



# Misuse and Misapplication in Statistical Data Analysis

– A topic that never goes out of style

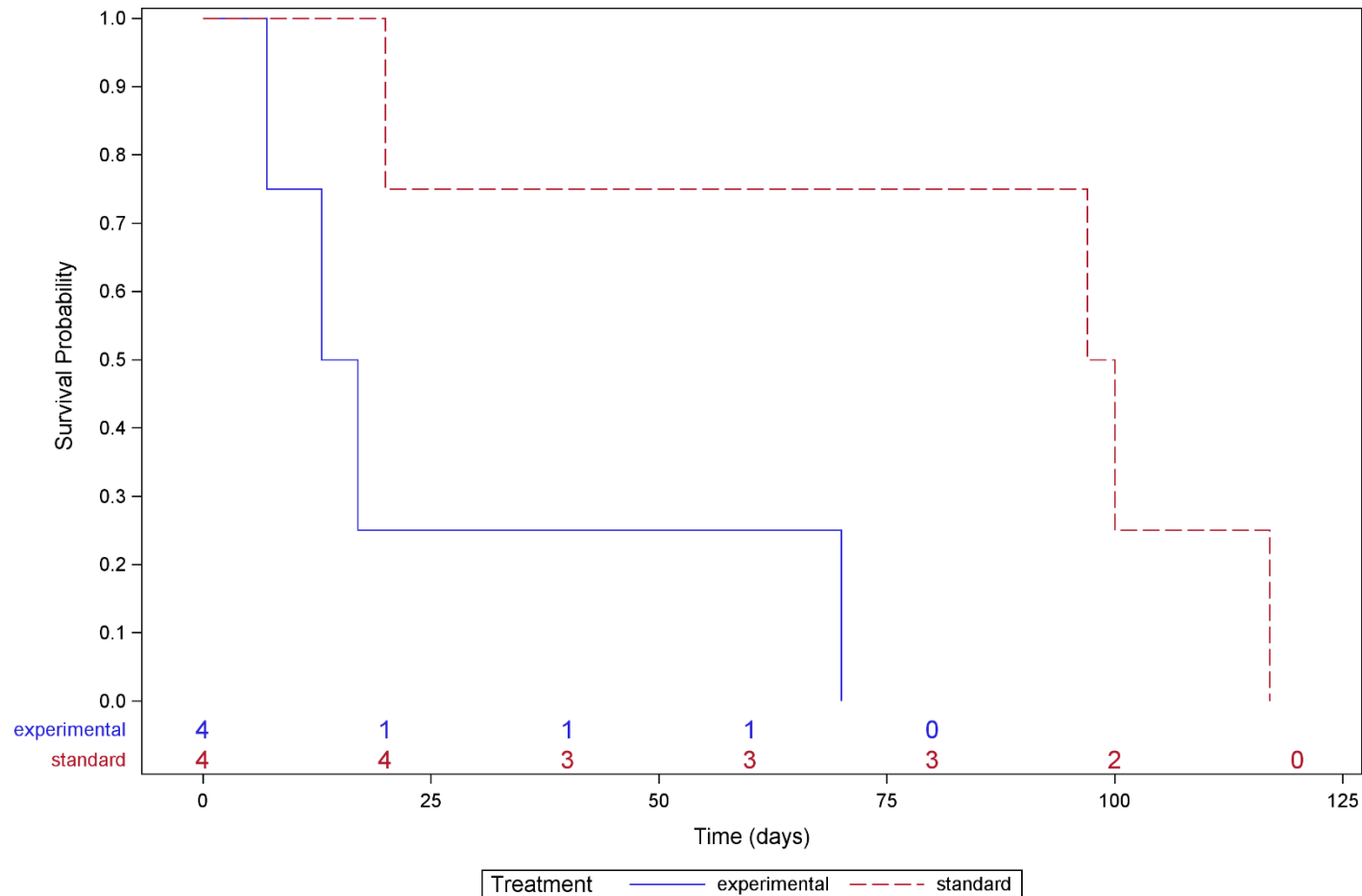
Corinna Miede

PhUSE US Connect Baltimore  
Analytics, Big Data & Statistics - Stream  
26-Feb-2019



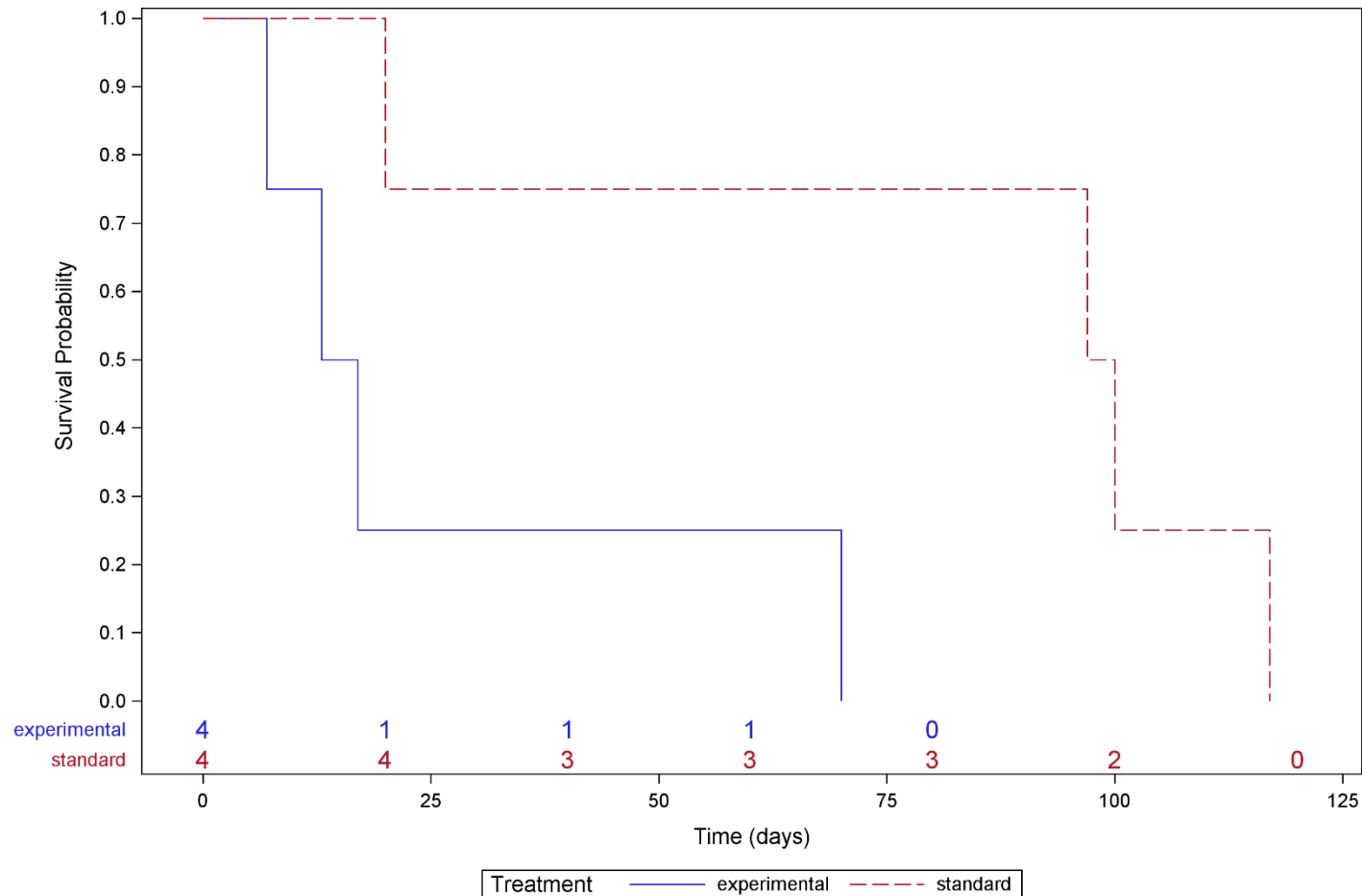
# Motivation

## Kaplan-Meier plot of overall survival



# Motivation

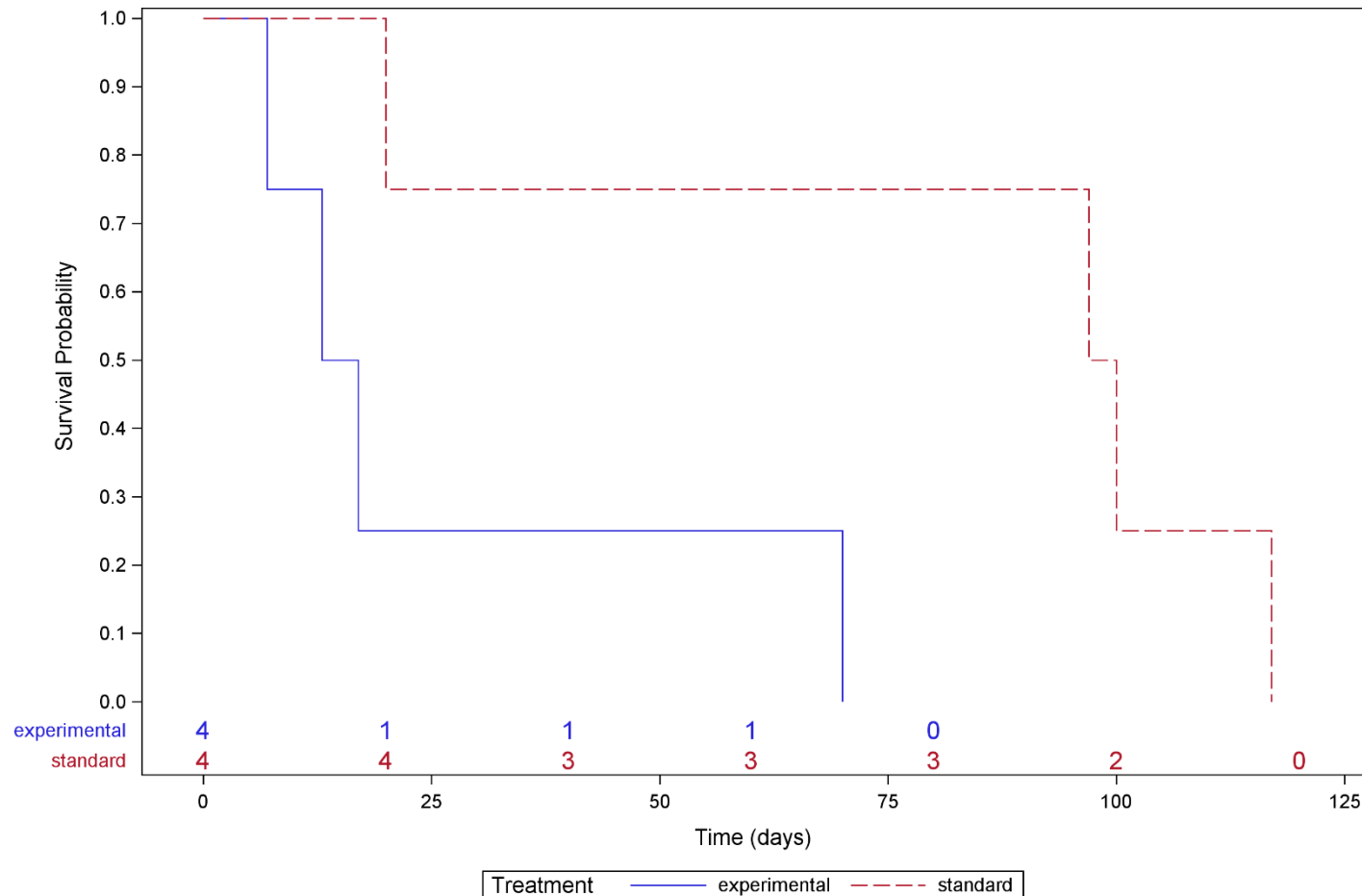
## Kaplan-Meier plot of overall survival



- All patients died?

# Motivation

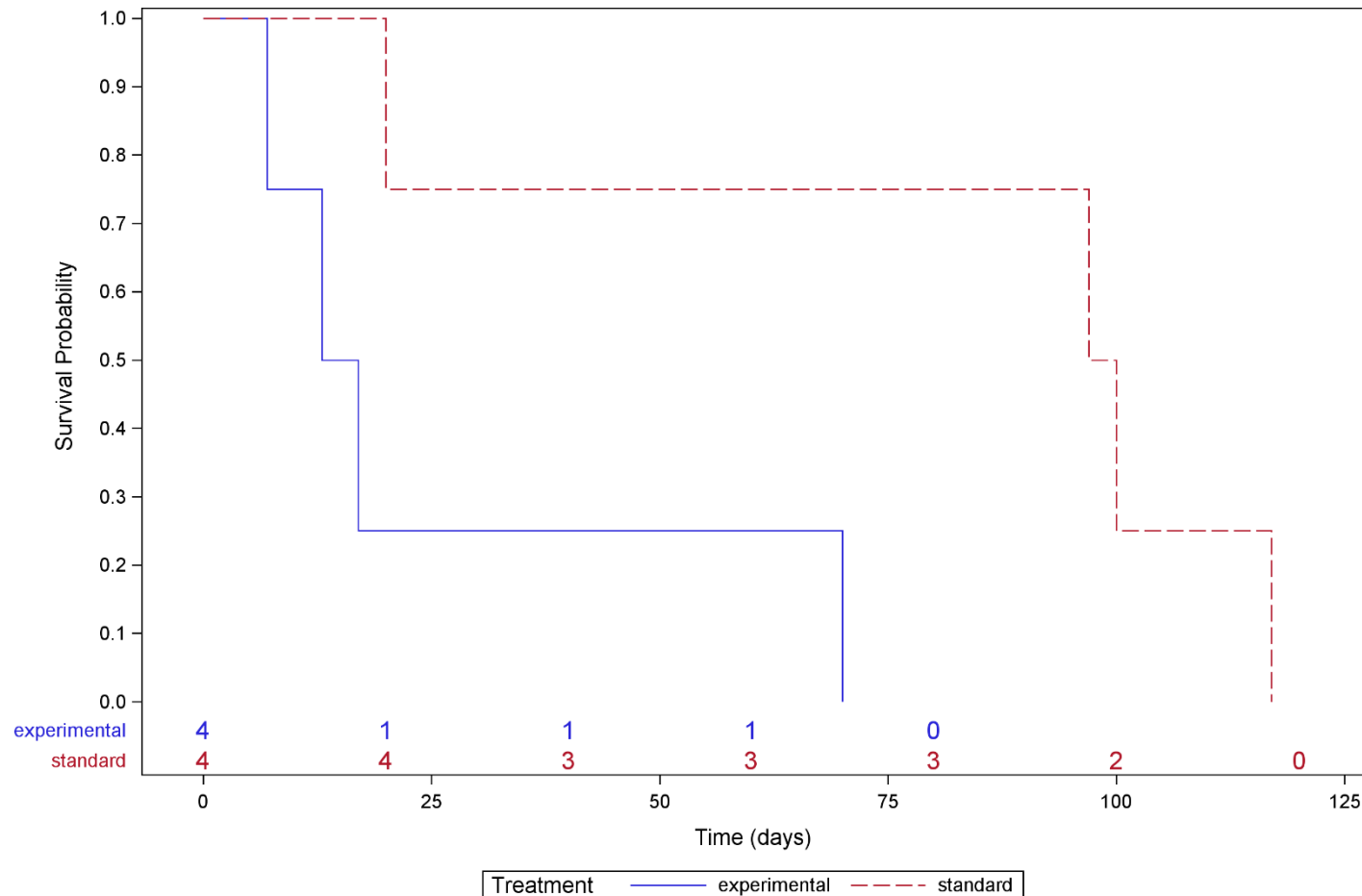
## Kaplan-Meier plot of overall survival



- All patients died?
- Only 4 patients in each treatment group?

# Motivation

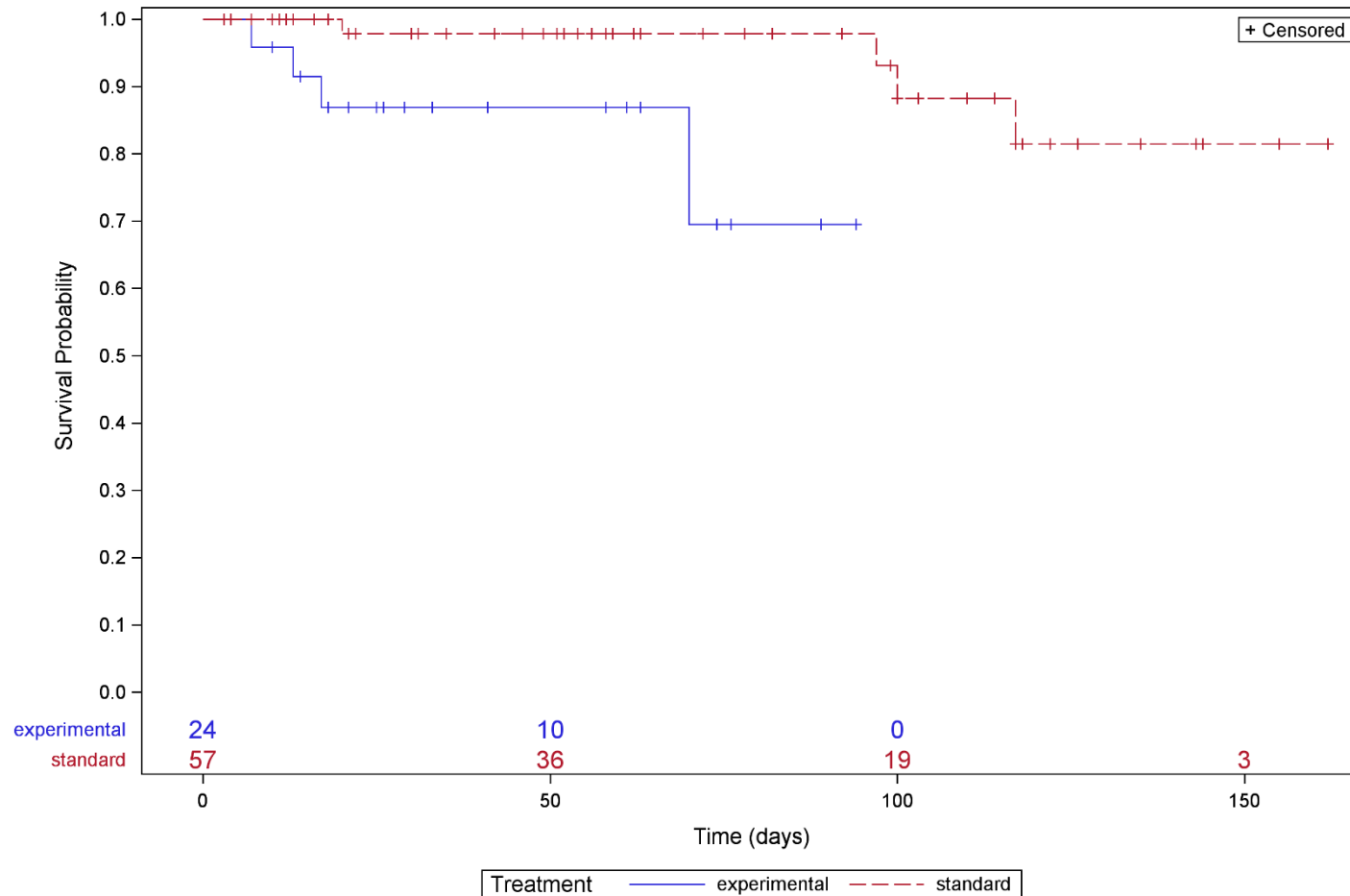
## Kaplan-Meier plot of overall survival



- All patients died?
- Only 4 patients in each treatment group?
- Only patients with events are included!

# Motivation

**Correct** Kaplan-Meier plot of overall survival



# Motivation

## Challenger disaster



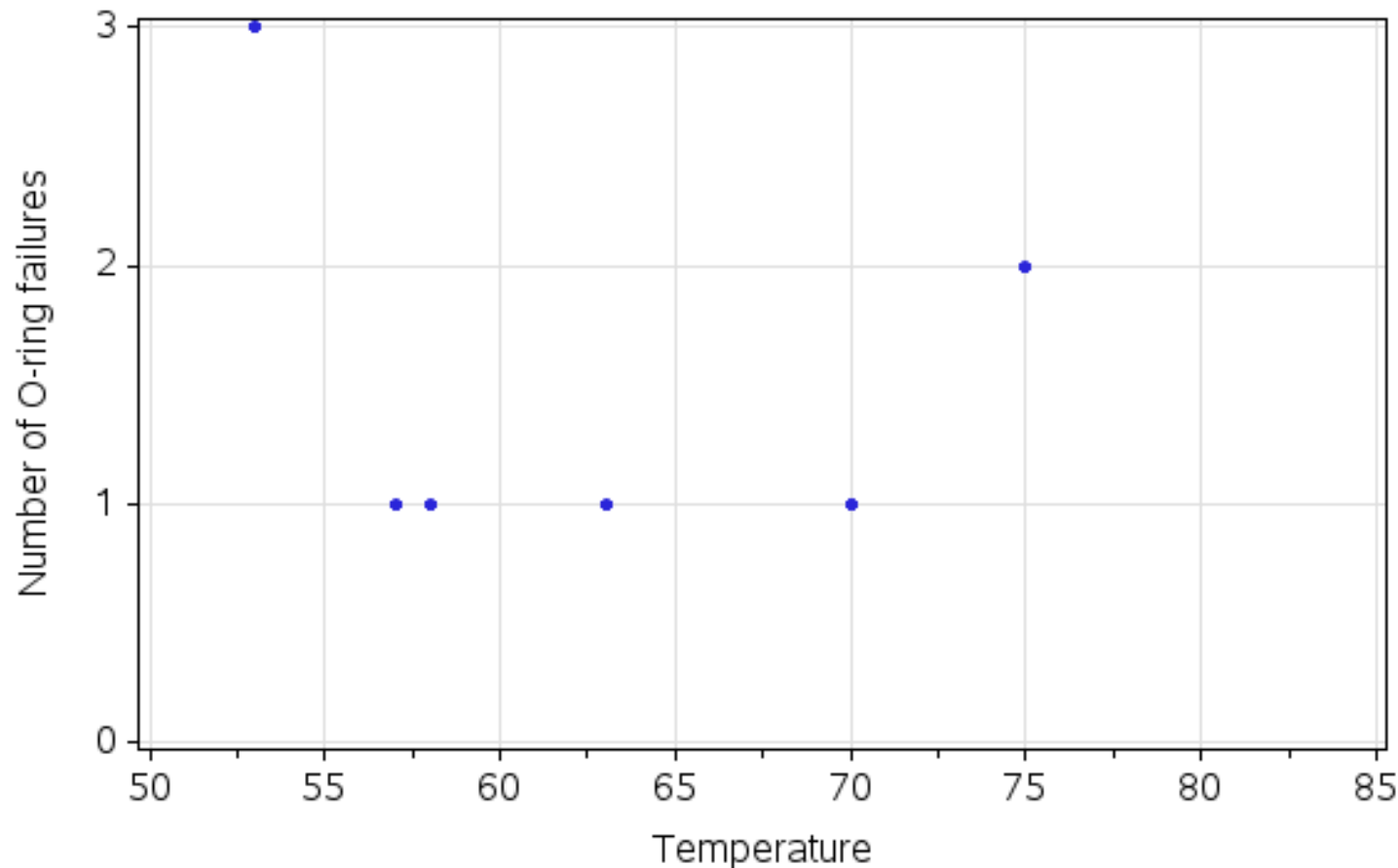
- By Kennedy Space Center - <http://grin.hq.nasa.gov/ABSTRACTS/GPN-2004-00012.html>, Public Domain, <https://commons.wikimedia.org/w/index.php?curid=530754>

# Motivation

Challenger disaster



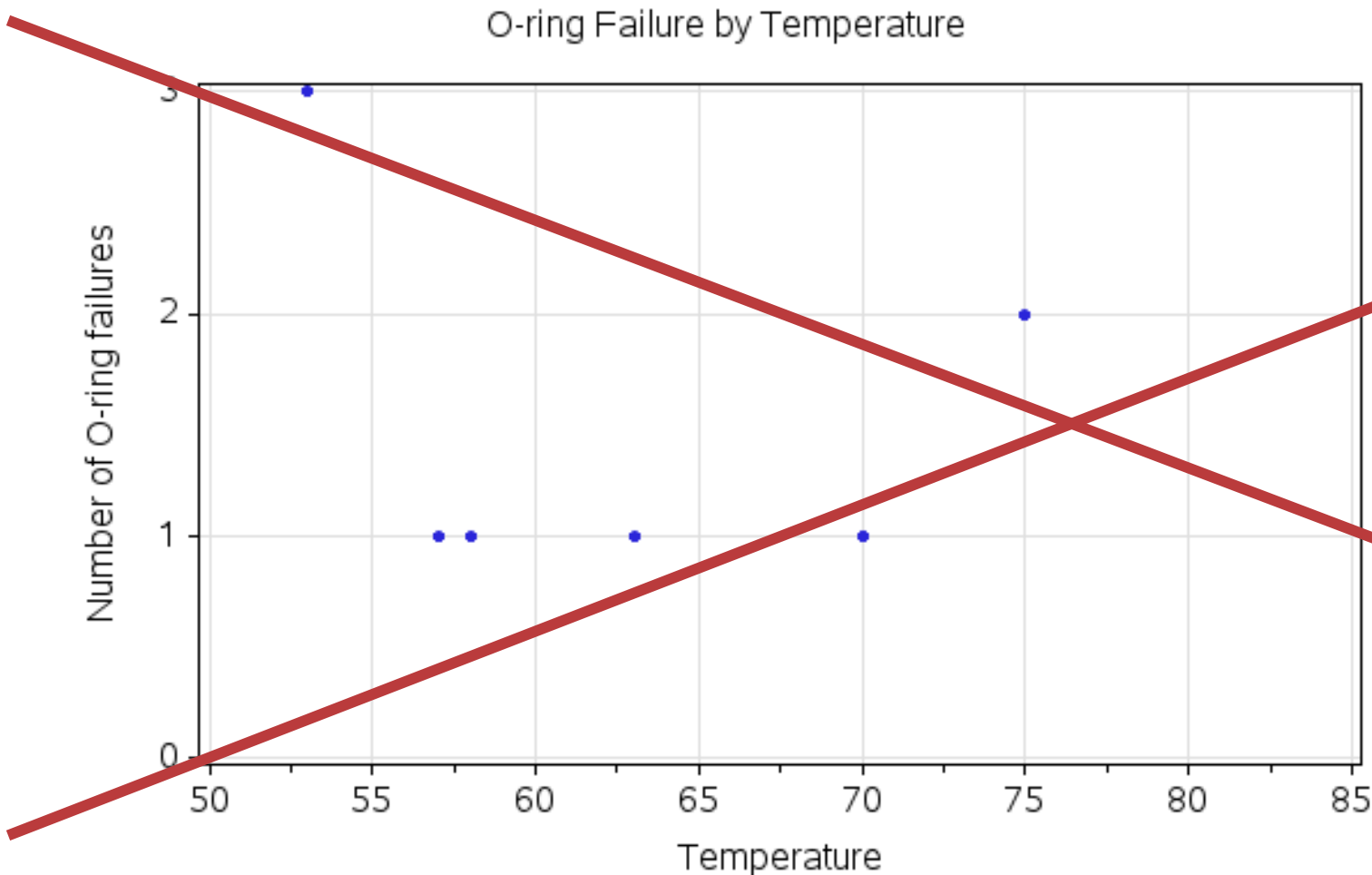
O-ring Failure by Temperature



- Data analyzed before space shuttle start
- Flights with incidents of O-rings
- No relationship between temperature and failure was detected
- Thus, decision for shuttle launch

# Motivation

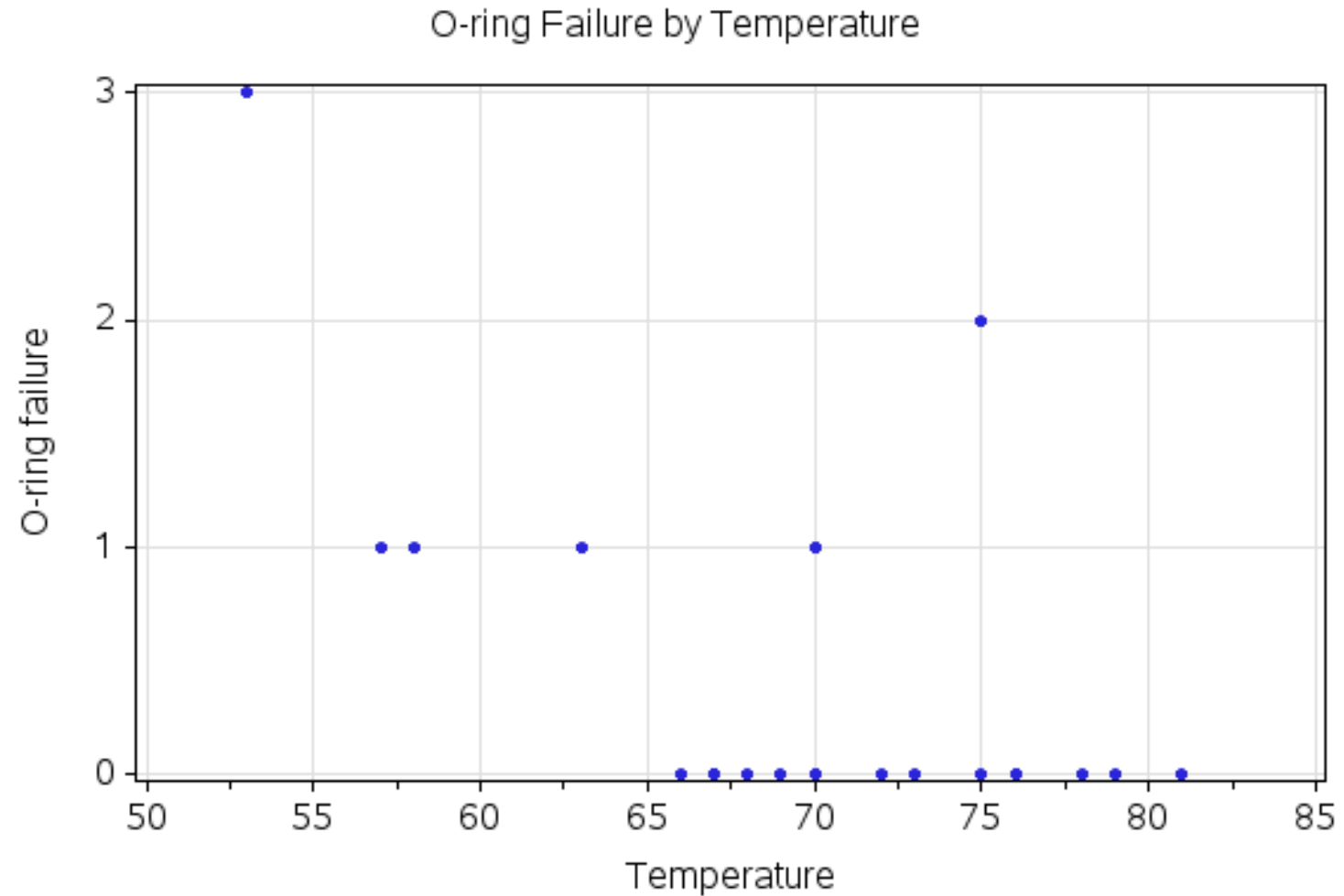
## Challenger disaster: wrong visualization



- Data analyzed before space shuttle start
- Flights with incidents of O-rings
- No relationship between temperature and failure was detected
- Thus, decision for shuttle launch

# Motivation

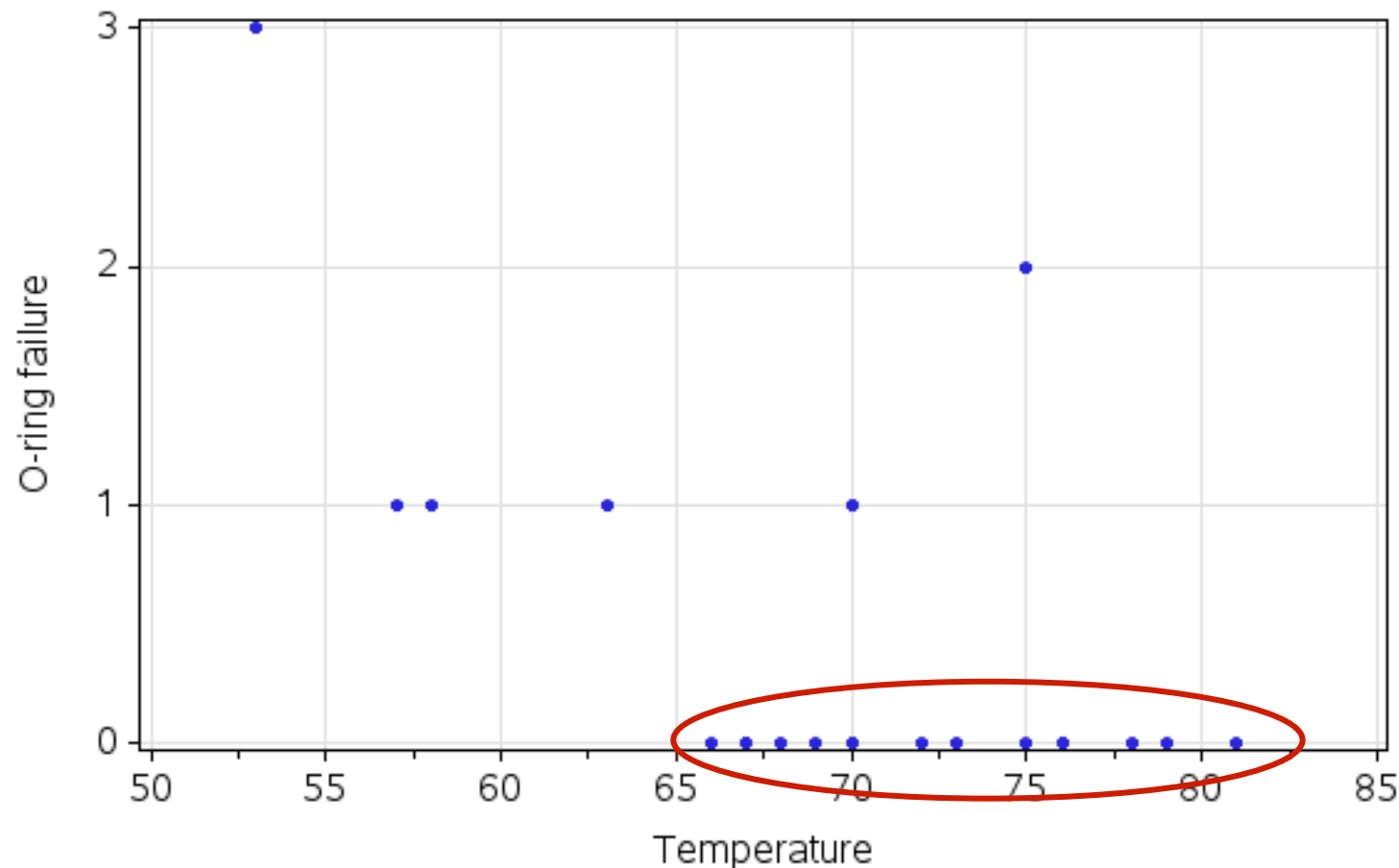
Challenger disaster: correct plot



# Motivation

## Challenger disaster

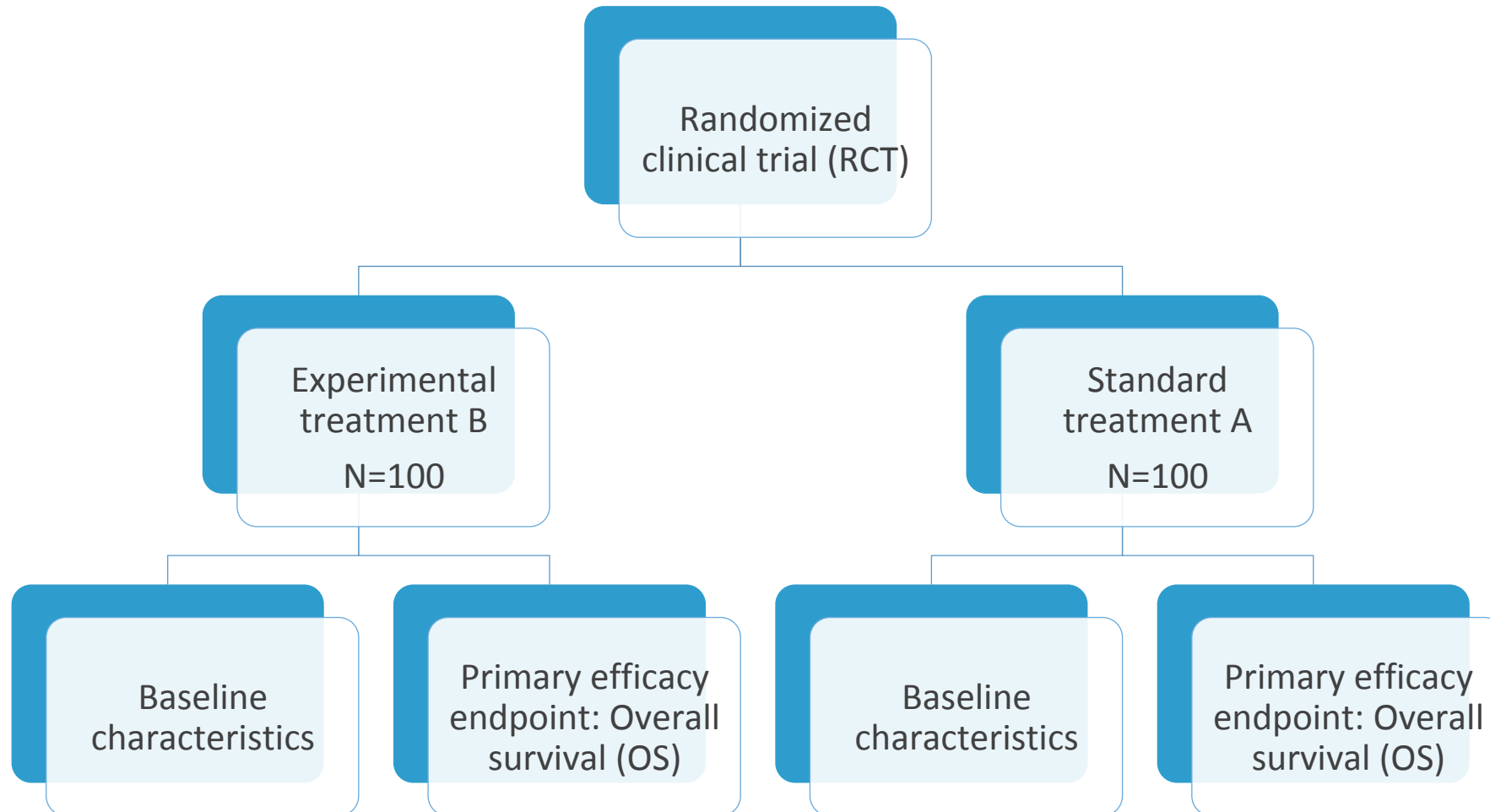
O-ring Failure by Temperature



- Include flights without any incidents of O-ring failure
- Relationship between temperature and failure is visible
- Temperature on the day of the launch was 31°F, so even colder than the minimum temperature in the plot
- They probably would have decided against shuttle launch

# Introduction

Let's introduce an Example Study



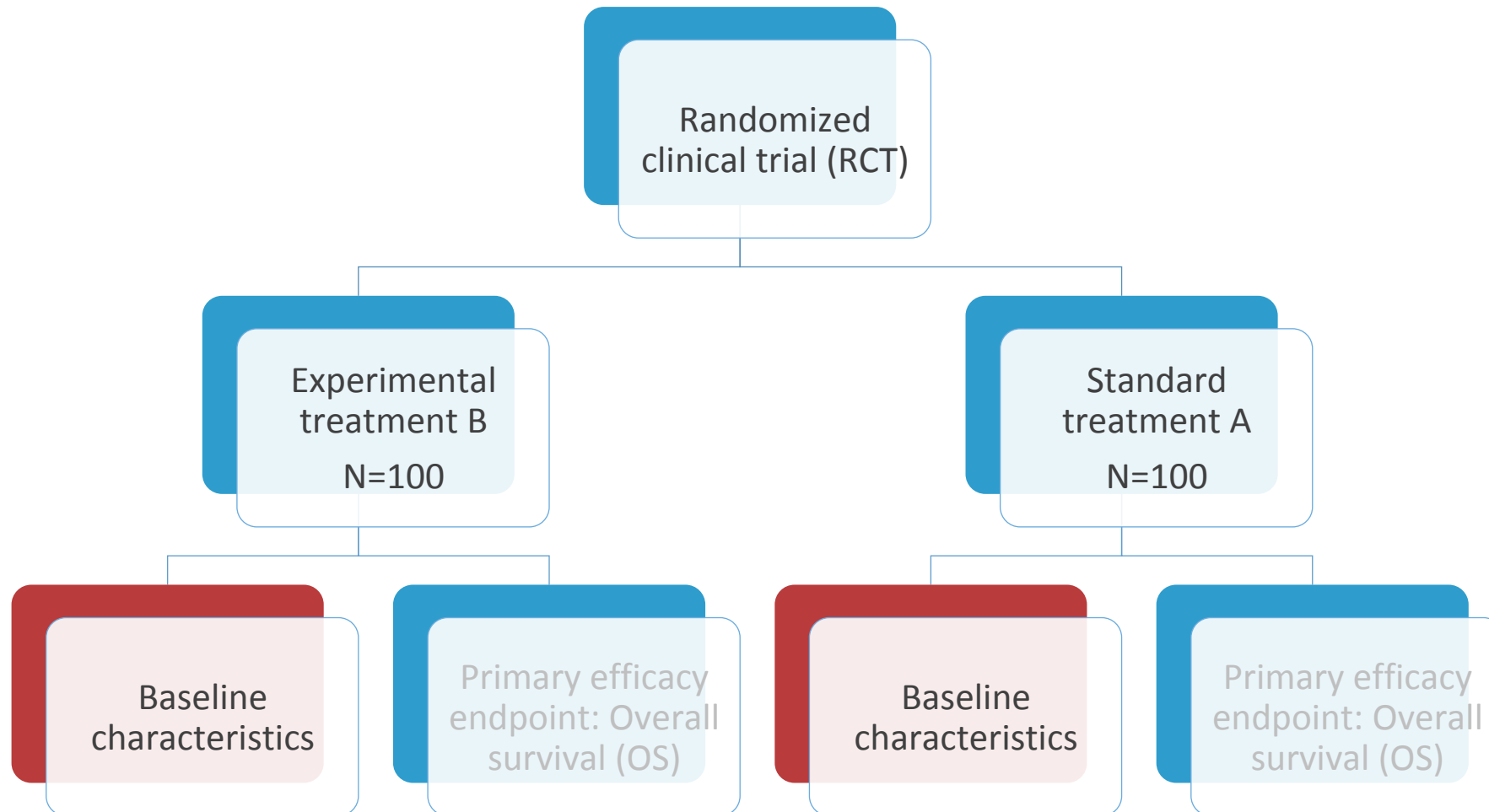
## Example Study: CTR analysis vs. Publications



- A clinical trial report (CTR) consists normally of several thousand pages with tables, listings, and figures as well as narratives and appendices
- After CTR is finalized, study results are published in medical journals
- A publication should summarize the results only on few pages
  - ➔ Goal is to save space and present result in a condensed way

# First Example

Baseline characteristics



# First Simple (Well-Known?) Example

Table of baseline characteristics



|                                | Experimental treatment B | Standard treatment A |
|--------------------------------|--------------------------|----------------------|
| Number of patients, n (%)      | 100 (100)                | 100 (100)            |
|                                |                          |                      |
| Age                            |                          |                      |
| Median (range)                 | 62.0 (32-81)             | 60.5 (28-81)         |
|                                |                          |                      |
| Male sex, n (%)                | 71 (71.0)                | 72 (72.0)            |
|                                |                          |                      |
| ...<other characteristics>     | ...                      | ...                  |
|                                |                          |                      |
| Bone marrow involvement, n (%) | 23 (23.0)                | 28 (28.0)            |

# First Simple (Well-Known?) Example

Table of baseline characteristics

|                                | Experimental treatment B | Standard treatment A |
|--------------------------------|--------------------------|----------------------|
| Number of patients, n (%)      | 100 (100)                | 100 (100)            |
| Age                            |                          |                      |
| Median (range)                 | 62.0 (32-81)             | 60.5 (28-81)         |
| Male sex, n (%)                | 71 (71.0)                | 72 (72.0)            |
| ...<other characteristics>     | ...                      | ...                  |
| Bone marrow involvement, n (%) | 23 (23.0)                | 28 (28.0)            |

# First Simple (Well-Known?) Example

Table of baseline characteristics: Wrong conclusion

|                                | Experimental treatment B | Standard treatment A |
|--------------------------------|--------------------------|----------------------|
| Number of patients, n (%)      | 100 (100)                | 100 (100)            |
| Age                            |                          |                      |
| Median (range)                 | 62.0 (32-81)             | 60.5 (28-81)         |
| Male sex, n (%)                | 71 (71.0)                | 72 (72.0)            |
| ...<other characteristics>     | ...                      | ...                  |
| Bone marrow involvement, n (%) | 23 (23.0)                | 28 (28.0)            |

- So, 77% and 72% without bone marrow involvement?
- Biopsy results from all patients available?
- What about missing data?

# First Simple (Well-Known?) Example

Table of baseline characteristics

|                                | Experimental treatment B | Standard treatment A |
|--------------------------------|--------------------------|----------------------|
| Number of patients, n (%)      | 100 (100)                | 100 (100)            |
| Age                            |                          |                      |
| Median (range)                 | 62.0 (32-81)             | 60.5 (28-81)         |
| Male sex, n (%)                | 71 (71.0)                | 72 (72.0)            |
| ...<other characteristics>     | ...                      | ...                  |
| Bone marrow involvement, n (%) | 23 (23.0)                | 28 (28.0)            |

- So, 77% and 72% without bone marrow involvement?
- Biopsy results from all patients available?
- What about missing data?

# First Simple (Well-Known?) Example

Table of baseline characteristics: Option 2



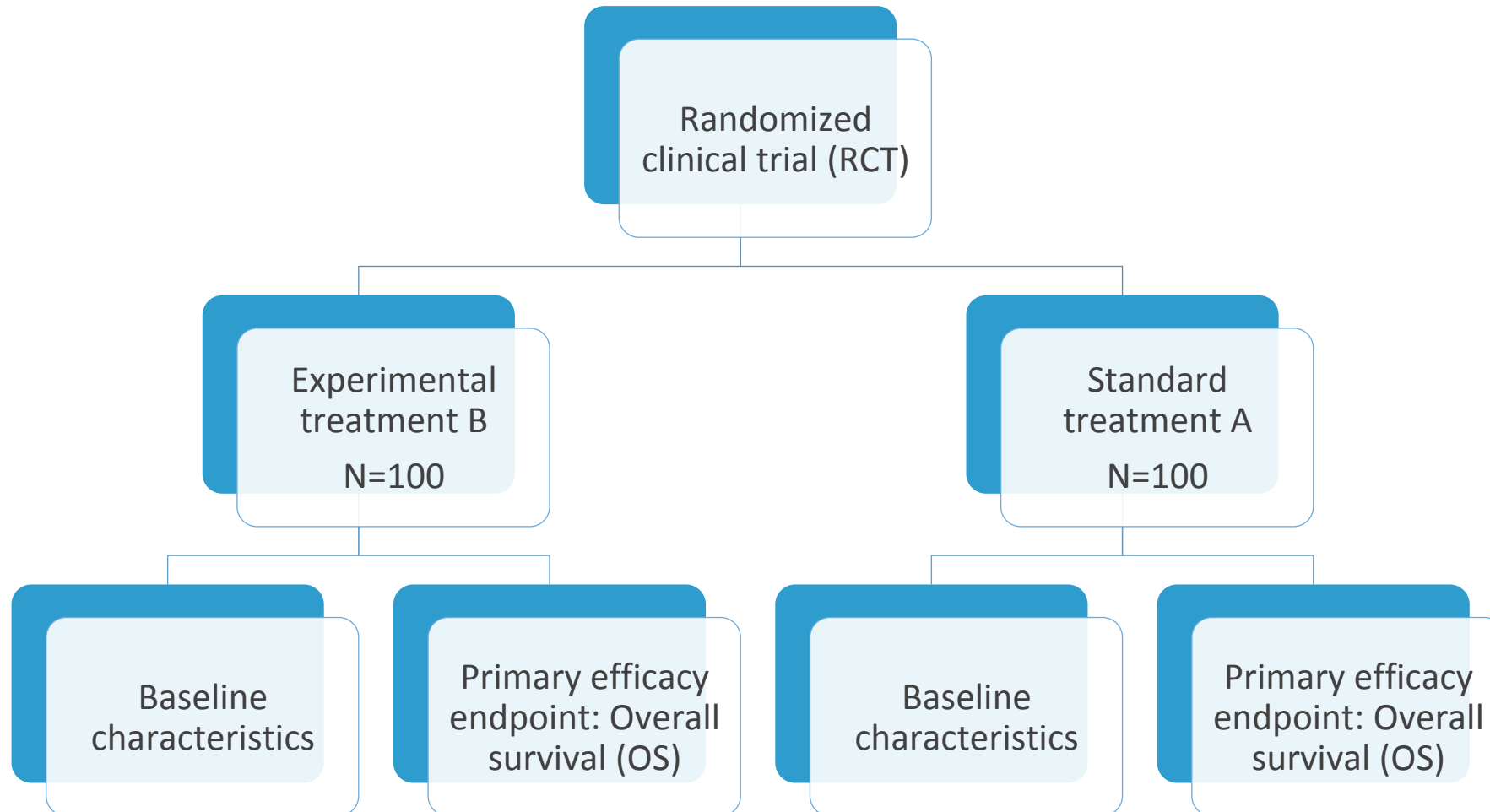
|                                 | Experimental treatment B | Standard treatment A |
|---------------------------------|--------------------------|----------------------|
| Number of patients, n (%)       | 100 (100)                | 100 (100)            |
| Age                             |                          |                      |
| Median (range)                  | 62.0 (32-81)             | 60.5 (28-81)         |
| Male sex, n (%)                 | 71 (71.0)                | 72 (72.0)            |
| ...<other characteristics>      | ...                      | ...                  |
| Bone marrow involvement*, n (%) | 23 (23.0)                | 28 (28.0)            |

- Percentages based on total number of patients
- So, 59% and 55% of patients without bone marrow involvement

\* 18 (18.0%) and 17 (17.0%) patients with missing information

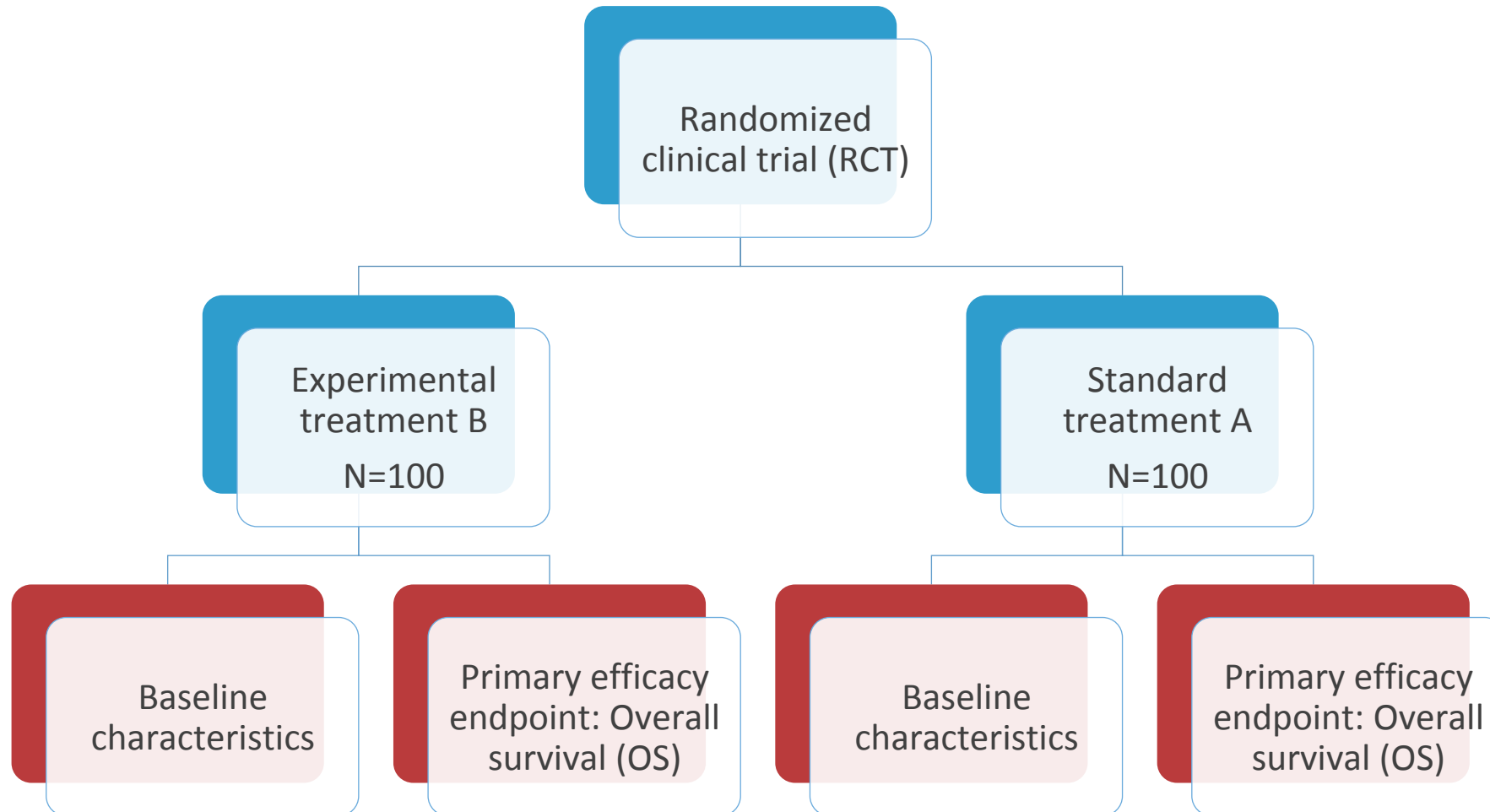
# Example: Subgroup Analysis

Question: Different treatment effect within subgroups?



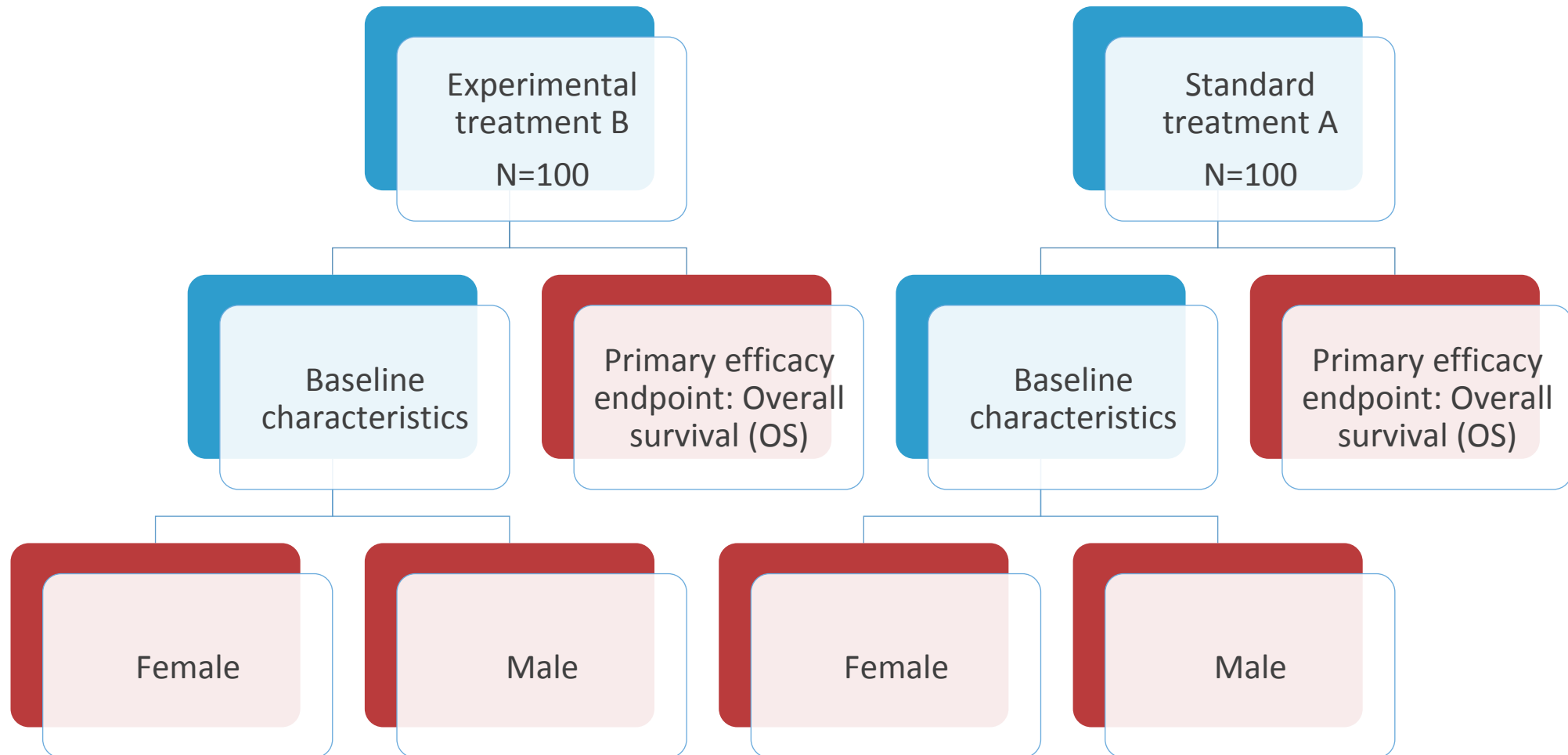
# Example: Subgroup Analysis

Question: Different treatment effect within subgroups?



# Example: Subgroup Analysis

Overall survival in female and male patients



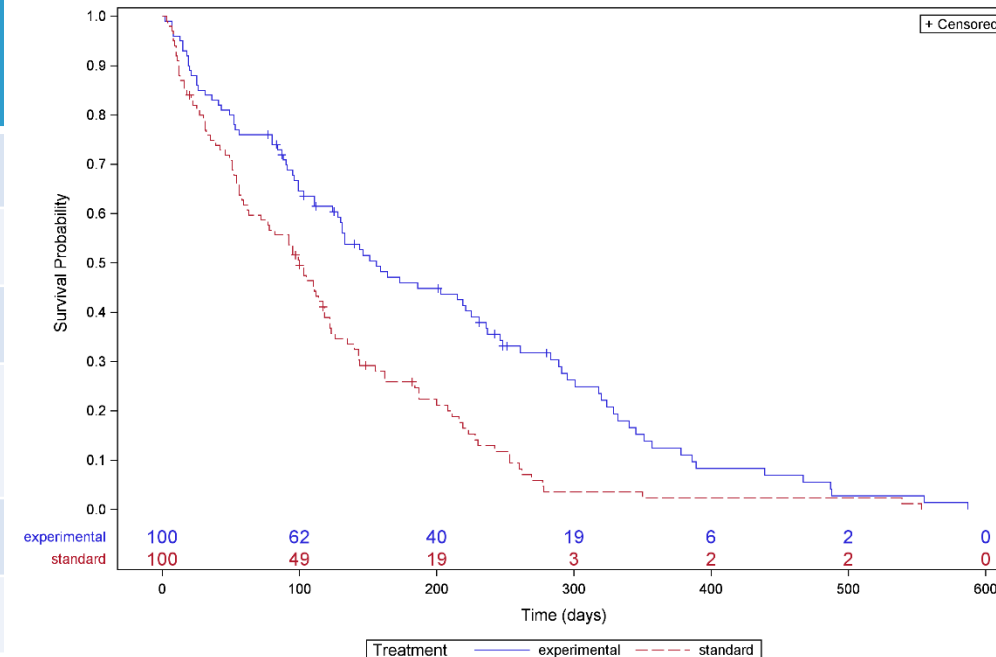
# Example: Subgroup Analysis

Main results in overall population

|                                          | Experimental<br>treatment B | Standard<br>treatment A |
|------------------------------------------|-----------------------------|-------------------------|
| Number of patients, n (%)                | 100 (100)                   | 100 (100)               |
| Overall survival                         |                             |                         |
| Number of patients with<br>events, n (%) | 87 (87.0)                   | 94 (94.0)               |
| Hazard ratio* (95% CI)                   | 0.54 (0.40, 0.73)           |                         |
| p-value*                                 | <0.0001                     |                         |

\* from Cox proportional hazard regression model

- Benefit of experimental treatment B over standard treatment A
- Is there a consistent benefit in male and female patients?



# Example: Subgroup Analysis

Overall survival in female and male patients



|                                       | Female                   |                      | Male                     |                      |
|---------------------------------------|--------------------------|----------------------|--------------------------|----------------------|
|                                       | Experimental treatment B | Standard treatment A | Experimental treatment B | Standard treatment A |
| Number of patients, n (%)             | 29 (100)                 | 28 (100)             | 71 (100)                 | 72 (100)             |
|                                       |                          |                      |                          |                      |
| Overall survival                      |                          |                      |                          |                      |
| Number of patients with events, n (%) | 25 (86.2)                | 26 (92.9)            | 62 (87.3)                | 68 (94.4)            |
| Hazard ratio* (95% CI)                | 0.64 (0.36, 1.13)        |                      | 0.51 (0.36, 0.73)        |                      |
| p-value*                              | 0.1241                   |                      | 0.0003                   |                      |

\* from Cox proportional hazard regression model

- Benefit of experimental treatment B over standard treatment A only for male patients?

# Example: Subgroup Analysis

## Limitations of subgroup analyses I

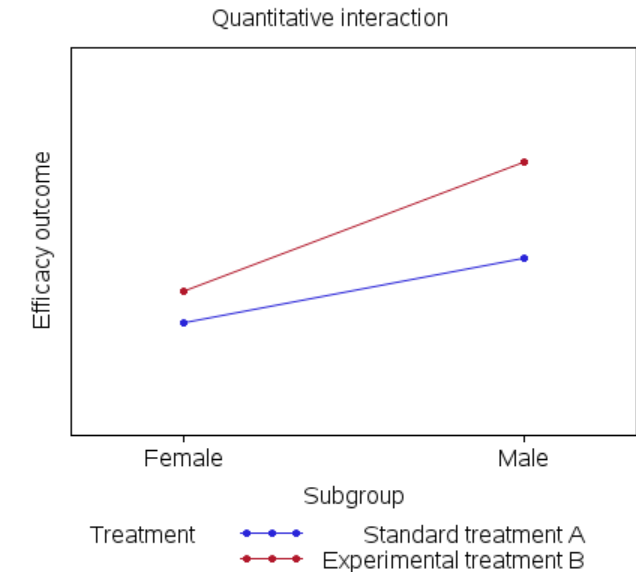
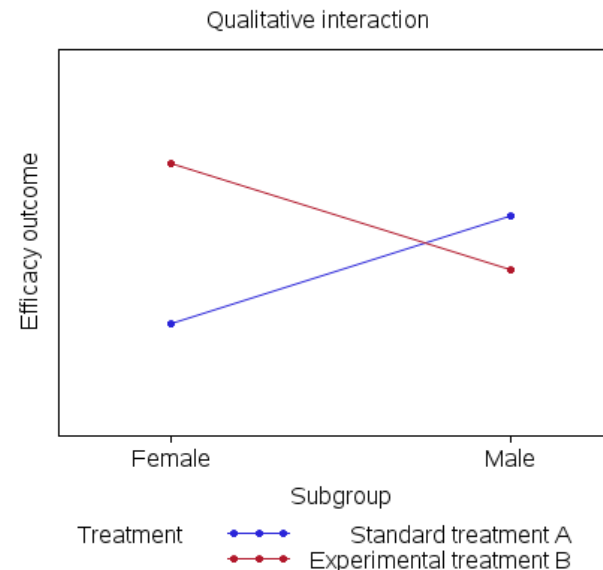
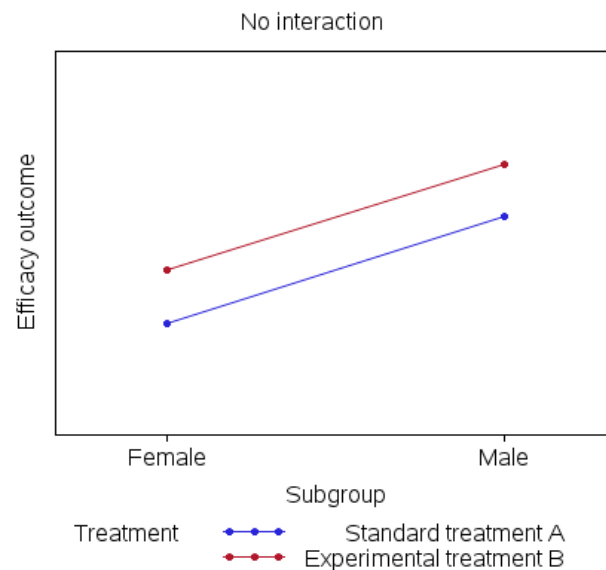
- Sample size much lower within the single subgroup categories and often not balanced between the categories (e.g. prevalence of disease is lower in females than in males)
- Study powered only for the analysis of the primary endpoint in the overall population, not for any subgroup analysis (even if pre-specified)
- Multiplicity: higher probability of incorrectly rejecting at least one null hypothesis (i.e. higher probability of incorrectly detecting a significant difference)

➡ Better approach: Interaction test

# Example: Subgroup Analysis

## Interaction test

- Interaction test: to check for an interaction between subgroup and treatment, i.e. if the treatment effect in one subgroup category is different to the treatment effect in the other subgroup category



# Example: Subgroup Analysis

Overall survival in female and male patients

|                                       | Female                   |                      | Male                     |                      |
|---------------------------------------|--------------------------|----------------------|--------------------------|----------------------|
|                                       | Experimental treatment B | Standard treatment A | Experimental treatment B | Standard treatment A |
| Number of patients, n (%)             | 29 (100)                 | 28 (100)             | 71 (100)                 | 72 (100)             |
|                                       |                          |                      |                          |                      |
| Overall survival                      |                          |                      |                          |                      |
| Number of patients with events, n (%) | 25 (86.2)                | 26 (92.9)            | 62 (87.3)                | 68 (94.4)            |
| Hazard ratio* (95% CI)                | 0.64 (0.36, 1.13)        |                      | 0.51 (0.36, 0.73)        |                      |
| Interaction p-value*                  | 0.3612                   |                      |                          |                      |

\* from Cox proportional hazard regression model

- No significant interaction, therefore no different treatment effect in males vs. females could be observed

# Example: Subgroup Analysis

## Limitations of subgroup analyses II

- The statistically not significant interaction p-value does not automatically imply that there is a consistent benefit of the experimental treatment B regardless of gender
- The type II error is unknown, which means you don't know the probability of failing to detect an interaction although there is actually an interaction
- The study is not powered for the interaction test, the number of patients may be too small to detect an interaction that is clinically relevant
- Multiplicity issue in case of more than one subgroup analysis
  - ➔ ■ Interaction p-value is an exploratory p-value
  - Do not focus only on the interaction p-value, look at HR, CI, number of available patients within subgroup categories

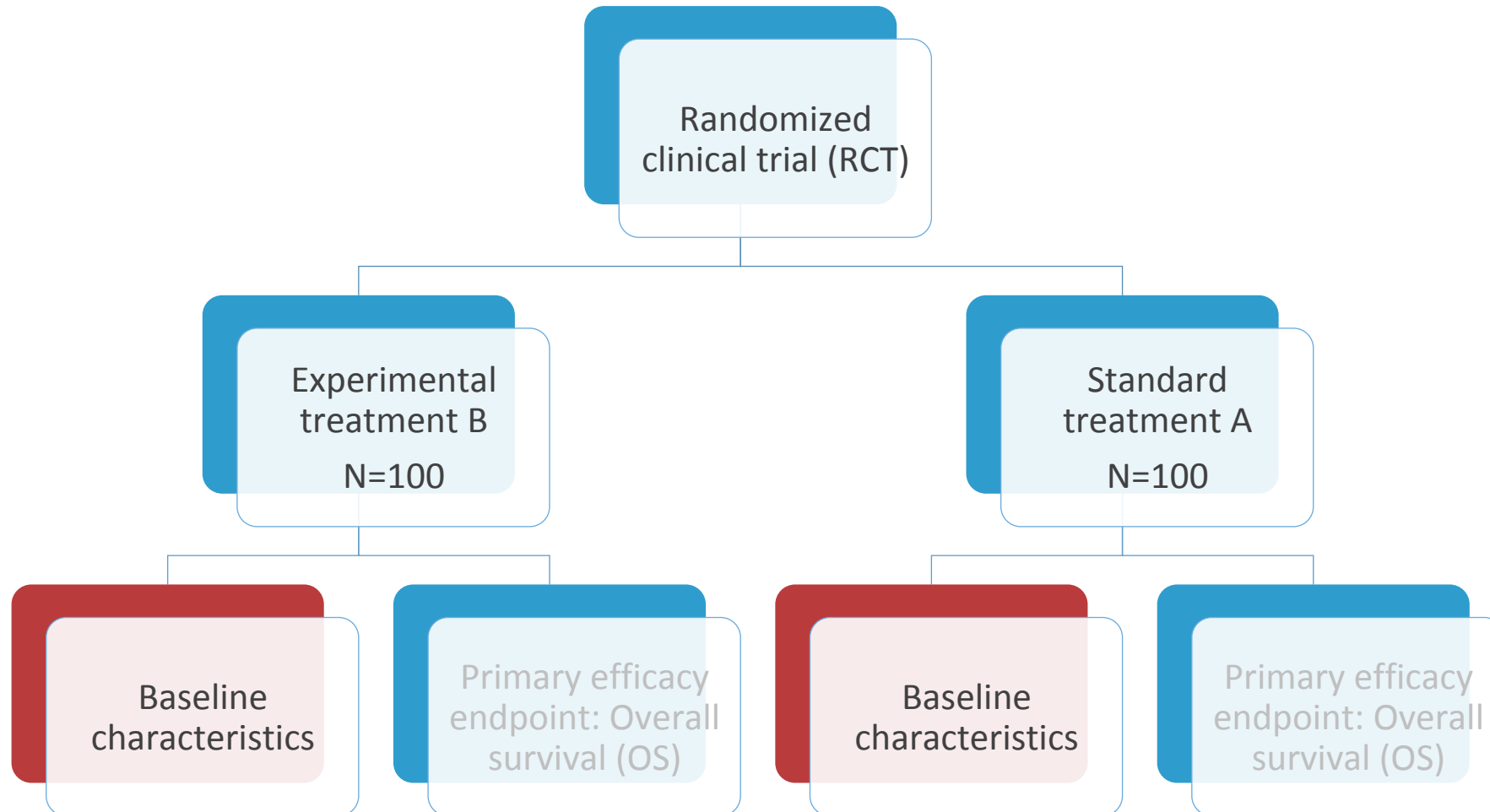
# Example: Subgroup Analysis

## Purpose of subgroup analyses

- Signal detection
  - Hypotheses generation
  - No strong conclusion should be drawn and any interpretation have to be done carefully
- ➔
- Words like "prove", "evident" or "demonstrate" should be avoided.
  - Instead it should be described what has been *observed*.
- See also “EMA Guideline on the investigation of subgroups in confirmatory clinical trials” (see references)

# Example: Comparability of Baseline Characteristics

Baseline characteristics



# Example: Comparability of Baseline Characteristics

Calculating p-values for treatment comparison



- Often requested by journals
- Reviewer performed a Chi-square test, which resulted in significant p-value for a baseline variable
- Reviewer's conclusion: imbalance favored treatment A
- *Is this reasonable? → NO!*

# Example: Comparability of Baseline Characteristics

Calculating p-values for treatment comparison



- If in a *randomized* trial the p-value is:

significant

- Just by chance
- Randomization went wrong

not significant

- you cannot conclude that the baseline variable did not affect the result of the primary endpoint
- Randomization has worked

# Example: Comparability of Baseline Characteristics

Calculating p-values for subgroup comparison



- Requested by a journal for a subgroup analysis: are baseline characteristics balanced between subgroup categories?
- Reviewer's assumption: selection of explanatory variables for a multivariable model was based on imbalances at baseline
- Imbalances in baseline variables are possible between subgroups
  - baseline characteristics are randomized, i.e. balance between treatments but not necessarily between subgroup categories (e.g. older patients have normally more concomitant diseases)
- *Is this reasonable? → NO!*

# Example: Comparability of Baseline Characteristics

Calculating p-values for subgroup comparison



■ If the p-value is:

significant

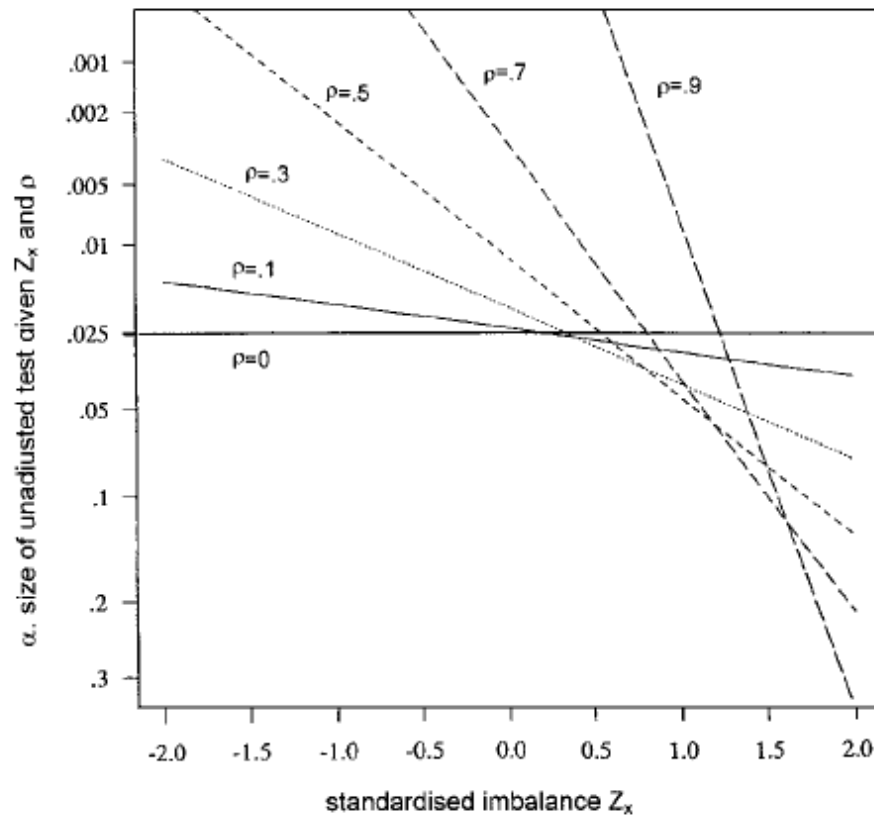
- Baseline variable is not balanced between subgroup categories
- Does not say that the baseline variable is related to the outcome

not significant

- Baseline variable is balanced between subgroup categories
- Does not say that the baseline variable is not related to the outcome

# Example: Comparability of Baseline Characteristics

## Covariate correlation vs. covariate imbalance

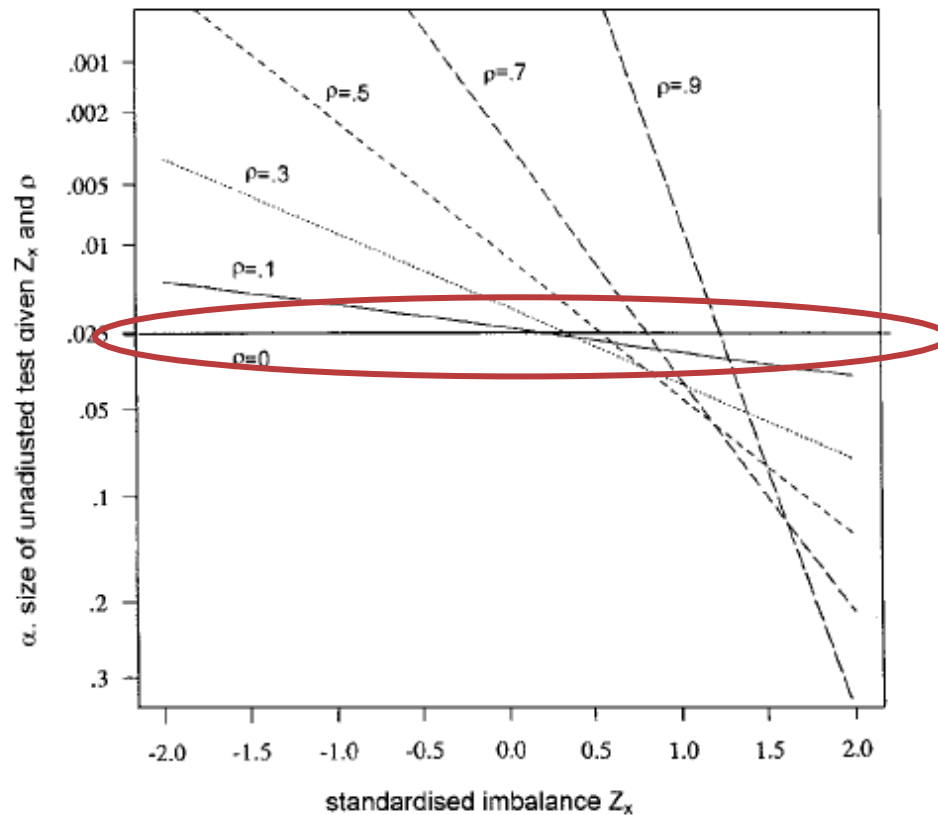


- From Pocock, S.J., Assmann, S.E., Enos, L.E. et al. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat Med, 21, pp. 2917-2930

Figure 2. The effect of standardized covariate imbalance  $Z_x$  and the covariate's correlation with outcome  $\rho$  on the conditional size of unadjusted one-sided test (true  $\alpha = 0.025$ ).

# Example: Comparability of Baseline Characteristics

## Covariate correlation vs. covariate imbalance

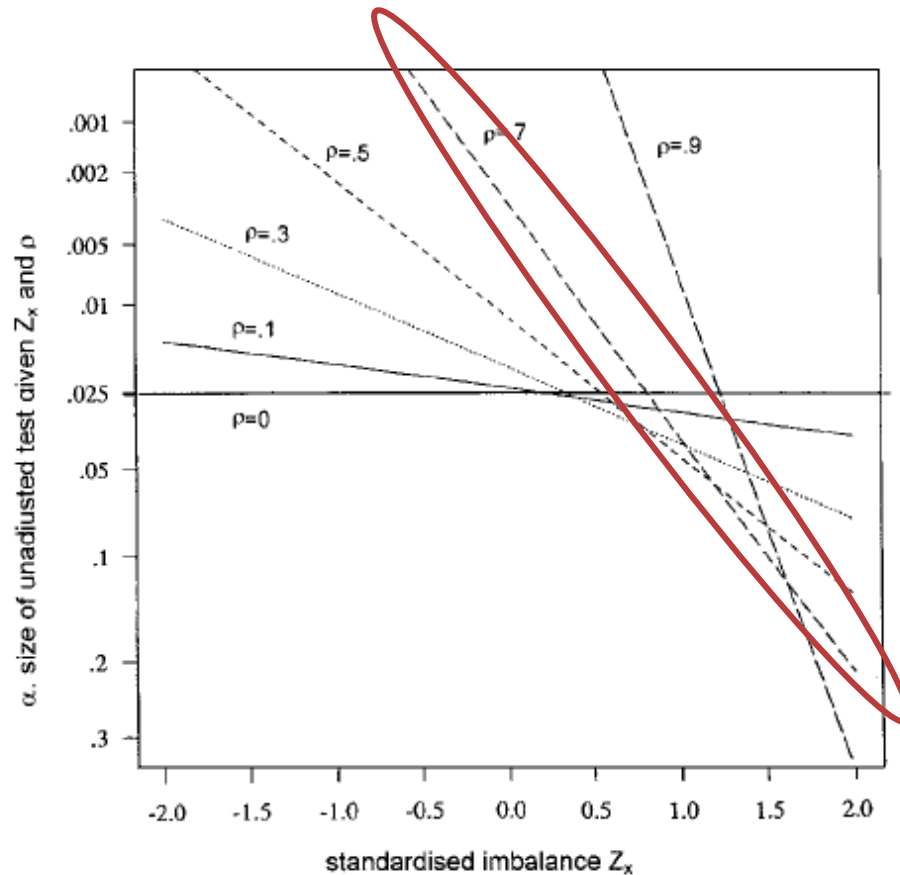


- From Pocock, S.J., Assmann, S.E., Enos, L.E. et al. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat Med, 21, pp. 2917-2930

Figure 2. The effect of standardized covariate imbalance  $Z_x$  and the covariate's correlation with outcome  $\rho$  on the conditional size of unadjusted one-sided test (true  $\alpha = 0.025$ ).

# Example: Comparability of Baseline Characteristics

## Covariate correlation vs. covariate imbalance



- From Pocock, S.J., Assmann, S.E., Enos, L.E. et al. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat Med, 21, pp. 2917-2930

Figure 2. The effect of standardized covariate imbalance  $Z_x$  and the covariate's correlation with outcome  $\rho$  on the conditional size of unadjusted one-sided test (true  $\alpha = 0.025$ ).

# Example: Comparability of Baseline Characteristics

## Further limitations/issues

- Multiplicity issue if several baseline variables are tested each at a level of e.g. 5%
  - Probability of incorrectly rejecting at least one null hypothesis, is (much) higher than 5%
- Focus on p-values = focus on statistical significance:  
Descriptive statistics may tell you more with regard to the *clinical relevance* than a p-value

# Summary

Please...

- Don't believe everything you see
- Check for handling of missing values
- Check if data is appropriately visualized
- Check if all data is included
- Check sample size (relevance of differences)
- Do an interaction test when performing subgroup analyses
- Distinguish between confirmatory and exploratory p-values
- Use words like “reasonable”, “considerable”, “relevant” instead of “significant” in case of exploratory p-values
- Use p-values sparingly – look at confidence intervals instead
- Consider both statistical significance and clinical relevance

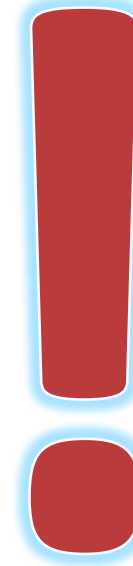
# Conclusion

What is essential

What is still necessary – even/especially when using data visualization and (new) analytical tools:

- Analysis plan: select appropriate statistical methods
- Validation of results
- Validation of underlying dataset
- (Statistical) output review

➔ Raise people's awareness of these essential facts



# THANK YOU!

Any questions?



## Selection of references

### Challenger disaster:

- Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989) Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure, *Journal of the American Statistical Association*, 84(408), pp. 945-957

### Comparability of baseline characteristics and covariate adjustment:

- Altman DG. Comparability of randomized groups. *The Statistician*. 1985;34:125-136
- Senn S. Testing for baseline balance in clinical trials. *Statistics in Medicine*. 1994;13:1715-1726
- Pocock SJ, Assmann SE, Enos LE et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med*. 2002;21:2917-2930

## Selection of references cont.

### EMA guidelines:

- European Medicines Agency: EMA/CHMP/539146/2013 – Guideline on the investigation of subgroups in confirmatory clinical trials. DRAFT  
[http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2014/02/WC500160523.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf). January 2014
- European Medicines Agency: EMA/CHMP/295050/2013 – Guideline on adjustment for baseline covariates in clinical trials.  
[http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2015/03/WC500184923.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500184923.pdf). February 2015

### CONSORT Statement:

- Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010. 2010;340:c332
- Moher D, Hopewell S, Schulz KF, Montori V, Gtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. J Clin Epi 2010; 63(8):e1-e37.

## Contact information



HMS Analytical Software GmbH  
Rohrbacher Str. 26  
69115 Heidelberg  
[www.analytical-software.de](http://www.analytical-software.de)  
[info@analytical-software.de](mailto:info@analytical-software.de)  
+49-6221-6051-0

Corinna Miede

Principal Statistician

+49- 6221 6051 196

[Corinna.Miede@analytical-software.de](mailto:Corinna.Miede@analytical-software.de)

Huteweg 4, 35096 Weimar (Lahn), Germany

