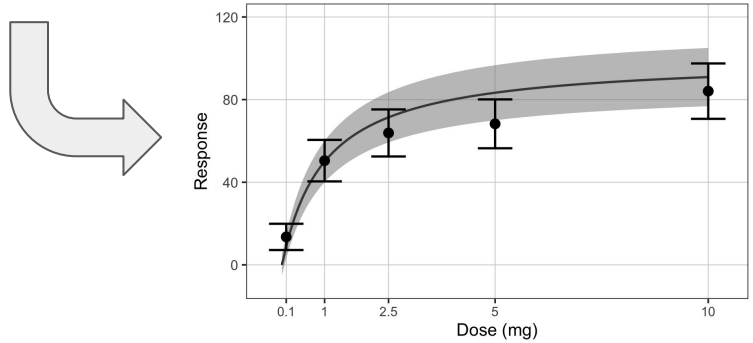
Effective visualisations

- 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

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>

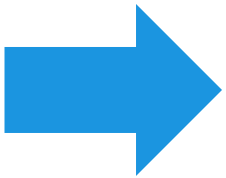
		<Overall>		<exposure 1> N = XXX		<exposure 2> N = XXX			
Characteristics	**Parameters or Categories**	**Value or n**	**%**	**Value or n**	**%**	**Value or n**	**%**	**SMD**	**p-value**
<Quantitative var> (<unit>) [b]	N	XXX	.	XXX	.	XXX	.	X.XX	X.XXX[a]
	Mean	XXX	.	XXX	.	XXX	.		
	Median	XXX	.	XXX	.	XXX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

Design Considerations

Seamless integration into analytics & reporting workflows

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

Integrate graphical principles in every analytics projects

Explore different visualisations of analytic data set

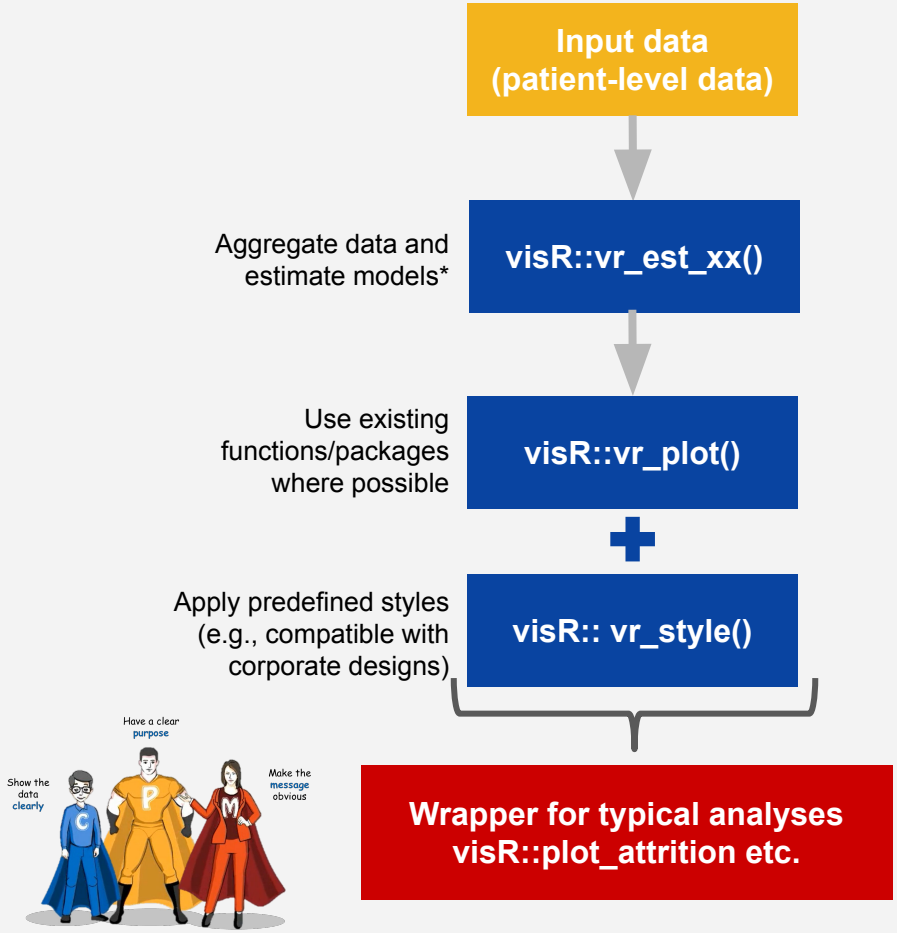
Suitable for analytics use cases in clinical/ medical development

Combination of ease of use with flexibility for complex analyses

Adaptable to target audiences without repeating core analysis

Package Architecture

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



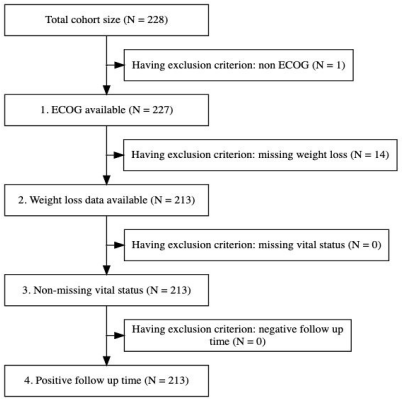
**Functions for survival models, p-values, confidence intervals,... make available separately & allow to call on patient-level data*

Typical Time To Event Analysis Workflow

1 Build Analysis Cohort

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

Attrition Chart



2 Baseline Characteristics

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

Summary Table

Overview over Lung Cancer patients

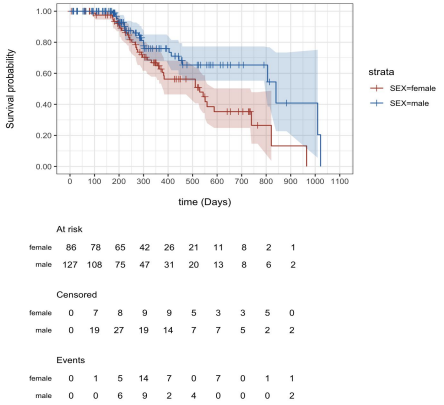
Baseline characteristics of study cohort stratified by treatment type

	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			

3 Estimate Survival Function

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

Kaplan-Meier Plot



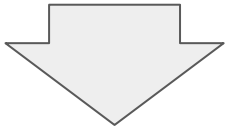
1 Build Analysis Cohort

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 censoring status",
                            "4. Positive follow up time"),
  criteria_conditions   = c("!is.na(SUBGR1)",
                            "!is.na(WEIGHT)",
                            "!is.na(CNSR)",
                            "AVAL >= 0"),
  subject_column_name   = 'USUBJID')


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(SUBGR1)	227.00	99.56	1.00	0.44
2. Weight loss data available	!is.na(WEIGHT)	213.00	93.42	14.00	6.14
3. Non-missing censoring status	!is.na(CNSR)	213.00	93.42	0.00	0.00
4. Positive follow up time	AVAL >= 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)

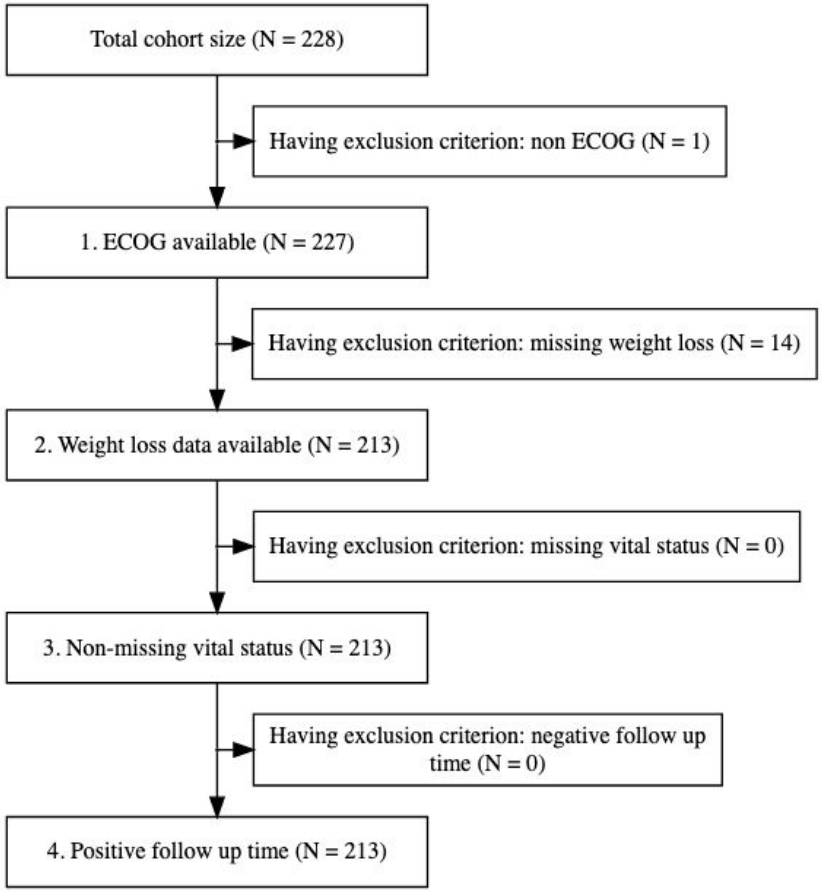
Quickly convert attrition table into flow diagram

```
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")
```

Rendering as gt table

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 DT 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)

3 Survival Analysis

- Based graphical principle findings from the Kmunicate study by Morris *et al.* conducted among 1176 researchers about what the perfect Kaplan-Meier plot should look like[1]
- 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.

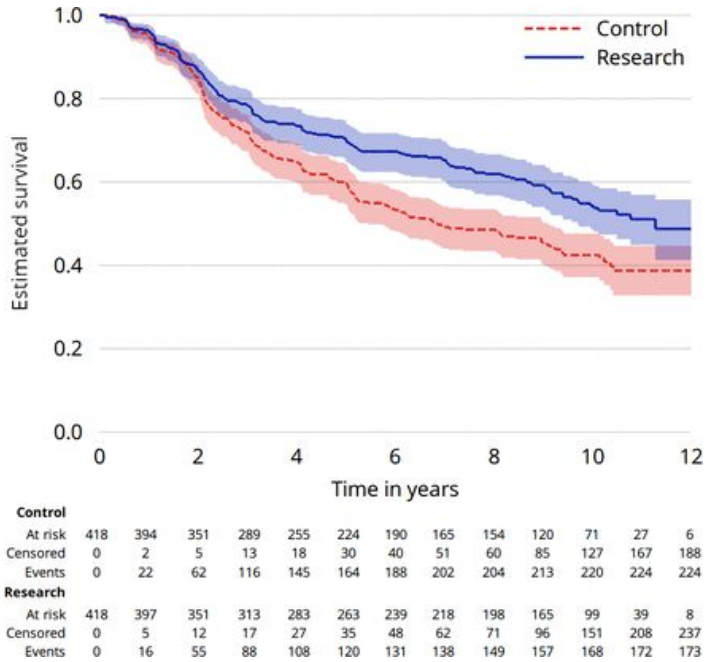
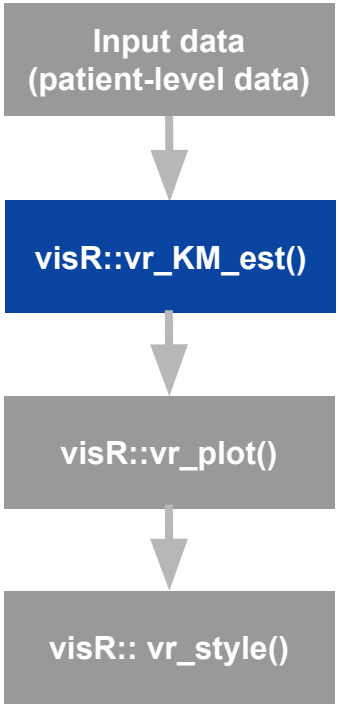


Fig. Best practice recommendation in KMunicate study include an extended risk table of numbers and Confidence Intervals (from Morris *et al.*, BMJ Open, 2019)

1 A package solely focussing on KM plots following the KMunicate study is now available: https://ellessenne.github.io/KMunicate-package/

3 Survival Analysis

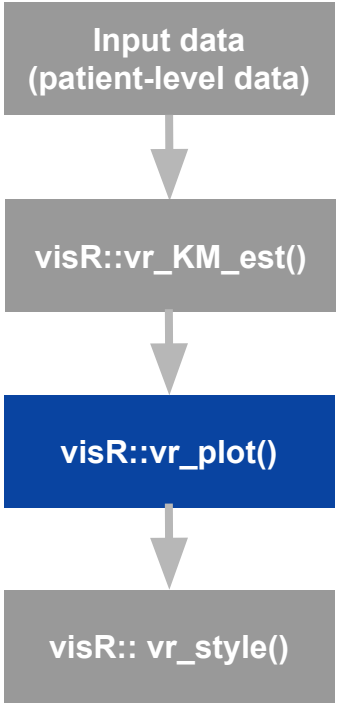



```
lung_cohort %>%
  vr_KM_est(strata = "SEX", conf.int = 0.90)
#> Call: survfit(formula = Surv(AVAL, 1 - CNSR) ~ SEX, data = lung_cohort,
#>     conf.int = 0.9)
#>
#>               n events median 0.9LCL 0.9UCL
#> SEX=female   86     37    529    382    740
#> SEX=male    127     25    840    806     NA
```

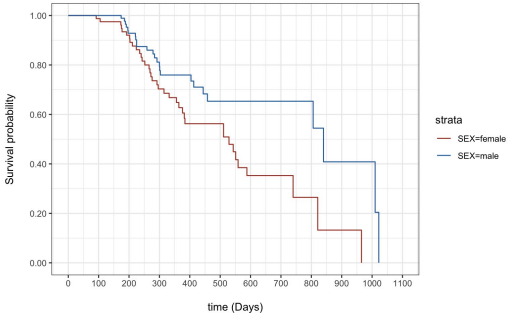
Features

- Use of controlled terminology (CDISC ADaM compatible) reduces number of required arguments over direct survfit-call
- Always assign strata (minimally “Overall”) for consistent output format between stratified and non-stratified analyses
  → simplify downstream analyses
- Retaining call information including name of source data set even when using purrr and %>%

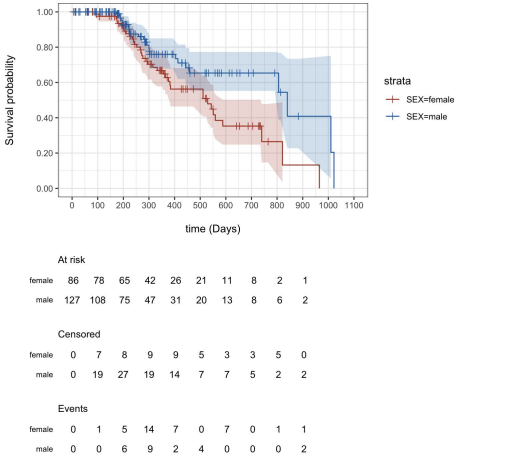
3 Survival Analysis




```
lung_cohort %>%
  vr_KM_est(strata = "SEX", conf.int = 0.90) %>%
  vr_plot(legend_position = "right", x_unit = "Days")
```




```
lung_cohort %>%
  vr_KM_est(strata = "SEX", conf.int = 0.90) %>%
  vr_plot(legend_position = "right", x_unit = "Days") %>%
  add_CI(alpha = 0.2, style = "ribbon", linetype = 3) %>%
  add_CNSR(shape = 3, size = 2) %>%
  add_risktable(min_at_risk = 0,
                display= c("n.risk", "n.censor", "n.event"),
                title = c("At risk", "Censored", "Events"),
                collapse = F
               )
```



Acknowledgements

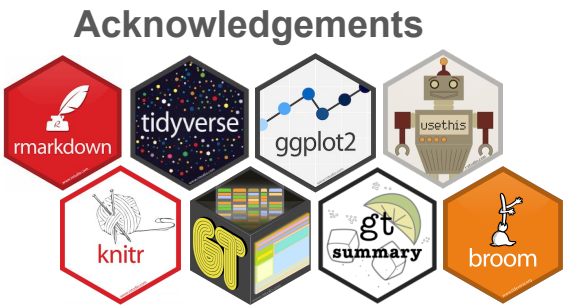
visR
http://openpharma.github.com/visR

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

visR Core Team
Rebecca Albrecht-Dietsch, Mark Baillie, James Black, Charlotta Früchtenicht, Steven Haesendonckx, Thomas Neitmann, Alexandra Papadopoulou, Diego Saldana, Tim Treis, Marc Vandemeulebroecke

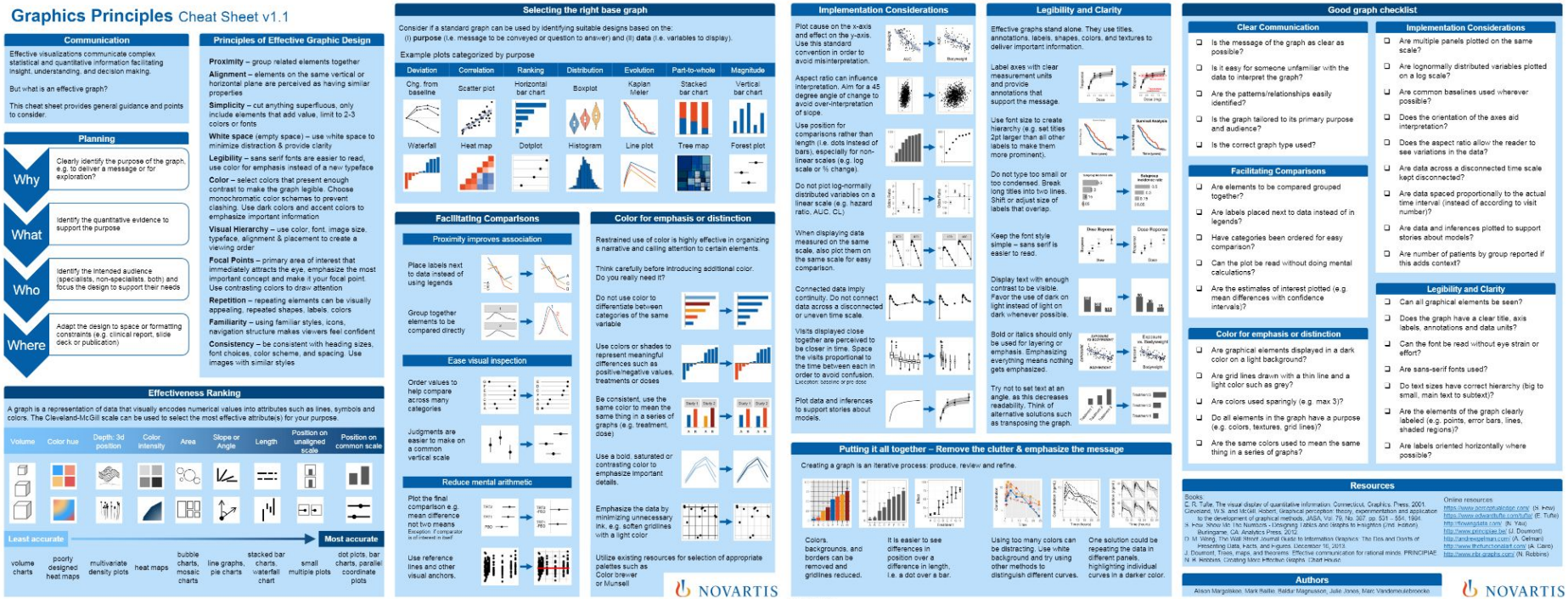
Acknowledgments
Thanos Siadimas, Pawel Kawski, Janine Hoffart, Baldur Magnusson, Alison Margolskee

Want to get involved?
Email: mark.baillie@novartis.com &
charlotta.fruechtenicht@roche.com



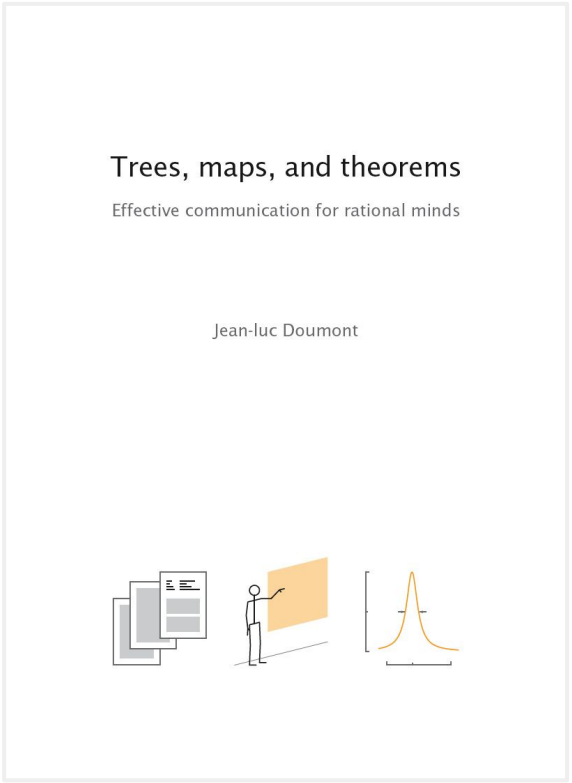
Survival, DT Datatables, kable & kableExtra, ggpubr, and many more

Principles for effective visual communication

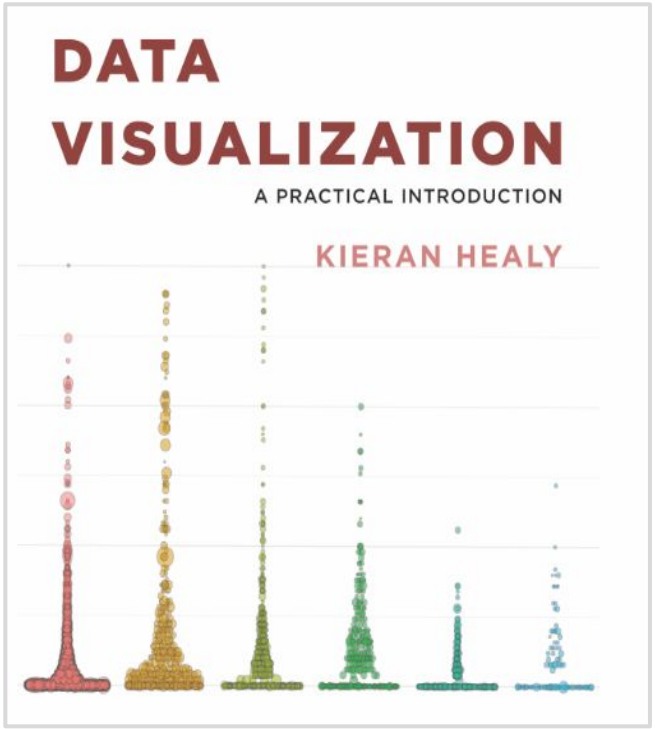


https://graphicsprinciples.github.io/

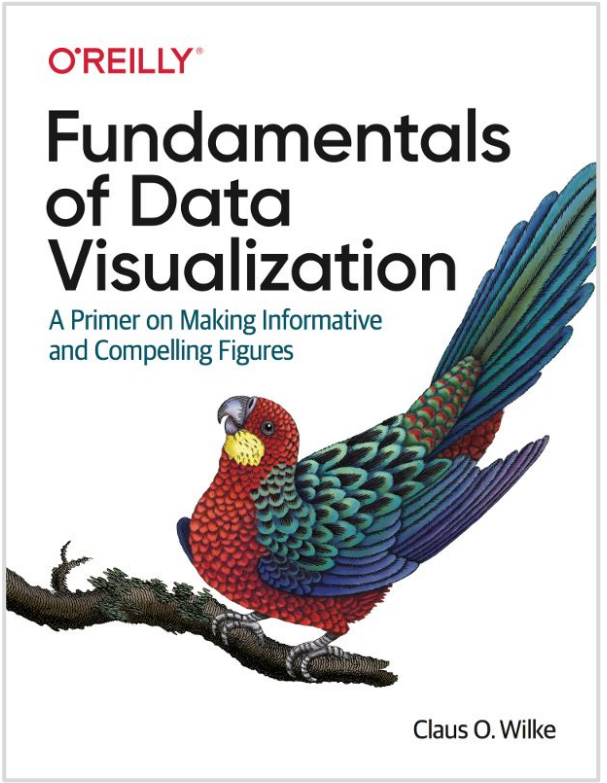
Where to find to out more?



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



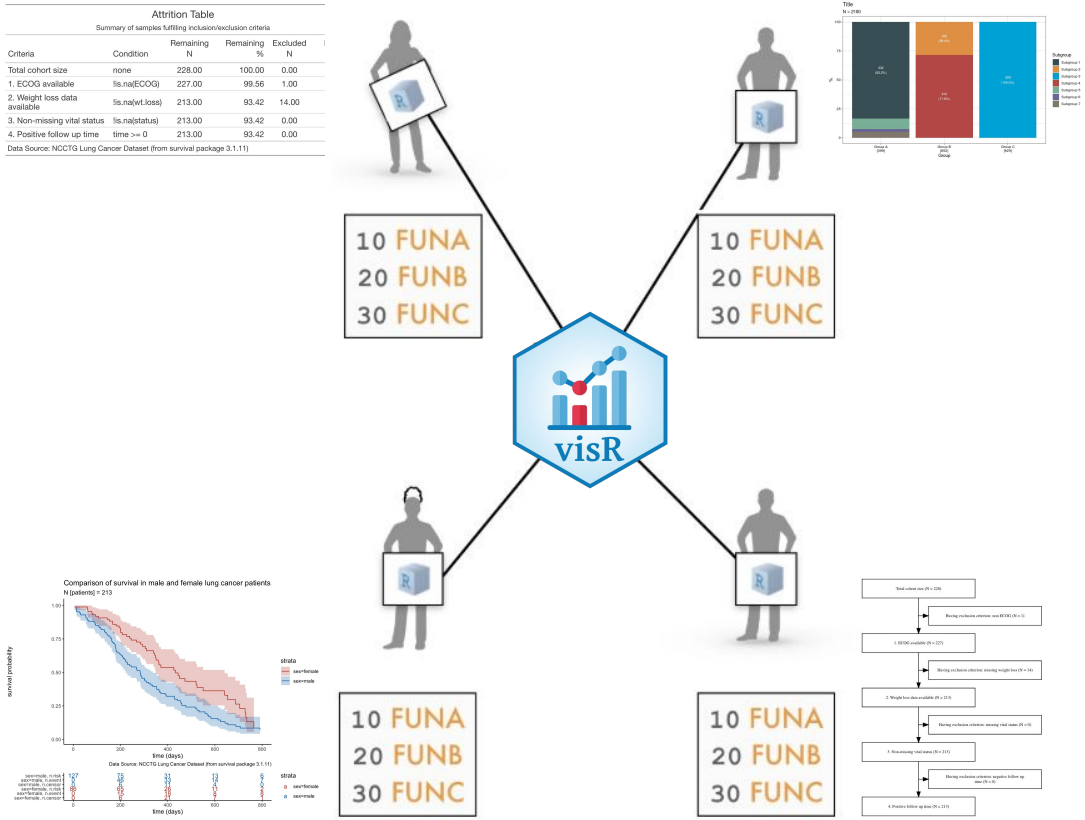
https://socviz.co/



https://serialmentor.com/dataviz/

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



Basic Architecture

- Distinct tasks implemented as independent functions:
  - **Estimator functions:** computes estimates, as well as upper and lower bounds, p-values, etc.
  - **Visualization functions:** visualizes data as a plot or a table (or something else).
  - **Style functions**: applies common theme and color palettes to all output
- Function arguments follow same vocabulary
  - CDISC

Input data model (a data.frame / tibble)

trt	time	status

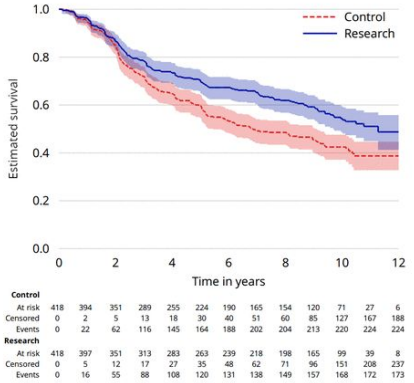
Estimator Function

Interim (estimate) data model

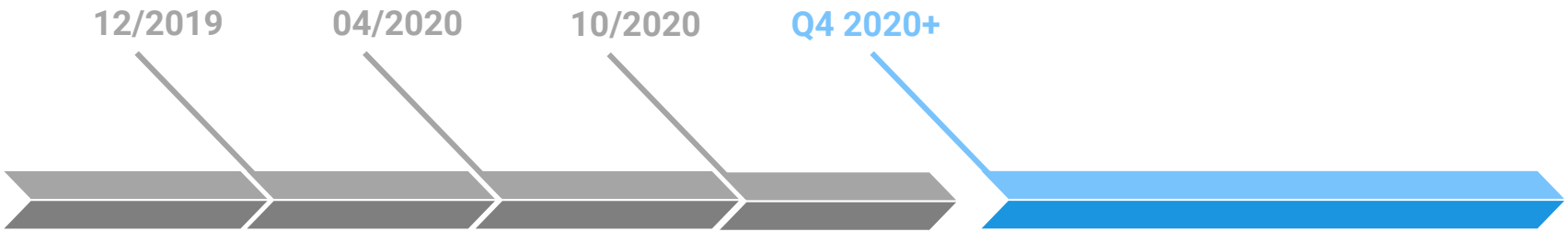
time	trt	estimate	lower	upper

Visualization Function + Style Function

Visualization



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

Official Launch

- Present prototype at RStudio webinar
- 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