

SAIL.R – Journey Towards R Software in Johnson & Johnson A Training Perspective

Paper PD08

Phuse Connect EU 2024, Strasbourg, France

10 - 13 November 2024

Yannick Vandendijck, Annemie Van Dijck, Daniel Hofstaedter, Katherine Wilde, Mia Qi & Rawan Rupnow

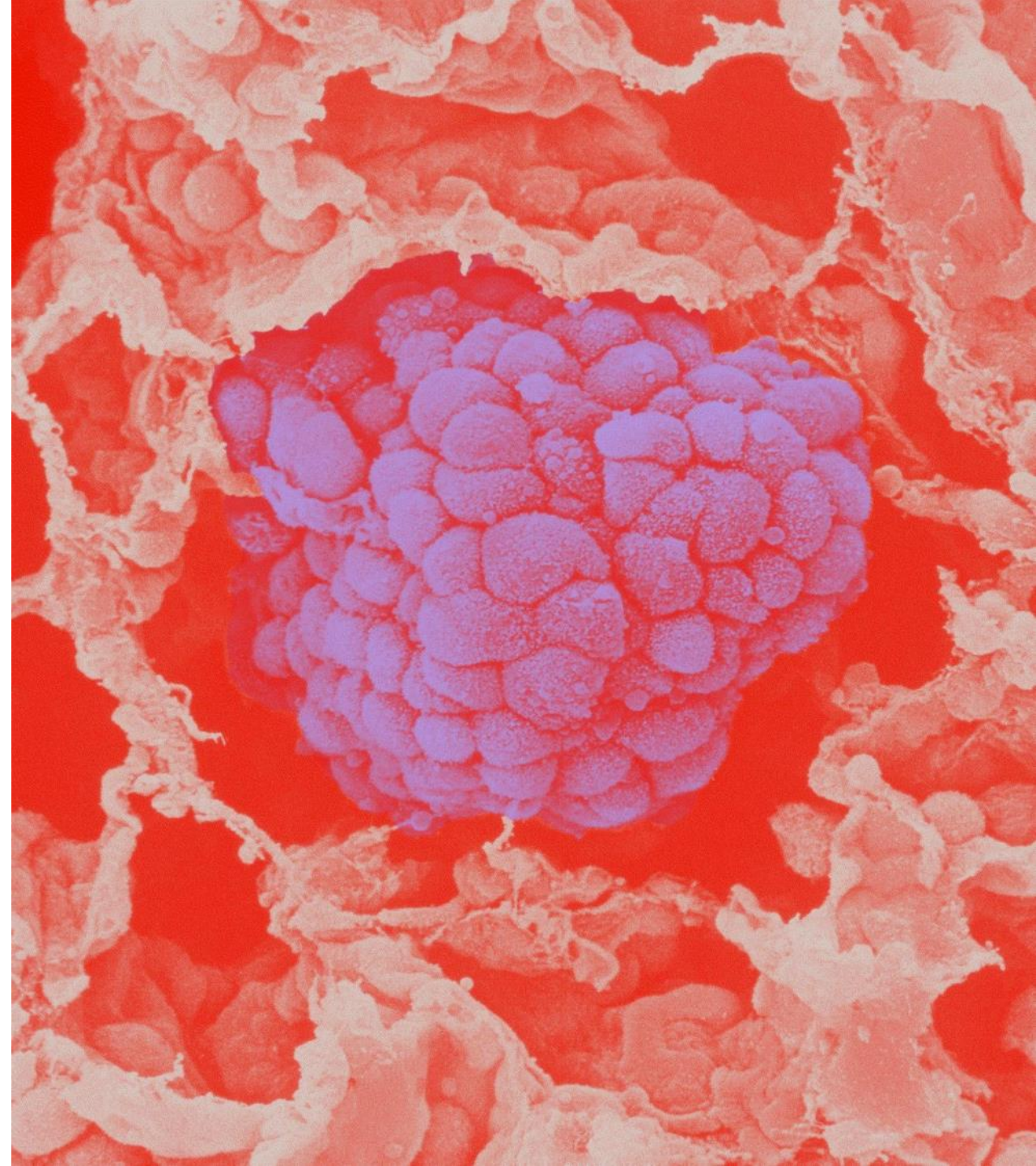
Principal Scientist Biostatistics

Johnson & Johnson, Belgium

The views and opinions expressed in this presentations are those of the speakers and do not necessarily reflect the views or positions of any entities they represent

Johnson&Johnson
Innovative Medicine

SAIL.R



Why R in our Toolbox?

- Partner with other pharma companies, collaborate in the open-source spirit
- Faster development of packages and functions of new statistical methodologies
- Wide range of packages/libraries
- Large online community (increases problem-solving speeds)
- Hiring opportunities: R is used in universities
- Impressive visualization and interactive opportunities
- Seamless integration with other data analysis software



Vision & Objectives



Vision

R-Based Submissions (TLFs) by 2025

**Use of R for interactive analysis,
visualization and exploration as a
business as usual**



Primary Goal

**Fully implementing R eco
system in developing
static/interactive analysis
(Tables, Listing and
Figures)
for data exploration and
production output delivery**



Secondary Goal

**Generating ADaM/analysis
datasets using R**

Regulatory Acceptance

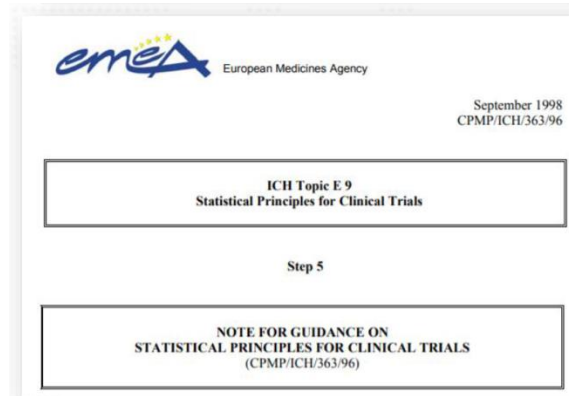
- FDA, EMEA, Japan PMDA: there is **no requirement** for a specific statistical software
- China NMPA: can accept R code but expects corresponding SAS code as well. A cross-pharma workgroup has been set up to prepare NMPA for R based submissions.



Statistical Software Clarifying Statement

FDA does not require use of any specific software for statistical analyses, and statistical software is not explicitly discussed in Title 21 of the Code of Federal Regulations [e.g., in 21CFR part 11]. However, the software package(s) used for statistical analyses should be fully documented in the submission, including version and build identification.

As noted in the FDA guidance, *E9 Statistical Principles for Clinical Trials* (available at <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm>), “The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available.” Sponsors are encouraged to consult with FDA review teams and especially with FDA statisticians regarding the choice and suitability of statistical software packages at an early stage in the product development process.



“The computer software used for data management and statistical analysis should be reliable[...].”

5.8 Integrity of Data and Computer Software Validity

The credibility of the numerical results of the analysis depends on the quality and validity of the methods and software (both internally and externally written) used both for data management (data entry, storage, verification, correction and retrieval) and also for processing the data statistically. Data management activities should therefore be based on thorough and effective standard operating procedures. The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available.

https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf

SAIL.R Initiative

Submission: Smooth and successful submission of R developed programs and outputs



Reporting Framework and Standards: Define and establish core standards used for TLGs



Change Managment: drive the adoption and implementation of R within the organization
➤ R Training



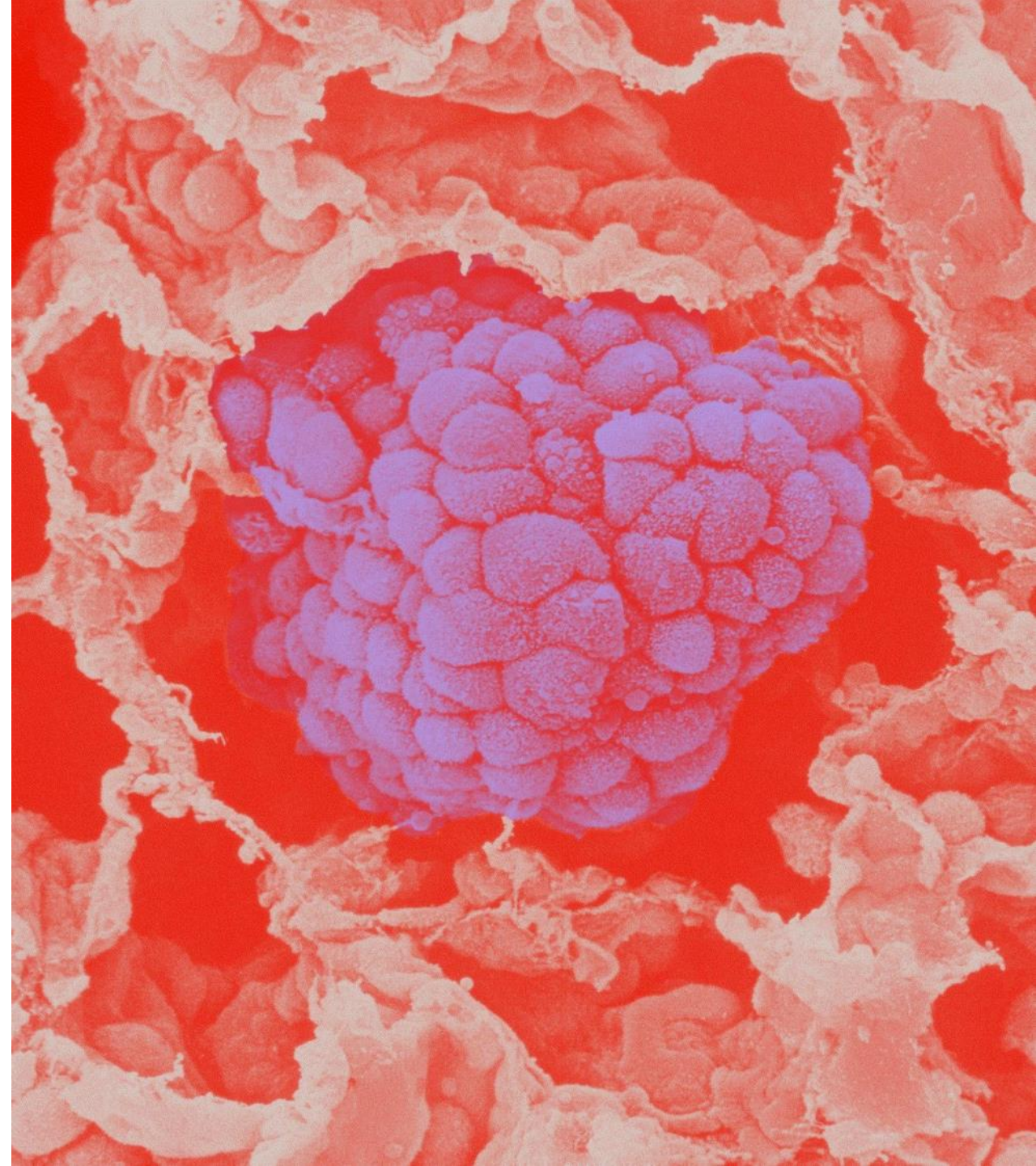
Interactive Visualization and Reports: Optimizing, improving and developing applications in R for data visualization , exploration and static TLGs



SAIL.R

R Training

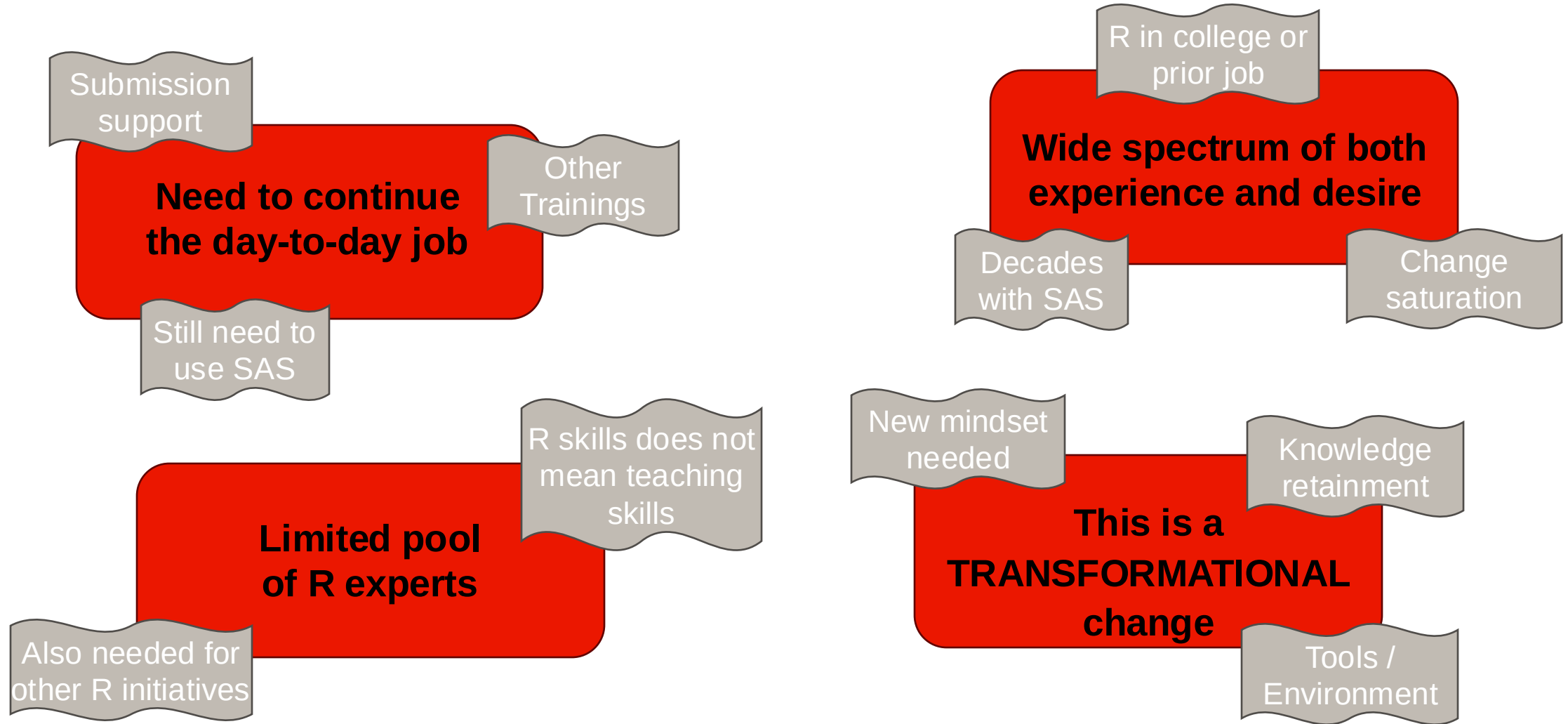
for statisticians



Objective

Training on the effective use of R for exploring, summarizing and visualizing clinical data

Challenges



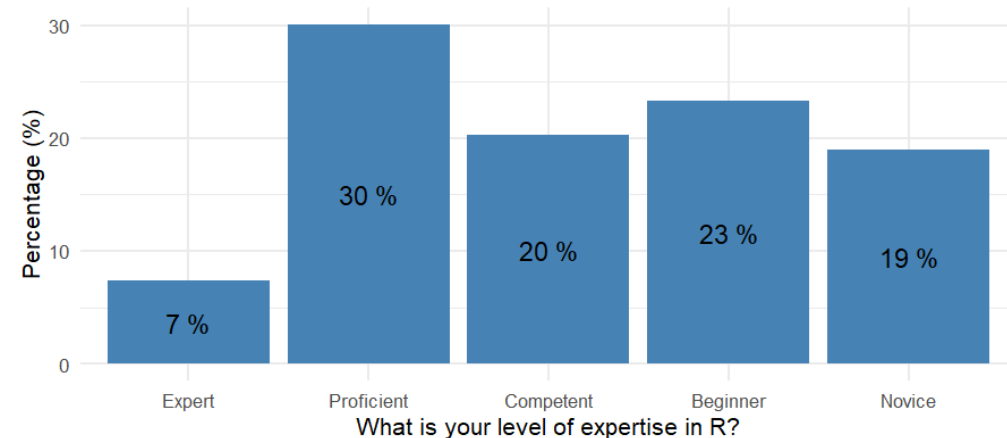
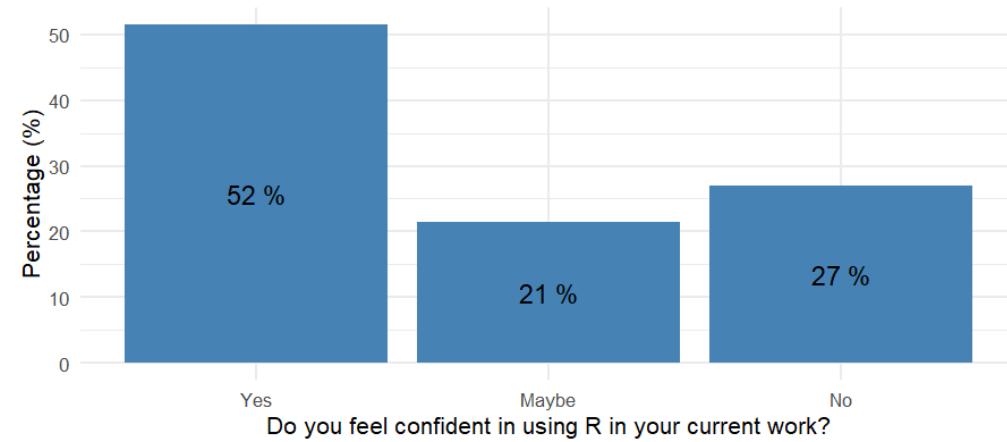
Strategy & Considerations



Audience:

- What skills do they need to be trained on?
- What skills do they already have?
- What are the competing priorities for the audience?
- ...

- Increasing interest in using R in recent years
- Target: statisticians working with clinical data
- Large cohort needed training (± 220)
- Very heterogeneous R proficiency
- Difficult to set aside time for training



Strategy & Considerations



Audience:

- What skills do they need to be trained on?
- What skills do they already have?
- What are the competing priorities for the audience?
- ...



Trainers:

- Do you have a pool of statisticians that are experienced in using R?
- Are these also experienced trainers?
- Will the trainers develop the training materials?
- What are the competing priorities of the trainers?
- ...

More *junior* colleagues with a lot of R expertise

- Develop training materials and provide the trainings
- Help them to become an effective teacher
- Same competing priorities as the audience
- Across therapeutic areas and groups (ED – LD) to develop network of R experts

Strategy & Considerations



Audience:

- What skills do they need to be trained on?
- What skills do they already have?
- What are the competing priorities for the audience?
- ...



Trainers:

- Do you have a pool of statisticians that are experienced in using R?
- Are these also experienced trainers?
- Will the trainers develop the training materials?
- What are the competing priorities of the trainers?
- ...



Environment:

- Do you already have a working R Environment for training? Vendor?
- Do you want to use proprietary data for training and/or practice?
- Traditional Classroom Approach or Leveraging Modern Technology?
- ...

Curriculum

1

QC using R

- Great start for R adaption
- Building blocks for QC
- Requires no prior R knowledge
- 8 lessons
- Participants can immediately start applying on their projects

2

Training on more automation & interactive exploration

- Innovative R tools useful for clinical statisticians
- Requires basic R knowledge

Curriculum

1

QC using R

[1] RStudio

- Setup environment for R
- RStudio interface
- Navigating in Rstudio

[2] R Programming Basics

- Intro to R objects
 - Vectors, factors, data frames,...
- Descriptive statistics
- Loading data
- R packages

[3] Data Manipulation 1

- *dplyr* package
 - Filter, select, rename, summarize, group by,...
- Joining data

[4] Data Manipulation 2

- Pivoting data
- Dates and times
- Compare datasets
- Tidytlg package

Curriculum

1

QC using R

[5] Visualization 1

- Introduction to ggplot2
- Important geom functions
- Building a good-looking figure
- Changing colors, axis, labels
- Multiple panels

[7] Modelling 1

- Binary endpoints
- Continuous endpoints

[6] Visualization 2

- More examples of ggplot2
 - Bar graph, boxplot, Line plot, Forest plots
- Tidytgl package

[8] Modelling 2

- *mmrm* package
- Survival analysis

Curriculum

2

Training on more automation & interactive exploration

[9] R Markdown and interactive plots

- Introduction of R Markdown
- Interactive plots using *plotly*

[10] TLR and Statistical report in R

- In-house innovative R tools
 - Top Line Results report
 - Statistical Report

Implementation

- Traditional “classroom” approach leveraging modern technology
- Multiple classes per topic organized
- Topics were provided sequentially
- Total duration of training was $\pm$ 6 months

Class duration



Class size



Programming environment:

- Consistent version of R & RStudio
- Consistent set of packages
- All examples on SDTM and ADaM data (clinical test data from CDISC)

Materials

- Quick wins for the audience
- Relatable to the job
- Audience actively engaged
- Each training followed a consistent and standardized structure

3 R vs SAS

SAS	R
<pre>* proc sort data=adsl; by trt01p; run;</pre>	<pre>adsl_new <- adsl %>% arrange(TRT01P)</pre>
<pre>* data adsl_new; set adsl; rename USUBJID = SUBJ SAFFL = TRTFL; run;</pre>	<pre>adsl_new <- adsl %>% rename(SUBJ = USUBJID, TRTFL = SAFFL)</pre>

1 Slides

The `table()` function provides counts for character or factor variables. Create a factor variable to specify the ordering of the output.

Note the difference in ordering between the factor and the character

```
> table(adsl$TRT01P) #TRT01P was turned into a factor with specified ordering.  
Placebo  Kanesline low dose Kanesline high dose  
      84          84          84  
> table(adsl$TRT01P) #TRT01P is still a character vector.  
Placebo Kanesline high dose Kanesline low dose  
      84          84          84
```

To get proportions, we can use the `prop.table()` function on the table created with the `table()` function. Specify row proportions or column proportions for 2x2 tables via the `margin` argument in the `prop.table()` function.

```
#counts for categorical variables  
tbl1 <- table(adsl$TRT01P)  
tbl2 <- table(adsl$DOSECD, adsl$TRT01P)  
  
#get proportions for TRT01P  
prop.tbl1 <- prop.table(tbl1)  
  
#margin=2 gives proportions by columns for 2x2 table  
prop.tbl2 <- prop.table(tbl2, margin=2)
```

2 Demo code

```
168 #Calculate the average height, weight and BMI per group for the IIT population  
169 tbl <- adsl %>%  
170   filter(ITTFL == "Y") %>%  
171   select(USUBJID, TRT01P, HEIGHTBL, WEIGHTBL) %>%  
172   mutate(derived_bmi = WEIGHTBL / (HEIGHTBL / 100) ^ 2) %>%  
173   group_by(TRT01P) %>%  
174   summarise(  
175     AvgHeight = mean(HEIGHTBL, na.rm = TRUE),  
176     AvgWeight = mean(WEIGHTBL, na.rm = TRUE),  
177     AvgBMI    = mean(derived_bmi, na.rm = TRUE)  
178   )  
179
```

4 Class Polls

Which function would I use to create a new variable in my data frame?

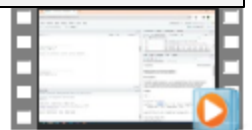
select
arrange
mutate
rename

5 Exercises

Exercise – Box Plots

1. Using the `adsl` dataset: Use `ggplot2` to map `TRT01A` (actual treatment for period1) to the x-axis and `DURDIS` (duration of disease by months) to the y-axis.
2. Create a boxplot. Set the outliers to show in color blue and shape to 1.
3. Define appropriate labels for your title, the x and y axes.
4. Change the theme.

6 Videos



✔ Lesson2_R
Programming
Basics.mp4

Example 1

Data Manipulations 1

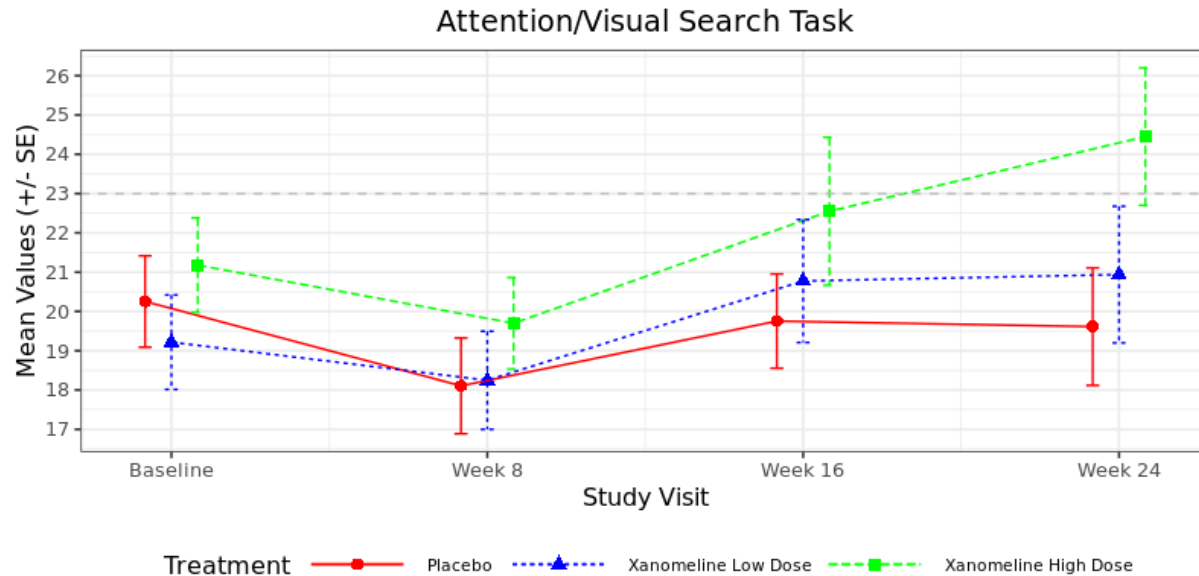
- *dplyr* package
- Functions such as `filter()`, `select()`, `mutate()`, `rename()`, `summarize()`, `group by()`, etc. are introduced separately
- Finally, all functions are combined to show an example in which:
 - `adsl` dataset is filtered on ITT patients
 - BMI variable is calculated
 - Summary statistics are calculated by treatment group

Example 1

```
tbl = adsl %>%  
  filter(ITTFL == "Y") %>%  
  select(USUBJID, TRT01A, HEIGHTBL, WEIGHTBL) %>%  
  mutate(BMIn = WEIGHTBL / (HEIGHTBL / 100) ^ 2) %>%  
  group_by(TRT01A) %>%  
  summarise(  
    n = n(),  
    nHeight = sum(!is.na(HEIGHTBL)),  
    AvgHeight = mean(HEIGHTBL, na.rm = TRUE),  
    nWeight = sum(!is.na(WEIGHTBL)),  
    AvgWeight = mean(WEIGHTBL, na.rm = TRUE),  
    nBMI = sum(!is.na(BMIn)),  
    AvgBMI = mean(BMIn, na.rm = TRUE)  
  )
```

Filter ITT population
This step is not required
Calculate BMI variable
Summary stats by treatment
Calculate summary stats

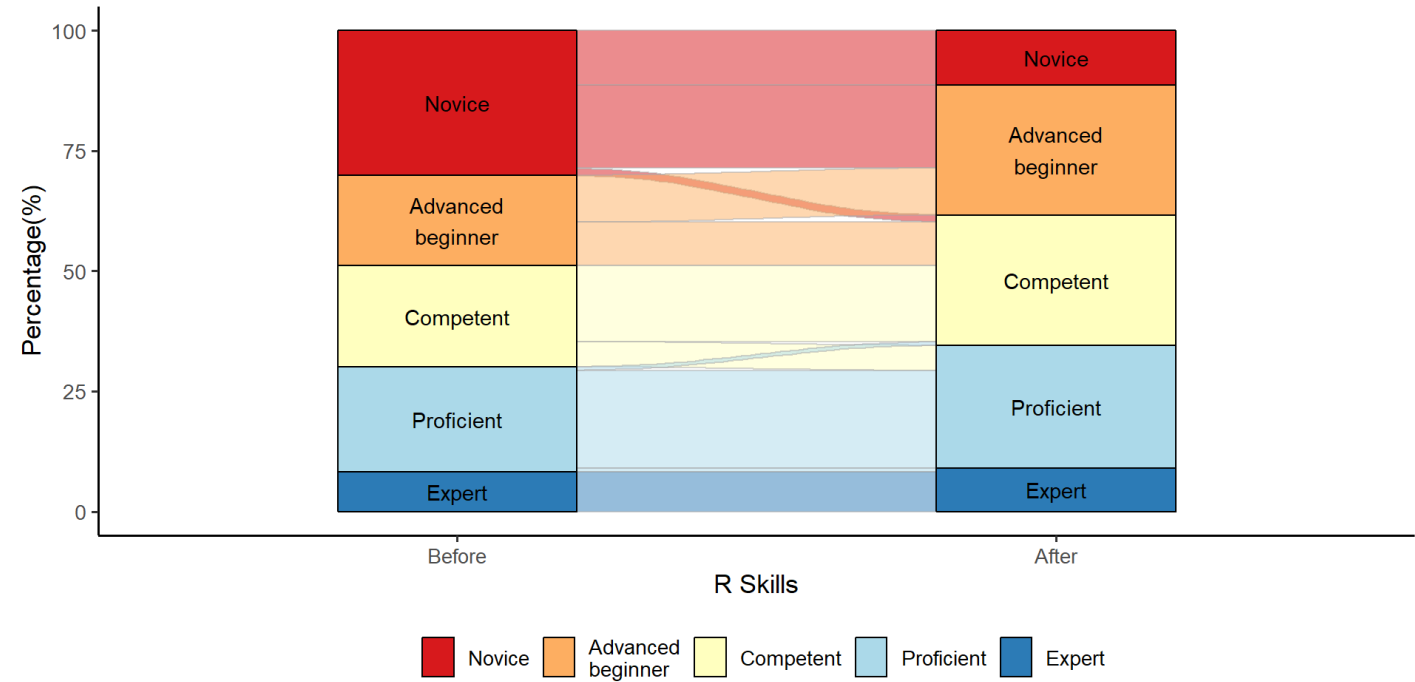
Example 2



- ```
ggplot(data = summ_data,
 aes(x = AVISITN, y = mean, group = TRTPf)) +
 geom_point(size = 2.5) +
 geom_line()
```
- ```
ggplot(data = summ_data,  
       aes(x = AVISITN, y = mean,  
           col = TRTPf, linetype = TRTPf,  
           shape = TRTPf)) +  
  geom_point(size = 2.5) +  
  geom_line()
```
- ```
ggplot(data = summ_data,
 aes(x = AVISITN, y = mean,
 ymin = mean - se, ymax = mean + se,
 col = TRTPf, linetype = TRTPf,
 shape = TRTPf)) +
 geom_point(size = 2.5) +
 geom_line() +
 geom_errorbar(width = 1)
```

# Evaluation

- 40+ classes
- Approx 180 individuals trained
- Global coverage including US, EU, AP



## Comments:

“I used what I learned on real data from my study”

“The trainers were experts in R. It was a big plus for the training.”

“This series of trainings were very well organized and informative”

“I found the live version of the training not to be very helpful”

“More time between topics to have more time to digest and practice by our self”

“More examples”

# Training Support & Reinforcement

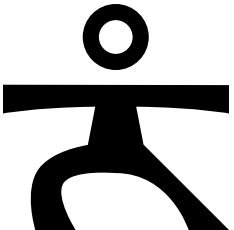


## QC of Tables, Listings and Figures

- All clinical statisticians encouraged to utilize R for basic QC on their trial work
- Network of experts per disease area

## GenAI

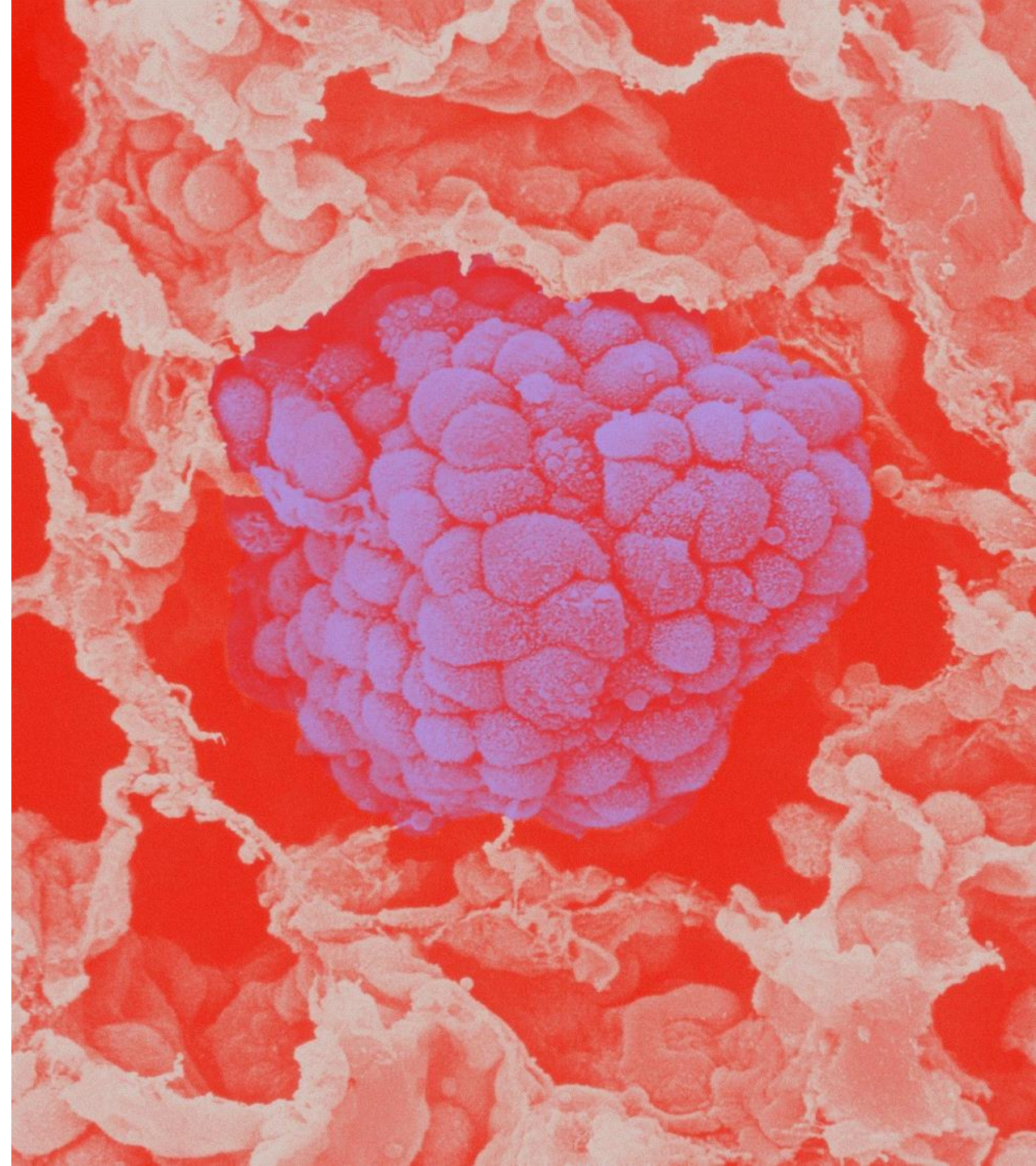
- Separate training provided on how genAI can help in R coding
- Shows how to enhance R learning
- Warning: also show pitfalls of GenAI



## SAIL.R Trips

- Recurring meetings
- Devoted to a small pre-specified training topic.
- Open forum, driven by the community.

# Discussion



# Discussion

## ➤ **Statisticians vs Statistical Programmers**

- R proficiency levels are different (<20% confident in using R)
- R Training on building blocks comparable (Rstudio, Basics, Data Manipulations)
- Less/no emphasis on data modeling for programmers
- More emphasis on creating TLGs for programmers
- More emphasis on visualization and R Markdown for statisticians

## ➤ **eLearning Approach**

- Not have in-house R trainers or training expertise
- No time/resources to build and conduct the training
- Supplement in-house training

## ➤ **Continuous learning in R**

- Simulations, Functions, Power calculations, Swim plots, etc.

# 70-20-10 Learning model

## 70:20:10 LEARNING MODEL: EVERYTHING YOU NEED TO KNOW



# Key Takeaways



**When it comes to training, one size does not fit all**  
**Identify your goals**

- Understand your audience
- Know your resources



**Keep your trainings relatable to the job and interactive**



**Provide your audience with opportunities to reinforce the learnings**

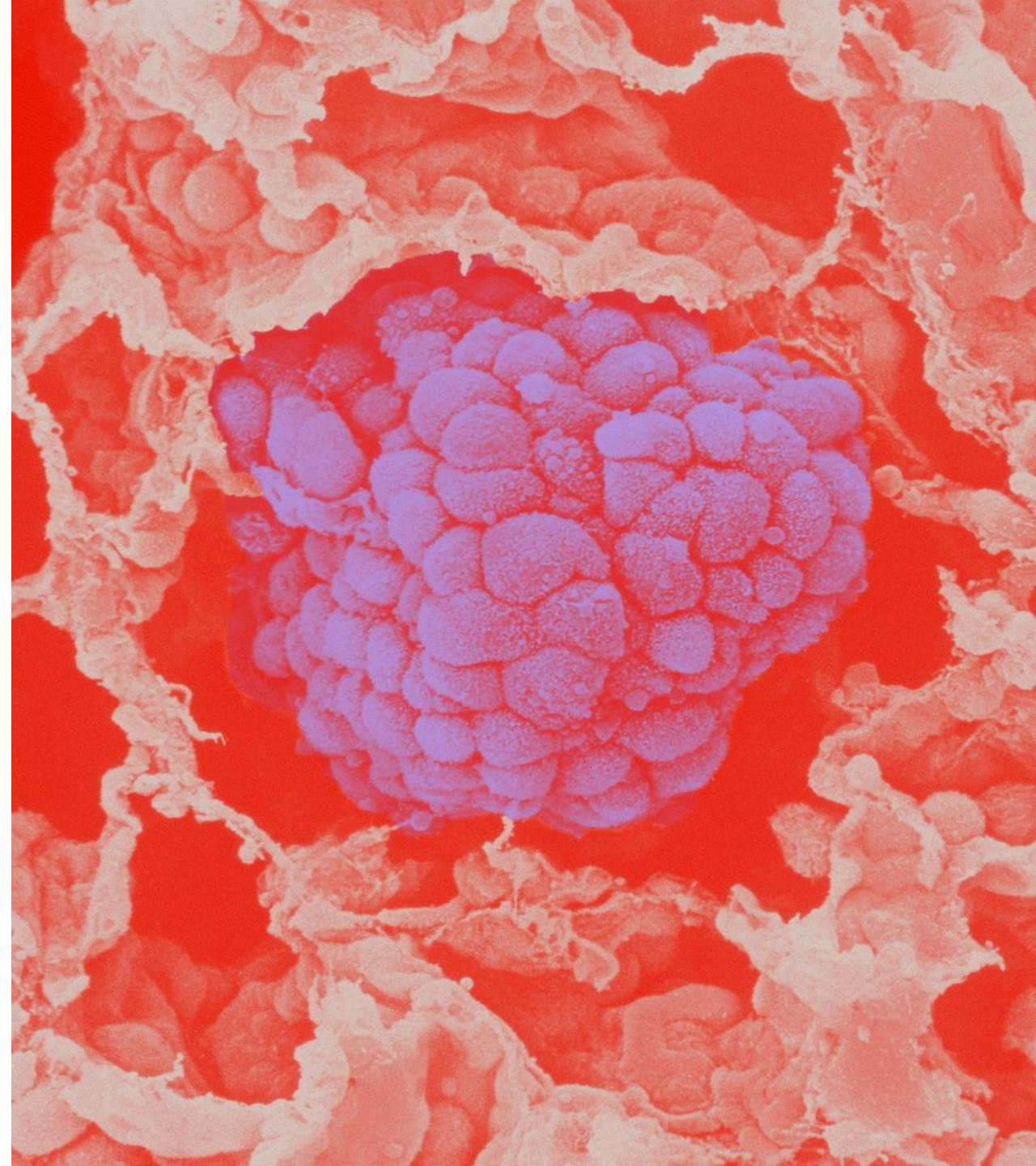


**Provide your audience with post-training support**



**Training is only one aspect to initiating the change. Consider an integrated plan**

Questions  
Comments  
Suggestions



# References

1. <https://www.r-project.org/> (2024)
2. <https://pharmaverse.org/> (2024)
3. Statistical Software Clarifying Statement. FDA (2015). Available at <https://fda.report/media/109552/Statistical-Software-Clarifying-Statement-PDF.pdf>
4. ICH harmonised tripartite guideline: statistical principles for clinical trials (E9) (1988). Available at [http://www.fda.gov/Drugs/GuidanceCompliance\\_RegulatoryInformation/Guidances/default.htm](http://www.fda.gov/Drugs/GuidanceCompliance_RegulatoryInformation/Guidances/default.htm)
5. Wickham H, François R, Henry L, Müller K (2022). dplyr: A Grammar of Data Manipulation. <https://dplyr.tidyverse.org>, <https://github.com/tidyverse/dplyr>.
6. Masel N, Haesendonckx S, Alexandra Papadopoulou P, Wang S, Miller E, Kosiba N, Ceney A (2023). tidytlg: Create TLGs using the 'tidyverse'. <https://pharmaverse.github.io/tidytlg/main/>
7. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
8. Sabanes Bove D, Dedic J, Kelkhoff D, Kunzmann K, Lang B, Li L, Stock C, Wang Y, James D, Sidi J, Leibovitz D, Sjoberg D (2023). mmrm: Mixed Models for Repeated Measures. <https://openpharma.github.io/mmrm/>.
9. <https://rmarkdown.rstudio.com/> (2024)
10. Sievert C (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC, Florida.