Paper AB11

Misuse and Misapplication in Statistical Data Analysis – A Topic that Never Goes Out of Style

Corinna Miede, HMS Analytical Software GmbH

Abstract

Data visualization and analytical tools are quite famous these days, especially in the field of data-driven science, as they facilitate creating nice plots or applying statistical tests intuitively and without code. That's why these tools seem time-saving and attractive not only for programmers and statisticians but also for other functions. However, they may also facilitate the misuse or misapplication of the data or statistical methods. But even if analyses are programmed conventionally, it is not guaranteed that the resulting analyses are correct and reliable. Sometimes requests from the clinical team or authorities are simply processed without questioning if these requests are reasonable. This can lead to false conclusions that may be published in manuscripts or shown at conferences. This presentation will show different examples of misuse and misapplication in statistical data analysis – a topic that has already been discussed many times but never goes out of style.

1 Introduction

For a classical confirmatory clinical trial it is required to specify the planned statistical analysis of the key endpoints in the clinical trial protocol before any actual data is available. Further details are later described in the statistical analysis plan [1]. The methods for the main analyses of the key endpoints is mostly predetermined by guidelines and previous studies that investigated the same indication or treatment. In addition, a sample size calculation has been done, so that at least for the primary endpoint the statistical analysis can be regarded as a sufficiently powered confirmatory analysis. Later on, after the study has been analyzed and the study report has been written, the main study results are published in a medical journal. In addition, further analyses are carried out to investigate the data in more detail. Certain subgroups are examined in order to check whether the treatment effect seen in the overall study population can also be observed in subgroups. There is a bit more flexibility with regard to the statistical methods, but often similar methods are applied as in the main study analysis. The results, however, need to be interpreted more carefully since the study is normally only powered for the primary endpoint in the overall population but not for any subgroup analysis. Therefore, these additional analyses can only be regarded as exploratory analyses.

An example randomized clinical trial is introduced in section 3 that is later utilized in section 3.1 to show a simple example of a common pitfall when presenting incomplete data. Section 3.2 illustrates important principles to follow when doing a subgroup analysis and discusses limitations. In addition, section 3.3 refers to the well-known misunderstanding when comparing baseline characteristics by means of calculating p-values. Although it is well-known and there are many publications about this topic, it is still not unusual to receive such requests from reviewers of medical journals.

However, let's first start in the the following section 2 with some basic examples of simple visualization errors caused by incorrect data selection.

The topic of misuse and misapplication in statistical data analysis is not new but seems to never go out of style because the following examples are based on recent personal experience. Furthermore, there are a lot of (new) data visualization and analytical tools these days that are advertized as efficient, easy to use, fast, powerful, interactive and secure, suggesting that these

innovative tools are not only time-saving but would also yield always reliable results. Therefore, they are quite attractive for programmers and statisticians as well as for other functions. But they may also facilitate the misuse or misapplication of the data or statistical methods since these tools may be prone to wrong data selection or misapplication of a statistical methods. Therefore, this paper would like to refresh the awareness that a careful handling of statistical methods and interpretation of results including validation is still mandatory regardless whether innovative or conventional tools are applied.

2 Visualizing Data

A famous example of how wrong data analysis can lead to terrible consequences is the Challenger space shuttle disaster in 1986. On 28 January 1986 the space shuttle exploded shortly after the launch and all 7 crew members died. Without going into the technical details of the causes leading to the disaster, a main issue was a malfunction of O-rings caused by low outside temperatures, so that hot gas could flow out of the tanks, which led to explosion [2]. Actually, the potential problem of the O-rings was discussed before the launch of the space shuttle and data was analyzed to check whether there is a correlation between the temperature and O-ring failure [3]. Only a few data points were available, so that a reliable conclusion was not possible. Nevertheless, a correct analysis of the limited data would have revealed a signal of a correlation. Figure 1 shows how the data was wrongly visualized shortly before the shuttle launch.

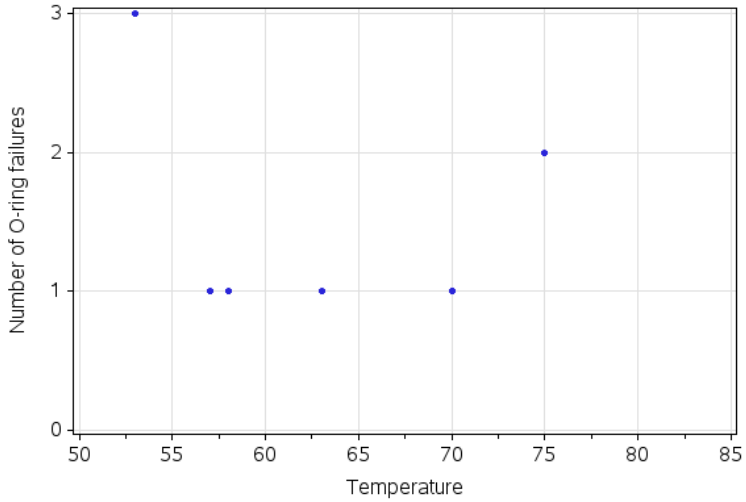

Figure 1: Wrong Visualization of O-ring Failure by Temperature (Data available in [3])

In contrast to Figure 1, the data is correctly presented in Figure 2. Figure 1 includes only data from flights with incidents of O-rings, whereas Figure 2 considers additionally flights without any incidence of O-ring failure. In Figure 1 no relationship between the temperature and O-ring failure could be seen. However, based on Figure 2 a signal is visible as all previous flights without any O-ring failures occurred at temperatures greater than 65°F without any exception. The technical causes that lead to the disaster are of course more complex and the impact of the temperature on the O-rings wasn't the only reason. Nevertheless, this correct and simple Figure 2 could have raised more serious concerns, so that the shuttle launch would potentially have been postponed especially because the temperature on the day of the shuttle launch was only 31°F.

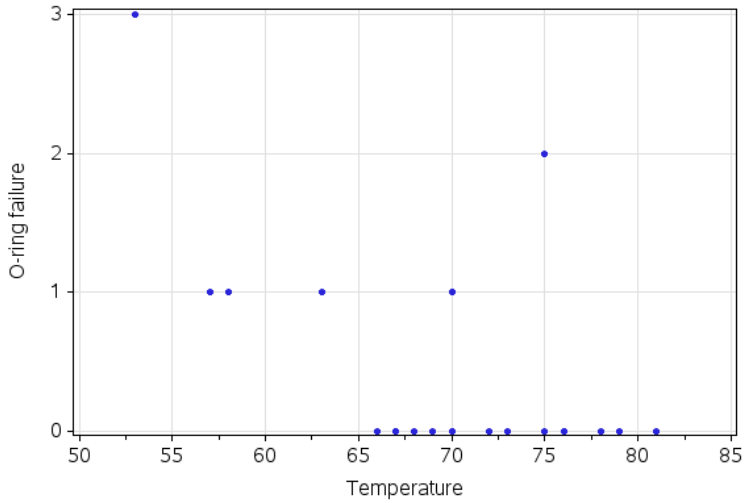

Figure 2: Correct Visualization of O-ring Failure by Temperature (Data available in [3])

A similar example seen within the field of a clinical study analysis is the visualization of the overall survival in a Kaplan-Meier plot. In the Kaplan-Meier plot displayed in Figure 3 the curves end at a probability of zero, thus all patients obviously died. However, the number of patients at risk at the bottom of Figure 3 indicates that not all patients have been included in the figure.

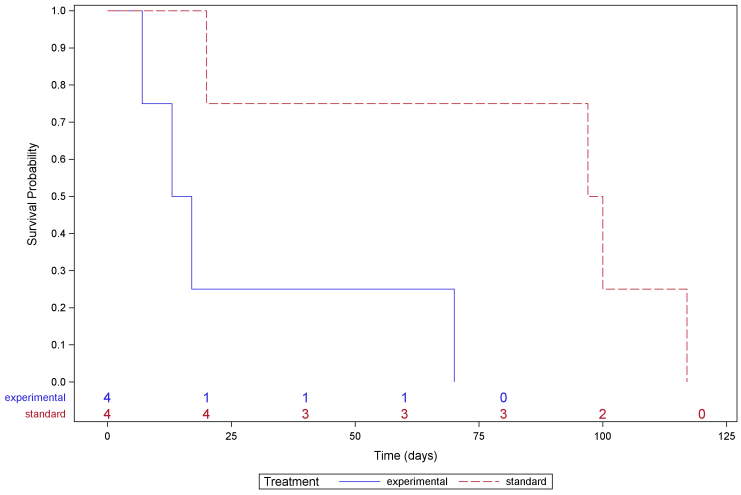

Figure 3: Incorrect Kaplan-Meier Plot of Overall Survival

In fact, only those patients with an event are taken into account but the Kaplan-Meier approach is actually a method that can handle censored data. Therefore, the patients without an event need to be considered as well to analyze the overall survival of the study population correctly. Figure 4 provides the correct overall survival and shows - in contrast to Figure 3 - that the majority of patients are still alive and that only few patients died.

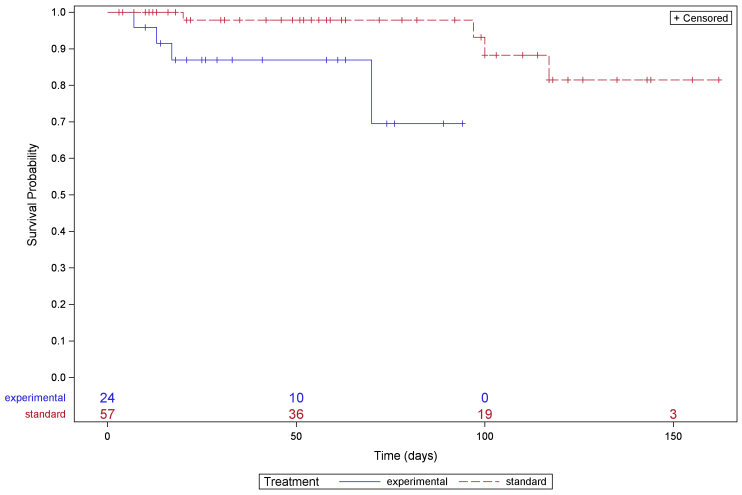

Figure 4: Correct Kaplan-Meier Plot of Overall Survival

3 Example Study

Let's introduce a sample study as displayed in Figure 5 that serves as basis for the following subsections. In a randomized clinical trial the overall survival (OS) of an experimental treatment B is compared with a standard treatment A.

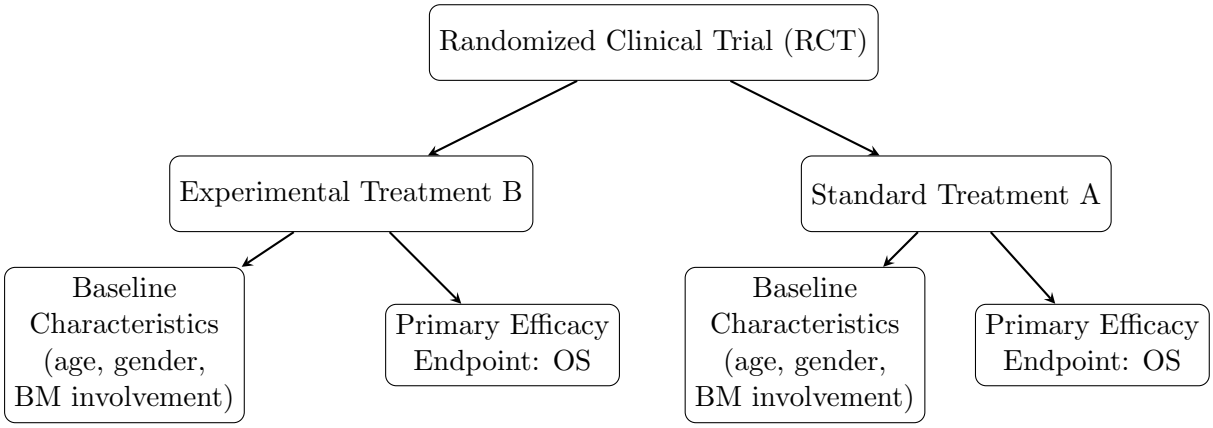

Figure 5: Example Study of a Randomized Clinical Trial

Some baseline and disease characteristics like gender, age and bone marrow involvement are collected. Let's assume that the sample size has been determined to achieve sufficient power (e.g. 80%) for the superiority comparison of OS between experimental treatment B and standard treatment A. The primary endpoint of OS is analyzed with a Cox proportional hazard regression model. The hazard ratio (HR) as well as the corresponding Wald 95% confidence interval (CI) and two-sided p-value are calculated.

3.1 Presenting Data with Missing Observations

When reporting or publishing the results of a clinical trial, tables presenting frequency distributions and/or descriptive statistics are always included. A clinical trial report including the source tables, listings, and figures as well as narratives and appendices has quite often several thousand pages, whereas a publication should summarize the results only on few pages. Most

journals have defined some requirements and limitations with regard to number of words or number of tables and figures. Therefore, the goal is to save space whenever possible. For example, in case of a binary variable only one category is included in the summary table because the number and percentage of patients in the other category can be calculated from the respective numbers of the reported category as long as the total number of patients is known. However, some bias may be introduced if missing data is not considered appropriately. In the following Table 1 showing some baseline characteristics of the example study as introduced in section 3, the included results for bone marrow involvement imply that there are 77% and 72% of patients in the experimental treatment group A and standard treatment group B, respectively, without bone marrow involvement. It is quite likely that there are some patients with missing data as the assessment of bone marrow involvement may be assessed by means of a biopsy, which was potentially not possible in all patients.

If this is actually the case, the above mentioned conclusion on the number of patients without

Table 1: Summary of Baseline Characteristics: Bad Example

	Experimental treatment B	Standard treatment A
No. of patients, n (%)	100 (100)	100 (100)
Age, years		
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)

bone marrow involvement would be wrong. Table 2 provides one option to avoid this misinterpretation without adding an additional row into the table but by providing additionally the number of patients with non-missing data and calculating the percentage based on this number. Thus, only 72% and 66% of patients, respectively, have a disease without bone marrow involvement if the number of patients with non-missing data is considered as basis for the percentages.

In case you prefer to calculated the percentages consistently based on the overall number

Table 2: Summary of Baseline Characteristics: Good Example (Option 1)

	Experimental treatment B	Standard treatment A
No. of patients, n (%)	100 (100)	100 (100)
Age, years		
N	100	100
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 (%)	23/82 (28.0)	28/83 (33.7)

of patients in each treatment group, Table 3 illustrates a second option that allows a correct interpretation. The information on patients with missing data is included in a footnote, which normally also saves space especially if there is more than one variable with missing data. Thus, there are only 59% and 55% of patients, respectively, without bone marrow involvement

considering the overall number of patients in each treatment group. Actually, both options

Table 3: Summary of Baseline Characteristics: Good Example (Option 2)

	Experimental treatment B	Standard treatment A
No. of patients, n (%)	100 (100)	100 (100)
Age, years		
N	100	100
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)

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

provide all required numbers, so that either percentages (based on overall population or only on non-missing data) could manually be calculated, whereas this is not possible with the data included in Table 1. However, the option as shown in Table 2 is not appropriate if the data cannot be considered as missing at random. For example, if the bone marrow biopsy was not performed because the physician suspected that there is no bone marrow involvement for particular patients. Consequently, most of these patients potentially have in fact no bone marrow involvement, which would lead to an underestimation of the percentage of patients without bone marrow involvement if the percentages are based on non-missing data only. Hence, the option as presented in Table 3 should be preferred - at least if a deviation from the "missing at random" assumption is possible. Generally, it is recommended to consistently report the numbers correctly, even in case of a very low percentage of patients with missing data that wouldn't change the interpretation.

3.2 Subgroup Analysis

In the randomized clinical trial introduced in section 3, the analysis of the primary endpoint reveals a beneficial effect of treatment B over a standard treatment A. Table 4 summarizes the

Table 4: Main Results from Cox Proportional Hazard Regression in Overall Population

	Experimental treatment B	Standard treatment A
No. of patients, n (%)	100 (100)	100 (100)
No. of patients with event, n (%)	87 (87.0)	94 (94.0)
HR* (95% CI)	0.54 (0.40, 0.73)	
p-value*	<0.0001	

main results in the overall population for the primary endpoint of overall survival. The corresponding Kaplan-Meier plot of overall survival is displayed in Figure 6. Additional subgroup analyses are performed to investigate whether a consistent treatment effect of treatment B over A is also observed in certain subgroups, e.g. in female and male patients. The results of the Cox proportional hazard regression model within the single subgroup categories of gender are shown in Table 5. The resulting Wald p-value is <0.05 for the male subpopulation and >0.05 for females. That there is a benefit of the experimental treatment B over treatment A only in males

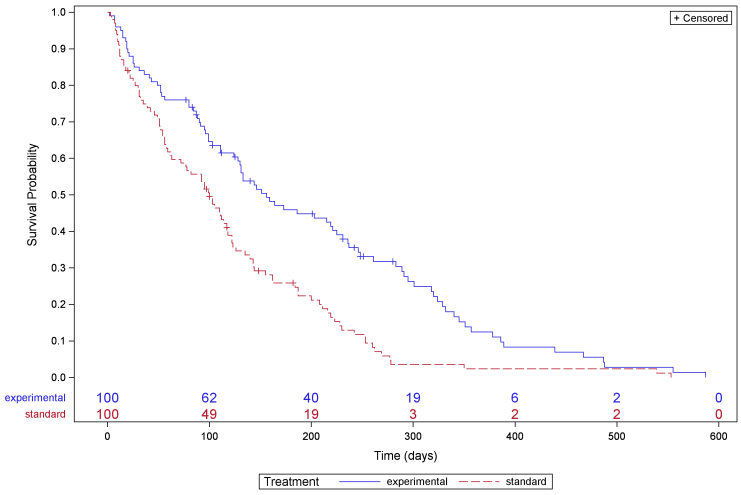

Figure 6: Kaplan-Meier Plot of Overall Survival in Example Study

is however a typical misinterpretation. Instead, the limitations of subgroup analyses need to be considered when interpreting their results. The sample size in each category is much lower as in the overall population, i.e. the study is only powered for the primary endpoint in the overall population, but not for a subgroup analysis, even if the subgroup analysis was pre-specified. In addition, the sample sizes are not balanced between the two categories, e.g. the prevalence of the underlying disease may be lower in females than in males. Furthermore, there is a mulit-plicity issue, so that without applying appropriate methods to control the family-wise error rate, the probability of incorrectly detecting at least one significant difference, is higher than the assumed global alpha level of 5%. Therefore, a simple analysis within the single categories

Table 5: Results from Cox Proportional Hazard Regression by Gender

	Female		Male	
	Experimental treatment B	Standard treatment A	Experimental treatment B	Standard treatment A
No. of patients, n (%)	29 (100)	28 (100)	71 (100)	72 (100)
No. of patients with event, n (%)	25 (86.2)	26 (92.9)	62 (87.3)	68 (94.4)
HR* (95% CI)	0.64 (0.36, 1.13)		0.51 (0.36, 0.73)	
p-value*	0.1241		0.0003	

is not an appropriate approach but an interaction term should be added into the statistical model instead. With the help of an interaction test, an interaction between the treatment and the subgroup, i.e. if the treatment effect in one subgroup category is different to the treatment effect in the other subgroup category, is explored. Figure 7 illustrates a qualitative and quantitative treatment by subgroup interaction in comparison to no interaction. In case there is no interaction between treatment and subgroup the curves are in parallel whereas the curves either do not run in parallel or even cross in case of a quantitative or qualitative interaction between treatment and subgroup. In other words, an interaction means that the treatment effect is more pronounced in one subgroup category compared to the other subgroup category or even the opposite treatment effect is observed.

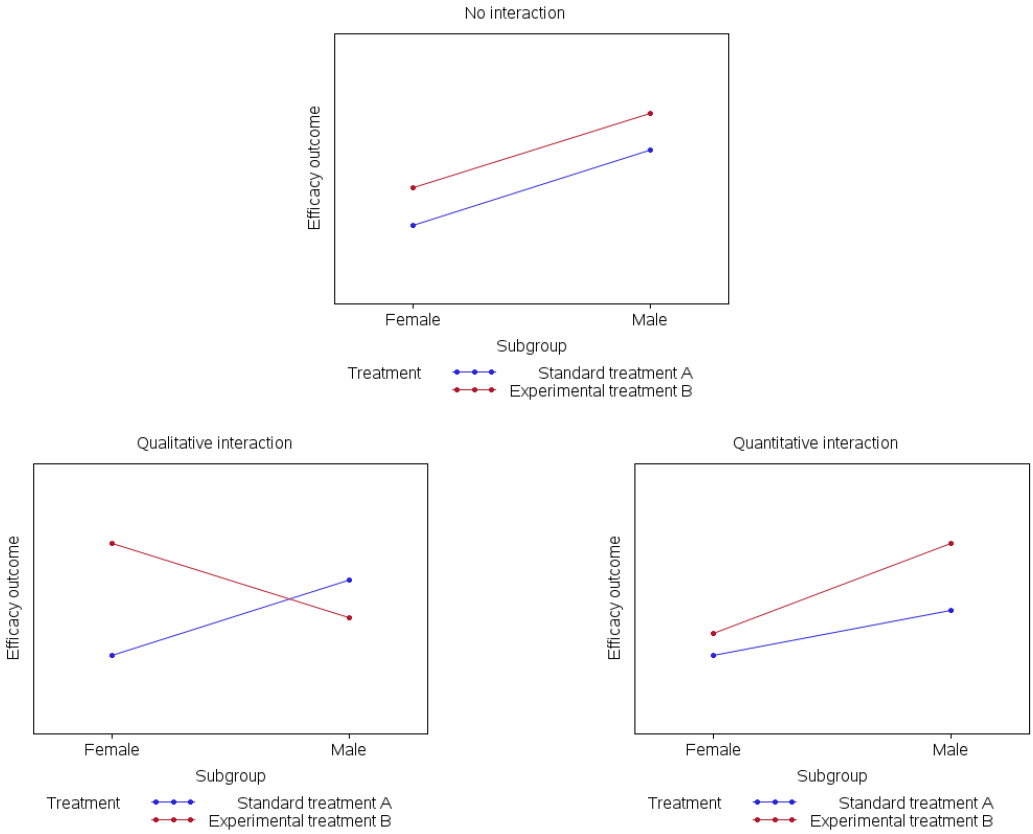

Figure 7: Interaction between Treatment and Subgroup

Table 6 provides the interaction p-value for an interaction between treatment and gender instead of the single Wald p-values from Table 5. The interaction p-value is >0.05, thus no statistical significant interaction between treatment and gender can be observed.

Table 6: Results from Cox Proportional Hazard Regression by Gender incl. Interaction p-value

	Female		Male	
	Experimental treatment B	Standard treatment A	Experimental treatment B	Standard treatment A
No. of patients, n (%)	29 (100)	28 (100)	71 (100)	72 (100)
No. of patients with event, n (%)	25 (86.2)	26 (92.9)	62 (87.3)	68 (94.4)
HR* (95% CI)	0.64 (0.36, 1.13)		0.51 (0.36, 0.73)	
Interaction p-value*	0.3612			

Limitations

The inclusion of an interaction term into the statistical model is the recommended approach for a subgroup analysis as also described in the EMA guideline on the investigation of subgroups in confirmatory clinical trials [7]. Nevertheless, there are still some limitations that need to be considered when interpreting results from subgroup analyses. The statistically not significant interaction p-value from Table 6 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 erroneously not detecting an interaction. As 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. Thus, the interaction p-value should be considered as an exploratory p-value. Additionally, if more than one subgroup analysis is performed or more than one endpoint is analyzed, there is again a problem of multiple testing, which is another reason, why these subgroup analyses should be regarded as exploratory analyses. Anyway, it is not appropriate to focus only on the interaction p-value but also the hazard ratio, its confidence interval, and the number of available patients within the single categories should be reported [7]. Considering all these results together helps with the interpretation and to differentiate between potentially more reliable results and those that are likely chance findings, e.g. due to a very small sample size. Generally, 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*. Hence, subgroup analyses can serve as signal detection or help to generate hypotheses.

3.3 Comparability of Baseline Characteristics

Baseline characteristics are analyzed in a clinical trial to describe the study population. In a randomized controlled clinical trial the results are generally reported by treatment group. Quite often medical journals request to calculate p-values for treatment comparison. However, these tests do not make sense for a randomized controlled clinical trial for various reasons. First of all, these tests are methods to assess whether treatment groups have randomly been assigned. In case of a randomized clinical trial any observed difference in baseline variables between the treatment groups occur per definition by chance and the treatment groups are balanced over all randomizations [9]. This means a significant p-value would be a chance finding or the randomization was in fact no real randomization (e.g. if there is a systematic issue in the randomization process). In addition, if several baseline variables are tested each at a level of e.g. 5%, there is a multiplicity issue, so that the family-wise error rate, i.e.the probability of incorrectly rejecting at least one null hypothesis, is (much) higher than 5%. Furthermore, journal reviewers often challenge a possible relationship between a baseline variable and the outcome variable, based on the resulting p-values. However, the results of these tests do not allow any conclusion whether a balance or an imbalance between the groups might had an impact on the results [8]. Especially, in case of non-significance you cannot conclude that the baseline variable did not affect the endpoint results. Also, in case of an imbalance between the groups no conclusion is possible that the baseline variable has an impact on the result the endpoint. Nevertheless, these conclusion are often done by the reviewers. Beside the various statistical publications that disuss this topic intensively (e.g. [9], [8], [4]) a group of experts developed the CONSORT (Consolidated Standards of Reporting Trials) statement [10]. Its "Explanation and Elaboration (E&E) Document" [11] provides a detailed guidance and explanations for correctly reporting randomized controlled clinical trials and refers - among many other items - to the above mentioned issues.

When comparing the subgroup categories (e.g. older vs. younger patients), the baseline characteristics are not necessarily balanced between the categories or imbalances are even expected (e.g. older patients have normally more concomitant diseases than younger patients). The same holds true for a treatment comparison in a non-randomized clinical trial. Nevertheless, if p-values for subgroup comparison or treatment comparison in a non-randomized clinical trial are calculated, there is still a multiple test problem. Moreover, a conclusion whether a baseline variable is related to an outcome variable based on these p-values is not possible in this situation either. The latter fact is especially important when exploring certain baseline variables by means of a covariate-adjusted model. Journal reviewers tend to request that selection of covariates to be included into a multiple regression model should be based on the observed imalances of baseline variables, which is not an appropriate approach. Pocock et al [8] describes

why "baseline imbalance" and "outcome correlation" are two different things and illustrates this with the following Figure 8, which is based on a treatment comparison by means of an analysis of covariance adjusting for the covariate x assuming a desired test size of 0.025 (i.e. true α). The test size of the corresponding unadjusted test is plotted versus values of the standardized imbalance Z_x (for more formal details, see [8]). As illustrated in Figure 8, a covariate that is

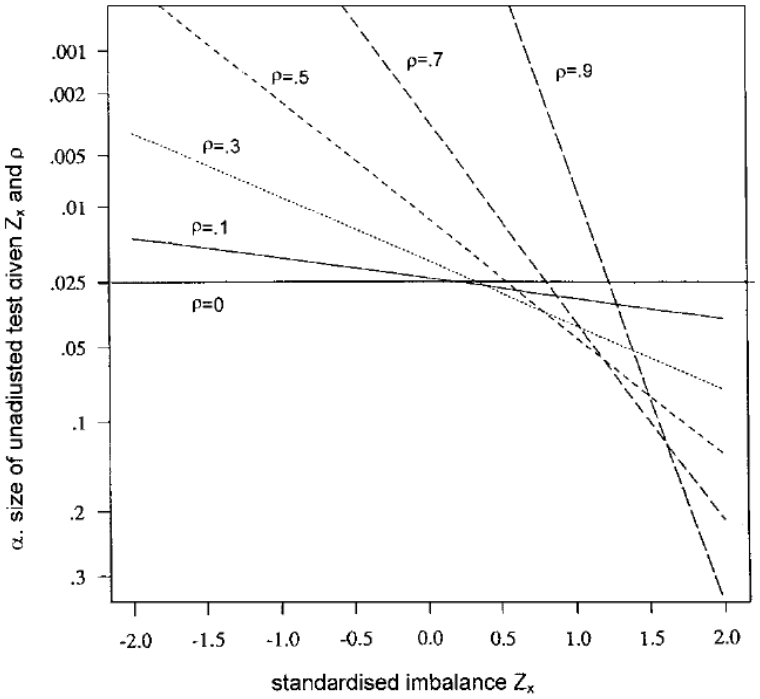

Figure 8: The effect of standardized covariate imbalance Z_x and covariate's correlation with outcome ρ on the conditional size of an unadjusted one-sided test, true α=0.025 (from [8])

not correlated with the outcome (i.e. ρ=0) has no impact on the size of the unadjusted test, regardless of if this covariate is balanced or highly imbalanced between the treatment groups (i.e. regardless of the treatment imbalance Z_x). On the other hand, if a covariate is completely balanced between treatment groups (i.e. Z_x=0) but correlated with the outcome (e.g. ρ=0.7) the resulting unadjusted α would become much lower than the true α of 0.025 [8]. Thus, the driving factor behind a covariate's impact on the outcome is the correlation of the covariate with the outcome variable, not the covariate imbalance. Generally, a covariate-adjusted model is an appropriate method to account for covariates that are correlated with the outcome both in a randomized and a non-randomized clinical trial. However, the selection of covariates for a covariate-adjusted model should be based on prior medical knowledge instead of any observed baseline (im)balance.

4 Conclusion

The described examples show only a small and subjective selection of potential misuse and misapplication in statistical data analysis. Nevertheless, they are based on recent personal experience. The aim of this paper is to raise awareness that the following actions are key regardless if an interactive analytical tool is used or a standard programming techniques:

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

Thus, in summary the following tips may help when interpreting results or analyzing clinical trials:

- 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 the sample size (relevance of a difference), look at confidence intervals instead of p-values.
- Do an interaction test when performing subgroup analyses instead of calculating p-values within the subgroup categories.
- Clearly distinguish between confirmatory and exploratory p-values.
- Use words like "reasonable", "considerable", "relevant" instead of "significant" in case of exploratory p-values.
- Generally, use p-values sparingly.
- Distinguish between statistical significance and clinical relevance, i.e. consider both when interpreting results - descriptive statistics may tell you more with regard to the clinical relevance than a p-value.

References

[1] The International Conference on Harmonization (ICH), Statistical principles for clinical trials: topic E9 (CPMP/ICH/363/96), https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf, 05 February 1998

[2] Report to the President by the Presidential Commission on the Space Shuttle Challenger Accident, https://spaceflight.nasa.gov/outreach/SignificantIncidents/assets/rogers_commission_report.pdf, June 1986, Washington, D.C

[3] Dalal, SR, Fowlkes EB, Hoadley B. Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. *Journal of the American Statistical Association*. 1989;84(408):945-957

[4] Altman DG. Comparability of randomized groups. *The Statistician*. 1985;34:125-136

[5] Dmitrienko A, Muysers C, Fritsch A, Lipkovich I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *Journal of Biopharmaceutical Statistics*. 2016;26(1):71-98

[6] 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. February 2015

[7] 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. January 2014

[8] 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

[9] Senn S. Testing for baseline balance in clinical trials. *Statistics in Medicine*. 1994;13:1715-1726

[10] 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

[11] Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche 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.

Acknowledgements

The author would like to thank Manuela Schmitz and Carolyn Cook from HMS Analytical Software for their valuable contribution to this paper as well as Martin Gregory from Merck for providing a LaTeX style and a template that have been used for writing this paper.

Contact Information

Corinna Miede
HMS Analytical Software GmbH
Huteweg 4
35096 Weimar/Lahn
Germany
Corinna.Miede@analytical-software.de