Published in final edited form as:

J Thorac Oncol. 2021 March ; 16(3): 375–380. doi:10.1016/j.jtho.2020.12.008.

Statistical Considerations for Subgroup Analyses

Xiaofei Wang[a], **Steven Piantadosi**[b], **Jennifer Le-Rademacher**[c], **Sumithra J. Mandrekar**[c]

[a]Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina

[b]Department of Surgery, Brigham and Women's Hospital, Boston, Massachusetts

[c]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota

Introduction

Randomized clinical trials (RCTs) are conducted to assess the effect of an experimental treatment on outcomes of a target patient population. Eligibility criteria for large trials are often broad to ensure the trial results can be generalized to a larger patient population. Subgroup analyses, either specified *a priori* or *post-hoc*, are performed to assess the treatment effect specific to a subgroup of treated patients. Regardless of whether a subgroup analysis is specified *a priori* or *post-hoc*, investigators must consider inflated false positive rates, chance differences in observed treatment effects, low power for the comparisons of interest, and interpretation of the subgroup results. Subgroup analyses in clinical trials and observational studies have been discussed in regulatory agency guidelines [1–3] and comprehensive review of this topic have been published in applied statistical journals [4–5]. This article reviews key statistical concepts associated with planning, conducting, and interpreting subgroup analyses in RCTs. It also highlights the pitfalls in conducting undisciplined subgroup analyses.

Defining Subgroups

In RCTs, subgroups are often defined by demographic variables, such as age, sex and race. Subgroups can also be defined by variables that are prognostic of clinical outcomes and/or predictive of better treatment effect, such as disease severity, prior therapies, genotype, and biomarker status. Using the same definition of subgroup enables comparison of outcomes between similar subgroups across different studies/clinical trials. If subgroups are defined by continuous variables, it is preferable to use well established or published cutoffs. For example, it is common in oncology to use age cutoffs of 40 years and 65 years to classify patients into age groups: <40 as adolescent and young adult, 40-65 as adult, and >65 as older adult. When common cutoffs are not available for a continuous biomarker, cutoff points can

Address for correspondence: Xiaofei Wang, Department of Bioinformatics & Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC 27710. xiaofei.wang@duke.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosure

The authors declare no conflict of interest.

be identified by data driven approaches such as simple percentiles (e.g. median), or visualization with statistical graphs. When multiple variables are expected to contribute to the definition of a subgroup, a continuous prediction score calculated from a multivariable prediction model may be used to categorize patients into low, moderate or high risk, indicating an ordinal increase in the risk of an adverse outcome or the severity for a disease condition. The cutoff points of a novel biomarker or a risk score are often chosen to maximize the difference in outcome or the difference in treatment benefits between the subgroups. When extensive preliminary data are available, statistical methods can be used to identify optimal boundaries to define subgroups based on high-dimensional variables.

Caution is necessary when defining subgroups by a data-driven approach. For example, are the subgroups plausible given what is known about the clinical, pharmacological, and biological mechanisms? Additional factors that affect the validity and the reproducibility of subgroup analyses are missing data and measurement error associated with the variables for searching and defining subgroups. Ill-defined and unreproducible subgroups could lead to biased and unreproducible estimates of treatment effects and hard-to-interpret findings. See Lipkovich et al (2017) [6] for a comprehensive review of data-driven methods to define subgroups using clinical trials data. Although subgroups can be defined using continuous variables, the loss of power and efficiency from categorization of continuous variables should be considered when defining subgroups in this manner [7].

Heterogeneity of Treatment Effect

The effect of a treatment may vary by baseline patient characteristics (e.g. gender, age, and race), tumor's molecular profile and genotype, and medical centers. Targeted agents or immunotherapy, for example, are expected to be effective for patients with certain genetic mutations or protein/genetic overexpression levels due to their mechanism of action. In such cases, a favorable treatment effect in a targeted subgroup is anticipated whereas the treatment would likely have no effect or even a deleterious effect in the complementary subgroup, or in the unselected population.

The statistical term to describe differential treatment effects by subgroups is the interaction between the treatment and the subgroup variable. If there is no interaction between treatment and subgroup, the treatment effect is the same across subgroups, as in Figure 1 (A). If there exist differential treatment effect across subgroups, two types of treatment-by-subgroup interaction have been described -- quantitative versus qualitative interaction. In quantitative interaction, the size of the treatment effect varies across subgroups, but the treatment effects in different subgroups are in the same direction. Figure 1 (B) shows an example of quantitative interaction where although the magnitude of the treatment effect differs between the two subgroups, they are in the same direction. In such cases, the therapeutic implications are unchanged by the interaction. In qualitative interactions, the treatment benefit is in favor of the experimental treatment in one subgroup but is unfavorable or neutral for the other subgroup. Figure 1 (C) shows an example of qualitative interaction where the treatment effects are in opposite directions. These circumstances carry important therapeutic consequences.

One example of qualitative interaction is from the IPASS trial for non-small cell lung cancer (NSCLC) [8]. As illustrated in Figure 2, gefitinib (an EGFR inhibitor) was associated with significantly better progression-free survival (PFS) compared to control (carboplatin plus paclitaxel) in EGFR mutants, while gefitinib was associated with significantly worse PFS compared to control in EGFR wild types. In this example, EGFR mutation is a predictive biomarker for gefitinib in NSCLC. A predictive biomarker, or more generically a predictive factor, gives information about the treatment effect of a new intervention relative to a control. In a contrast, a prognostic biomarker (factor) provides information about the patients overall cancer outcome, regardless of treatment. For example, the 21-gene recurrence score is associated with recurrence in tamoxifen-treated patients with node-negative, estrogen receptor (ER)–positive breast cancer [9]. It is important to point out that a predictive biomarker (factor) could have either quantitative or qualitative interaction with treatment. More discussion on prognostic and predictive biomarkers in cancer research can be found in Mandrekar and Sargent (2009) [10] and Ballman (2015) [11].

Meta-analysis evaluates the heterogeneity of treatment effects reported from multiple clinical studies and synthesizes the estimates of the treatment effects. When heterogeneity is not significant, pooling the information from multiple studies is preferred as it usually increases power. More information on meta-analysis can be found in Higgins et al (2019) [12].

Control of Type I Error Rate and Statistical Power

Statistical regression models are commonly used to test for treatment-by-subgroup interactions. The independent variables in these models include the main treatment effect, the variable defining the subgroups, and the treatment-by-subgroup interaction terms. Other prognostic or confounding variables of the outcomes can also be included in these models to adjust for their potential association with the outcomes of interest. The estimates of treatment-by-subgroup interactions from these models can be interpreted as the differential treatment effects, i.e. the differences in the treatment effects among various subgroups. When there is sufficient statistical evidence to indicate an interaction, it is important to state the treatment effects by subgroups and not use the overall average treatment effect on all patients to characterize the effect of treatment across all subgroups. For RCTs designed to evaluate the effect of an experimental treatment in the overall patient population, statistical tests to detect a treatment-by-subgroup interaction are often under-powered. Therefore, failure to detect a significant interaction does not automatically imply the absence of treatment effect heterogeneity.

Forest plots are a useful tool to visualize the treatment effect across subgroups [13–14]. They usually present confidence intervals around the point estimate of subgroup treatment effect and the size of the symbols used is proportional to the size of subgroups. The treatment effect estimates and standard errors used to create the forest plot should not be derived from fitting separate models for the subgroups, but from a model on the full set of data, also including the interaction term between the treatment and the subgroup variable if interaction is significant.

It is important to control for type I error rate when testing for treatment effects in multiple subgroups. The practice of conducting multiple statistical tests on a large number of subgroups to find a subgroup with significant treatment effect runs the high risk of inflating the overall type I error rate. For example, the chance of making at least one false positive claim for conducting 10 repeated tests at 5% significance level is 40% when there is no treatment effect. The Bonferroni method addresses the issue of multiplicity by lowering the critical value (making it harder) to declare statistical significance and therefore controls the false positive rate to its specified level. There are many other methods that can be used to control the overall type I error rate when subgroup analysis is of interest.

In the development of cancer targeted agents and immunotherapy, RCTs are conducted with an interest in demonstrating the treatment effect in all patients and/or in a targeted subgroup. A simple gating approach is to sequentially test for the overall treatment effect before evaluating the effect within subgroups and once the null hypothesis of the overall treatment effect fails to be rejected, no subgroup analysis should be conducted. One criticism of the gating approach is that one might miss the opportunity to identify a targeted subgroup of patients who may benefit from the new treatment based on prior preclinical or early clinical observations. An alternative to sequential approach is the Bonferroni method where the type I error rate is split between the overall test and the subgroup-specific tests. In general, when a specified level of overall type I error rate is allocated to testing multiple hypotheses with respect to different patient subgroups or endpoints and each test is conducted at its allocated significance level, the overall type I error rate will not exceed its specified level regardless if these tests are dependent or not. The Bonferroni approach is easy to implement but could be overly conservative, as it ignores the correlation between outcomes of patients in the subgroups and those of the overall patient group. More flexible multiple testing procedures have been proposed and they take the correlation and the testing order into account as well as allow recycling the significance level after rejecting a hypothesis to test the remaining hypotheses at increased significance levels. Reviews of these more flexible methods, such as the fallback procedure [15] and MaST procedure [16], can be found in Matsui et al (2014) [17] and Dmitrienko et al (2017) [18].

As mentioned, the power of treatment-by-subgroup interaction test is often low in a RCT in which the treatment effect of the overall population is of primary interest. For a test of treatment-by-biomarker interaction with equal numbers of subjects in all subgroups, the interaction test will have roughly four times the variance of an overall treatment effect. To compensate, the size of the trial would have to be increased accordingly which is seldom done. This increase may be mitigated if the interaction is of large size or qualitative, but tests for these can also have low power under certain situations [19–20]. Besides the size and nature of an interaction, the allocation ratio among subgroups also has an impact. Equal allocation yields the highest power. See Wang et al (2018) [21] for further discussion of strategies to optimize the power of treatment-by-subgroup interactions in biomarker-stratified clinical trials.

Confirmatory versus Exploratory Subgroup Analysis

Confirmatory subgroup analyses are intended to evaluate subgroup-specific treatment effects and require that the subgroup analyses be specified in the design of the trial. In confirmatory cases, the subgroups must be clearly defined and the endpoints delineated with a small number of hypotheses about subgroup-specific treatment effects. A plan for control of the type I error rate and adequate power for testing of the overall and subgroups treatment effects must be pre-specified. Pre-specified hypotheses of confirmatory subgroup analysis are based on strong biological evidence. Findings from confirmatory subgroup analysis may be used in applications for regulatory approval of the new treatment in the subgroups. See Ondra et al (2016) [22] for examples and reviews of confirmatory subgroup analysis in cancer clinical trials that test targeted agents and immunotherapy.

Unlike confirmatory subgroup analysis, exploratory subgroup analyses are intended to obtain preliminary evidence and to generate hypotheses for future investigation. Exploratory subgroup analyses are often conducted either in a post-hoc manner or are pre-specified at the design stage but without sufficient power to formally test for subgroup-specific treatment effects. Plan for pre-specified exploratory subgroup analyses should include subgroup definition, endpoints, and how the subgroup analyses will be carried out. An example of pre-specified exploratory subgroup analysis is the differential treatment effect between EGFR-positive and EGFR-negative patients in the IPASS trial, achieved through testing the treatment-by-biomarker interaction [8]. Results of exploratory subgroup analyses should be clearly labeled as such, and reported (in tables and forest plots) with point estimates and confidence intervals without p-values. If the investigators are so inclined to the report of p-values, they should be adjusted for multiplicity to control familywise type I error rate or false discovery rate. See Table 1 for a summary of the features of confirmatory and exploratory subgroup analysis. Lagakos (2006) [23] and Wang et al (2007) [24] provide guidelines for presentation of subgroup analysis in medical journals.

Conclusion

Reporting of treatment effects on subgroups of patients are common in the medical literature. Although knowing how well a new treatment works in patients with a specific biomarker or a specific combination of disease stage and histology are important for the patients and their clinicians to make informed treatment decisions, it is important to be cautious when interpreting the results of subgroup analyses. Common challenges with subgroup analyses include poor definitions, low statistical power, and inflated type I error due to multiple hypotheses testing. The decision to conduct a subgroup analysis should depend biological justification or on evidence of treatment heterogeneity from existing preliminary data. When a large number of unplanned subgroup analyses are conducted, the treatment may spuriously appear to be effective in one or more subgroups. One should always assess the validity of subgroup analysis results based on biological plausibility, sample size for the subgroup, proper type I error control and power.

Acknowledgement

The research work was partially supported by P01CA142538 (XFW) and R01AG066883 (XFW), and P30CA15083 (Mayo Clinic Comprehensive Cancer Center Grant; JLR and SJM).

References

1. Lewis JA (1999). Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. Statistics in Medicine, 18(15), 1903–1942. [PubMed: 10440877]
2. FDA Guidance for Industry (2012). Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products.
3. European Medicines Agency. (2014). Guideline on the investigation of subgroups in confirmatory clinical trials. EMA/CHMP/539146.
4. Dmitrienko A, Muysers C, Fritsch A, & Lipkovich I (2016). General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. Journal of Biopharmaceutical statistics, 26(1), 71–98. [PubMed: 26366479]
5. Alosh M, Huque MF, Bretz F, & D'Agostino RB Sr (2017). Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Statistics in Medicine, 36(8), 1334–1360. [PubMed: 27891631]
6. Lipkovich I, Dmitrienko A, D'Agostino RB (2017). Tutorial in Biostatistics: Data-driven subgroup identification and analysis in clinical trials. Statistics in Medicine. 36, 36–196.
7. Altman DG, & Royston P (2006). The cost of dichotomising continuous variables. BMJ, 332(7549), 1080. 10.1136/bmj.332.7549.1080 [PubMed: 16675816]
8. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, … & Nishiwaki Y (2009). Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. New England Journal of Medicine, 361(10), 947–957.
9. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, & Hiller W (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine, 351(27), 2817–2826.
10. Mandrekar SJ, & Sargent DJ (2009). Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. Journal of Clinical Oncology, 27(24), 4027. [PubMed: 19597023]
11. Ballman KV. (2015). Biomarker: predictive or prognostic? Journal of Clinical Oncology, 33:3968–3971. [PubMed: 26392104]
12. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, & Welch VA (Eds.). (2019). Cochrane handbook for systematic reviews of interventions. John Wiley & Sons.
13. Cuzick J (2005). Forest plots and the interpretation of subgroups. The Lancet, 365(9467), 1308.
14. Ou F, Le-Rademacher JG, Ballman KV, Adjei AA, & Mandrekar SJ (2020). Guidelines for Statistical Reporting in Medical Journals. Journal of Thoracic Oncology, 15(11), 1722–1726. [PubMed: 32858236]
15. Bretz F, Maurer W, Brannath W, & Posch M (2009). A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 28(4), 586–604. [PubMed: 19051220]
16. Freidlin B, Korn EL, & Gray R (2014). Marker sequential test (MaST) design. Clinical trials, 11(1), 19–27. [PubMed: 24085774]
17. Matsui S, Choai Y, & Nonaka T (2014). Comparison of statistical analysis plans in randomize-all phase Ill trials with a predictive biomarker. Clinical Cancer Research, 20(11), 2820–2830. [PubMed: 24691019]
18. Dmitrienko A, Millen B, & Lipkovich I (2017). Multiplicity considerations in subgroup analysis. Statistics in Medicine, 36(28), 4446–4454. [PubMed: 28762525]
19. Gail M and Simon R (1985). Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. Biometrics 41(2): 361–372. [PubMed: 4027319]
20. Piantadosi S & Gail MH (1993). A comparison of the power of two tests for qualitative interactions. Statistics in Medicine 12(13): 1239–1248. [PubMed: 8210823]

21. Wang X, Zhou J, Wang T, & George SL (2018). On Enrichment Strategies for Biomarker Stratified Clinical Trials. Journal of Biopharmaceutical Statistics, 28(2) 292–308. [PubMed: 28933670]

22. Ondra T, Dmitrienko A, Friede T, Graf A, Miller F, Stallard N, & Posch M (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. Journal of Biopharmaceutical Statistics, 26(1), 99–119. [PubMed: 26378339]

23. Lagakos SW (2006). The challenge of subgroup analyses–reporting without distorting. New England Journal of Medicine. 354 (16): 1667–9.

24. Wang R, Lagakos SW, Ware JH, Hunter DJ, & Drazen JM (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. New England Journal of Medicine, 357(21), 2189–2194.

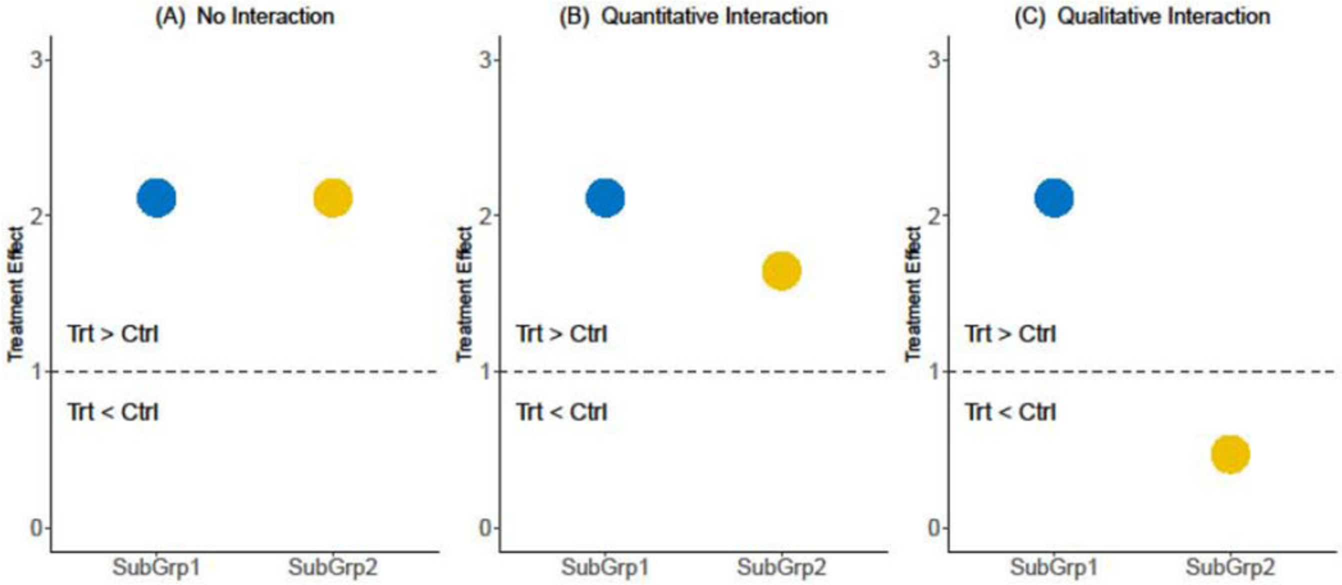

Figure 1.

Types of treatment-by-subgroup interaction – no interaction (A), quantitative interaction (B) and qualitative interaction (C). The Y-axis denotes the treatment effect, e.g. the Hazard Ratio (HR) for Ctrl vs. Trt. A treatment effect 1 indicates the treatment is as effective as the control; a treatment effect greater than 1 indicates the treatment is better than the control; and a treatment effect less than 1 indicates the treatment is worse than the control.

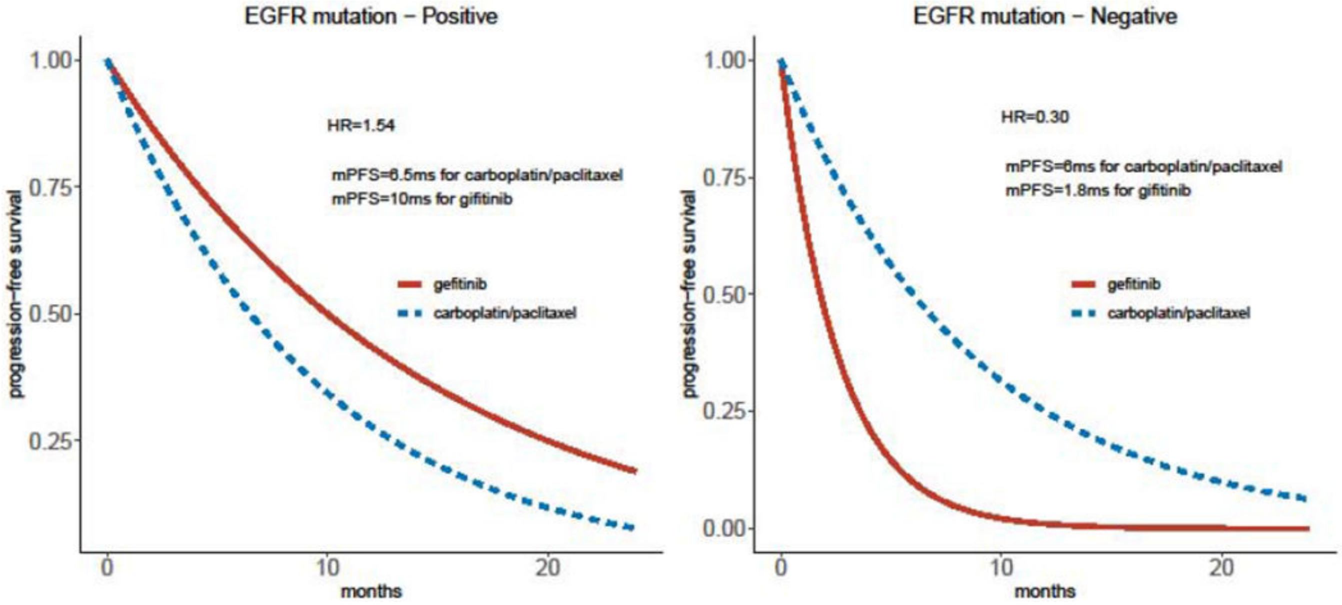

Figure 2.

Schematic plot of qualitative interaction between treatment and EGFR mutation observed in IPASS trial.

Table 1.

Comparison of Confirmatory versus Exploratory Subgroup Analyses

	Confirmatory Subgroup Analysis	Exploratory Subgroup Analysis
Purpose	confirm heterogeneity of treatment effect across subgroups	explore the possibility of treatment effect heterogeneity across subgroups
Pre-specified/Post-hoc	always pre-specified	pre-specified or post-hoc
Subgroup definition	pre-defined	pre-defined or to be discovered
Endpoints	always pre-specified	pre-specified or post-hoc
Null[1] and alternative[2] hypotheses	pre-specified	pre-specified or post-hoc
Number of subgroup analyses	limited	moderate or large number
Multiplicity	carefully chosen multiple testing procedure (MTP) for proper overall type I error rate[3] control	control overall type I error rate[3] at higher level e.g. 10%, or without any control
Size of subgroups	pre-determined	without planning
Power[5]	low type II error rate[4] sufficiently powered	high type II error rate[4] very low power
Interpretation	clear	unclear, sometimes impossible to interpret correctly
Reporting	mandatory	selective

[1] Null hypothesis (H_0): a statement about the effect of the new treatment that the investigators hope to not be true. For example, the new treatment is worse than or as effective as the control.

[2] Alternative hypothesis (H_1): a statement about the effect for the new treatment that the investigators hope to be true. For example, the new treatment is superior or not the same as the control.

[3] Type I error rate: The probability of making false claim that the new treatment is superior or not the same as the control (H_1 is true) when the new treatment is not as effective compared to the control (H_0 is true).

[4] Type II error rate: The probability of making a claim that the new treatment is the same as the control (H_0 is true) when the new treatment is more effective than the control (H_1 is true).

[5] Power: The probability of rejecting the null hypothesis (H_0) at a given size of effect for the new treatment (H_1 is true). Power is equal to 1 minus type II error rate.