Practical and statistical aspects of subgroup analyses in surgical neuro-oncology: A comprehensive review from the PIONEER Consortium

Dr. Jasper K.W. Gerritsen[1,2] MD PhD (Corresponding author), Dr. Philipp Karschnia[3,4] MD, Dr. Jacob S. Young[2] MD, Prof. Martin J. van den Bent[5] MD PhD, Prof. Susan M. Chang[2] MD, Dr. Timothy R. Smith[6] MD PhD MPH, Dr. Brian V. Nahed[7] MD FACS FAANS, Dr. Jordina Rincon-Torroella[8] MD, Dr. Chetan Bettegowda[8] MD PhD, Dr. Nader Sanai[9] MD, Prof. Sandro M. Krieg[10] MD PhD MBA, Prof. Takashi Maruyama[11] MD, Prof. Philippe Schucht[12] MD, Prof. Marike L.D. Broekman[13,14,15] MD PhD, Prof. Joerg-Christian Tonn[3] MD, Prof. Patrick Y. Wen[16] MD, Prof. Steven De Vleeschouwer[17] MD PhD, Prof. Arnaud J.P.E. Vincent[1] MD PhD, Dr. Shawn Hervey-Jumper[2] MD, Prof. Mitchel S. Berger[2] MD FACS FAANS, Prof. Rania A. Mekary[6,18] PhD, Prof. Annette M. Molinaro[2,19] PhD

[1]Department of Neurosurgery, Erasmus Medical Center, Rotterdam, The Netherlands

[2]Department of Neurosurgery, University of California, San Francisco CA, USA

[3]Department of Neurosurgery, Ludwig-Maximilian University Hospital, Munich, Germany

[4]Department of Neurosurgery, University Hospital Erlangen, Germany

[5]Department of Neurology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands

[6]Department of Neurosurgery, Brigham and Women's Hospital, Boston MA, USA

[7]Department of Neurosurgery, Massachusetts General Hospital, Boston MA, USA

[8]Department of Neurosurgery, Johns Hopkins University, Baltimore MD, USA

[9]Department of Neurosurgery, Barrow Neurological Institute, Phoenix AZ, USA

[10]Department of Neurosurgery, University Hospital Heidelberg, Germany

[11]Department of Neurosurgery, Tokyo Women's Medical University, Japan

[12]Department of Neurosurgery, Inselspital University Hospital Bern, Switzerland

[13]Department of Neurosurgery, Haaglanden Medical Center, The Hague, The Netherlands

[14]Department of Neurosurgery, Leiden University Medical Center, Leiden, The Netherlands

[15]Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, The Netherlands

[16]Department of Neuro-Oncology, Dana-Farber Cancer Institute, Boston MA, USA

[17]Department of Neurosurgery, University Hospitals Leuven, Leuven, Belgium

[18]Department of Pharmaceutical Business and Administrative Sciences, School of Pharmacy, MCPHS University, Boston MA, USA

[19]Department of Epidemiology and Biostatistics, University of California, San Francisco CA, USA

[*]Corresponding author

Jasper K.W. Gerritsen, MD PhD

Department of Neurosurgery, Erasmus Medical Center, Rotterdam, The Netherlands

Address: Dr. Molewaterplein 40, 3015 GD Rotterdam

Email: j.gerritsen@erasmusmc.nl

Author contributions

JKWG, SMC, MSB, RAM, and AMM were responsible for the study concept and study design. JKWG, PK, JSY, MJvdB, SMC, SDV, AJPEV, SHJ, MSB, RAM, and AMM drafted the manuscript. All authors revised the manuscript

Conflict of interest statement

Enclosed is a manuscript to be considered for publication in *Neuro-Oncology.* As the corresponding author, I state that the contents of this manuscript have not been published elsewhere (both in whole or in part), nor are they under consideration by another publisher.

I confirm that there are no known conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome.

I confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that all have approved the order of authors listed in the manuscript.

Jasper K.W. Gerritsen

Key points

- Subgroup analyses play an important role in personalized surgical treatment for brain tumor patients
- This paper reviews and summarizes for the first time in a comprehensive manner the most important practical and statistical considerations that are critical for this field

ABSTRACT

Subgroup analyses are essential to generate new hypotheses or to estimate treatment effects in clinically meaningful subgroups of patients. They play an important role in taking the next step towards personalized surgical treatment for brain tumor patients. However, subgroup analyses must be used with consideration and care because they have significant potential risks. Although some recommendations are available on the pearls and pitfalls of these analyses, a comprehensive guide is lacking, especially one focused on surgical neuro-oncology patients. This paper, therefore, reviews and summarizes for the first time comprehensively the practical and statistical considerations that are critical to this field. First, we evaluate the considerations when choosing a study design for surgical neuro-oncology studies and examine those unique to this field. Second, we give an overview of the relevant aspects to interpret subgroup analyses adequately. Third, we discuss the practical and statistical elements necessary to appropriately design and use subgroup analyses. The paper aims to provide an in-depth and complete guide to better understand risk modeling and assist the reader with practical examples of designing, using, and interpreting subgroup analyses.

Keywords

Subgroup analysis; Malignant glioma; Confounding; Multiplicity; Study design

Introduction

Neurosurgeons can choose from an array of surgical modalities and techniques to achieve maximal safe resection of brain tumors. For example, techniques can be used to improve extent of resection (e.g. intraoperative MRI[1,2], intraoperative fluorescence[3], Raman spectroscopy[4,5], or intraoperative ultrasound[6,7]), to prevent neurological deficits (e.g. evoked potentials[8–10], minimally invasive techniques such as LITT (Laser Interstitial Thermal Therapy)[11,12]), or both (awake or asleep brain mapping[13,14]). There is increasing evidence on the role of certain aspects of surgery for neuro-oncological patients, such as extent of resection, supramaximal resection, and intraoperative mapping[15–21]. At the same time, there have been notable improvements in classifying brain tumor patients based on molecular information[22–24]. These developments warrant new, comparative analyses of different surgical treatments in these newly classified patient subgroups.

Traditionally, studies have evaluated the benefits of these surgical treatments in overall cohorts of brain tumor patients[1,3,6]. Although this is a powerful approach to demonstrate the overall benefit of a treatment, it fails to inform the neurosurgeon for which *individual* patient it would be beneficial, and for which it is unlikely to improve outcome[25,26]. This means that the specific effects of these treatments in important subgroups of patients – e.g., based on age, preoperative functioning status, or molecular mutation profile – remain unclear. This, in turn, might lead to the underuse of these modalities due to uncertainty about their benefits, even though these approaches could potentially improve the outcomes of selected patients. For example, recent evidence shows that maximal and supramaximal may be more important for astrocytoma grade 2 patients than oligodendroglioma grade 2 patients.[20]

On the other hand, this might also lead to the overuse of invasive modalities in patients who might have benefitted from avoiding invasive procedures. This might be illustrated by a recent study demonstrating that only patients <65 years old benefitted from supramaximal resection.[16] Subgroup analyses can also provide insight into treatment effects in specific patient populations, informing risk stratification for prospective trials. For example, a clinical risk score for postoperative outcome in glioblastoma patients can decrease prognostic imbalance between study arms[15,26]. Last, post-hoc exploratory subgroup analyses can generate new hypotheses that can be validated in subsequent confirmatory studies[27].

It is essential to carefully design and interpret these subgroup analyses, as they are often exploratory post-hoc analyses with inherent statistical risks[28–30]. This highlights the need for a solid understanding of these analyses to ensure adequate design and interpretation. While certain aspects of subgroup analyses have been previously discussed in the literature, a comprehensive guide with practical pearls and pitfalls – especially one focused on surgical neuro-oncology patients – remains lacking. The fragmentation of existing recommendations can lead to improper use, selective reporting, and misinterpretation, which may harm clinical practice. This review aims to fill that gap by providing, for the first time, a comprehensive guide on planning, analyzing, and interpreting subgroup analyses in the context of surgical neuro-oncology studies. This guide has been put together by the members of the international PIONEER Consortium (Personalized Interventions and Outcomes in Neurosurgical Oncology Research Consortium), an updated name of the previously known ENCRAM Consortium (European and North American Consortium for Intraoperative Mapping in Glioma Patients) that is more representative of the expanded and changing scope of the Consortium. For this manuscript, we focus on subgroup analyses as stratified analyses of a study population of interest. We also evaluate various study designs and types of subgroup analyses, highlighting their advantages, challenges, pearls, and pitfalls in a practical setting.

Advantages and challenges of different study designs in surgical neuro-oncology

Randomized controlled trials (RCTs) are challenging for any surgical intervention, and this also holds true for neuro-oncology. The need for individual and clinical equipoise greatly limits the options for control groups and blinded randomization is nearly impossible[31]. Individual equipoise means that the treating physician is truly uncertain about which treatment might be the best for a specific patient. In contrast, clinical equipoise embodies a similar uncertainty but expanded to the profession[31–33]. As such, only a handful of randomized controlled trials have been published evaluating surgical modalities' benefit in glioblastoma patients[1,3,34]. There are a few actively enrolling RTCs for glioblastoma patients: for instance, the SAFE trial is investigating the potential benefits of awake craniotomy (NCT03861299)[35], the RESURGE trial is investigating re-resection in recurrent tumors (NCT02394626), and the BOLD and G-SUMIT trials are comparing supramaximal resection to other resections (NCT04243005 and NCT04737577, respectively). **Table 1** summarizes key RCTs' design and study protocol and prospective cohort studies on resection for newly diagnosed diffuse gliomas.

It is undisputed that RCTs have major strengths: it cannot be overstated that they are the only study design that can account for both known *and* unknown confounders due to randomization, unlike observational studies that can account for only known confounders. Additionally, they can be combined with blinding techniques (single-blind, double-blind, triple-blind), although complete blinding is not feasible for surgical studies in practice[36]. When preceded by appropriate power analyses, they are the best approach for comparing the overall benefit of different treatment arms with high internal validity[37]. However, several issues make designing and completing RCTs for investigating surgical neuro-oncological modalities particularly difficult[38]. Critically, the lack (or perceived lack) of equipoise often results in highly selective study recruitment, slow accrual and lack

of external generalizability[39–42]. Moreover, many RCTs suffer from slow accrual due to highly selective inclusion of patients with exceptional high-performance statuses. As a result, a considerable proportion of RCTs fail to be completed across neuro-oncology[43,44]. The effect of a lack of patient acceptance to be randomized in the context of brain tumor surgery is another barrier. Critically, RCTs are extremely costly and require significant resources and time to complete relative to observational studies with carefully selected propensity matched groups[45]. Finally, RCTs cannot include updates on the techniques when the trial is active (also called "episodic design"), which can limit the inclusion of new technology or fail to account for a change in equipoise during trial enrollment, potentially making any results outdated before they are published. Given these limitations, RCTs may be best applied to questions for very specific subgroups of patients in which there is a true equipoise of treatment, where broad generalizability is not necessary, and a large multicenter effort can accelerate patient recruitment. RCTs can be followed by post-hoc exploratory, hypothesis-generating subgroup analyses and subsequently, large observational studies with predefined subgroups to assess the risk-benefit ratio for selected patient subgroups. When there is a lack of equipoise and a large heterogeneity in clinical practice, prospective observational studies coupled with matching methods might be better suited to answer the research question. Notably, it has been shown that well-designed observational studies may find a similar magnitude of the associations between exposure and outcome as RCTs[46,47]. **Figure 1** evaluates the methodological strengths and weaknesses of randomized versus observational study designs in surgical neuro-oncology.

Reasons to use subgroup analyses

Generally, there are four main reasons for subgroup analyses, as described comprehensively by Rothwell in 2005[48]. The first reason is possible heterogeneity in treatment effects related to risk. This poses an additional challenge because study arms are often skewed in the risk distribution within the study arms. This means that a disproportionate (low) number of patients is responsible for a

disproportionate (high) risk. As a consequence, the treatment effect might be over-interpreted for low-risk patients[49]. The long-standing debate on biopsy versus resection for elderly glioblastoma patients could be an example[50,51]. Within this subgroup, the location of the tumor causes possible heterogeneity in treatment effect: deep-seated, butterfly, or basal ganglia tumors have vastly different risk profiles than cortical tumors. This difference in risk warrants an additional subgroup analysis to adequately compare resection versus biopsy. The second reason is closely related: instead of differences in risk, there are potential differences in pathophysiology among the patients. Our previous publications might be practical examples[14,19]. We demonstrated that awake mapping and maximal resection only led to longer survival outcomes in patients with MGMT methylated tumors, but not in MGMT unmethylated tumors. We hypothesized that the cytoreduction synergized with the adjuvant treatment only in MGMT methylated patients.

The third reason is an clinically important question regarding the practical application of the treatment, such as differences in risk-benefit ratios across patients due to age categories or surgery timing. For instance, Molinaro and co-authors found that supramaximal resection conferred a survival advantage only among patients 65 years or younger.[16] Young and co-authors found that a longer time between diagnosis and surgery did not negatively impact survival in glioblastoma patients[52]. Importantly, they concluded that “future studies are needed to explore subgroups for whom time-to-surgery may impact clinical outcomes”. The fourth reason consists of the underuse of the treatment in specific groups. In neurosurgical oncology, for example, tumors presumed inoperable by some surgeons or centers may benefit from surgical resection. Krieg et al showed that a safe resection was possible in most of these patients in expert centers.[53] Additionally, Southwell et al showed that the use of intraoperative mapping could even lead to not only a safe but also a maximum resection in a high percentage of these patients[54].

The following paragraphs will elaborate on three methods of subgroup analysis: [1] prospective, confirmatory subgroup analyses (inferential subgroup analysis), [2] prospective, exploratory subgroup analyses (consistency assessment, supportive subgroup analysis), and [3] post-hoc, exploratory subgroup analyses[55,56]. These three types of subgroup analysis[55] **(Figure 2A)** are powerful methods that can be employed for both randomized and observational study designs but need to be designed, performed, and interpreted cautiously **(Figure 4)**.

The three types of subgroup analysis

[1] Prospective, confirmatory subgroup analyses

With prospective subgroup analyses, the researchers define the subgroups they are interested in and their expected direction of treatment effect *before* the study starts (also called pre-defined subgroups). Often, these subgroups are defined by systematically evaluating previous evidence and identifying all potentially relevant factors (in a clinical or biological sense) and their potential interactions. The primary aim of the prospective, confirmatory subgroup analysis is to determine the efficacy of a certain treatment, surgery, or other intervention within a specific patient population. This requires comprehensive methodological planning and correcting for multiple testing to study causal relationships (inferential analyses).

There are two general risks when planning prospective subgroup analyses: type I and type II errors. Type I errors (false positives) are often caused by studying too many different subgroups and carries the risk of overinterpretation of the results[29,57,58]. Type II errors (false negatives) are frequently the consequence of inadequate power because the subgroups were not considered during the sample

size calculation. This leads to underpowered analyses and unreliable results[29]. This is an inherent weakness with post-hoc subgroup analyses (elaborated upon later); however, prospective studies (discussed here) can pre-empt this problem by considering the subgroups when planning the study's sample size.

Trial design is one of the key considerations when planning a prospective subgroup analysis. In a fixed design trial, the study design is documented before the start of the study and no modifications are allowed during the trial. Often, these trials only include the specific subgroup that the researcher is interested in, for example to study the benefit of resection in elderly patients >80 years of age (single population)[51]. On the other hand, studies can include multiple subgroups simultaneously (multi-population). For example, the currently accruing SUPRAMAX study investigates the benefit of supramaximal resection in three predefined subgroups: age, MGMT methylation status, and preoperative KPS status[59].

The alternative to a fixed design is an adaptive trial design. This means that the study design can be modified during the study by using an interim analysis at a time that was decided before the start of the trial.[60,61] It is up to the statistical and clinical committees to decide about the timing of the interim analysis. This is done in order to detect an early trend in the data to decide whether the trial needs to be stopped or modified. The timing of this varies by the research question, study team/statistician, the specific design, the expected recruitment and event rates, length of time to assess the outcome, and other ethical considerations, without inflating the type I error rate or compromising power.[62] Notably, if the analysis is conducted too early, it may lack sufficient power to detect any statistical significance and may erroneously lead to premature stopping. Hence, controlling for increased type I and type II errors while calculating sample size is needed.[63,64] An example is adaptive randomization, in which the group allocation is altered based on the interim analysis and more (or even all) patients are allocated to the treatment group that is performing better ("drop-the-loser")[65]. To our knowledge, neurosurgical trials have yet to use an adaptive

design, but these are already utilized in neuro-oncological trials. For example, Rahman et al used adaptive randomization in their study to compare abemaciclib, neratinib, and CC-115 (with chemotherapy plus radiotherapy as control arm) for their efficacy as adjuvant treatment in newly diagnosed glioblastoma patients[66]. They started the trial with a 1:1:1:1 randomization allocation, which was then automatically adapted during the trial. Based on the interim results, CC-115 had inferior PFS than the other treatment arms. Consequently, the randomization probability for this treatment arm decreased from 25% to 16% and as a result, 12 instead of 18 patients were included in the CC-115 arm.

Adaptive trial designs can be advantageous for subgroup analyses in neurosurgical studies[65]. For example, they can reduce the required sample size with the same power by a drop-the-loser design or re-estimating the sample size based on the interim analysis. Decreasing or eliminating randomization to inferior treatment arms also helps accrual and acceptance of patients to be randomized, which are notable issues for neurosurgical trials[65]. A second advantage is the possibility to combine in one study both exploration and confirmation of 1) a treatment effect (e.g., effect of resection vs biopsy in thalamic glioma patients explored and confirmed) or 2) a covariate threshold (e.g., age cut-off value of 65, 70, and 75 years old for supramaximal explored, selected, and confirmed).

A third advantage of adaptive designs is subgroup enrichment[67,68]. This means the subgroup of patients most likely to benefit from the treatment is selected. Although this is often a two-step approach in which the first study post-hoc identifies this subgroup, and a second study is prospectively "enriched" for this subgroup, these two steps can be combined into one study with an adaptive design. This allows individual patients quicker access to the most effective treatment and decreases unnecessary exposure to less effective treatments. Potential downsides of adaptive designs are the higher costs due to added support requirements from data managers and statistics

staff, and the risk of operation bias because caregivers and surgeons might learn which treatment is better during the trial.

[2] Prospective, exploratory subgroup analyses (consistency assessment)

The second type of subgroup analysis is the consistency assessment. Its primary aim is to examine if a treatment benefit demonstrated in an RCT, can be extrapolated to one or multiple predefined subgroups (treatment effect homogeneity). For instance, when a new surgical strategy is implemented and its effects are measured in the overall cohort, it could be useful to evaluate if the effects are the same in all the relevant subgroups. The RCT by Stummer et al on using 5-ALA in glioblastoma surgery is an example of this type of analysis[3]. They found that the 6-month progression-free survival was higher in the 5-ALA arm than in the conventional surgery arm (41.0% vs. 21.1%, p = 0.003). After stratifying these analyses for the predefined subgroups based on age, they found that this beneficial effect was consistent for younger and older patients (≤55 years vs <55 years). To make the consistency assessment results easily interpretable, making a forest plot can be helpful. A forest plot summarizes the treatment effects across different subgroups in one comprehensive figure. This makes the results across the subgroups easier to digest for the reader and allows for examination if indeed there is treatment effect homogeneity[56]. However, because these results can be underpowered, they must be interpreted cautiously.

[3] Post-hoc, exploratory subgroup analyses

The third type of subgroup analysis is the post hoc, exploratory analysis. Because it is sometimes impossible to predefine all potentially important subgroups, some must be examined post-hoc (after completing the study). This can be true for electronic health records, claims databases, and registry studies. Importantly, even though the subgroups are identified post-hoc, the methods to identify

them can be defined prospectively. Alternatively, relevant patient subgroups can be identified post-hoc with a systematic (disciplined) subgroup search[55,69]. This search aims to find patient subgroups that have a stronger effect of the surgery or treatment than other patient subgroups. It is important to note that post-hoc analyses cannot be used for inferential (causal) conclusions: their findings always need to be validated in a subsequent prospective confirmatory study. Furthermore, they are prone to confounding and should only be used as hypothesis-generating analyses.

The GLIOMAP study is an example of a study that uses post-hoc, exploratory subgroup analyses, in this case to stratify for baseline prognostic imbalance. For this study we compared the survival of awake craniotomy versus asleep resection using Kaplan-Meier curves[14]. For the overall cohort and some of the matched subgroups, we observed “crossing curves”. This simply means that the Kaplan-Meier curves crossed each other and can indicate non-proportional hazards. One way to address this issue is to stratify the analysis for an important prognostic variable, molecular factors (MGMT methylation status and IDH mutation status). After stratifying the analyses for these molecular factors, the survival curves separated without cross over. Other reasons to use post-hoc analyses are to investigate if certain subgroups may benefit from the intervention (when the overall effect was negative), those that may benefit the most from the intervention (when the overall effect was positive), or have a different safety profile.

Therefore, the primary aim of post-hoc exploratory subgroup analyses is to investigate these differences between subgroups (treatment effect heterogeneity) and to generate new hypotheses.

In statistical terms, treatment effect heterogeneity is defined as *interaction:* the treatment effect differs between subgroups because there is an interaction between the treatment and the subgroup covariate[70]. The GLIOMAP study might illustrate this[14]. One of the study's aims was to evaluate if awake craniotomy was predictive of complete resection of the contrast-enhancing tumor. We used

multiple multivariable logistic regression analyses in the overall cohort and subgroups for age, preoperative NIHSS score, and preoperative KPS. The regression analyses showed that awake craniotomy was predictive of complete resection in the overall cohort (OR 1.88, p=0.013), and the subgroups aged <70 (OR 1.86, p=0.028), aged ≥70 (OR 2.31, p=0.012), NIHSS 0-1 (OR 1.97, p=0.038), and KPS 90-100 (OR 2.44, p=0.0080). However, no significant effect was found in the subgroups of NIHSS ≥2 (OR 1.35, p=0.66), and KPS ≤80 (OR 2.19, p=0.18). We then used an interaction term between the treatment (awake craniotomy) and the subgroup covariate (age, NIHSS, or KPS) to study treatment effect heterogeneity among the subgroups. The interaction term is coded as a "*" between the treatment and subgroup covariate. The null hypothesis of the interaction is that test the treatment effect is the same for different covariate values, for example, younger or older patients. In our analysis, the interaction terms were nonsignificant for awake*age (p=0.77), awake*NIHSS (p=0.50), and awake*KPS (p=0.47). This indicates no statistical difference in the treatment effect of awake craniotomy on complete resection across these subgroups. In other words, the association between awake craniotomy and complete resection is similar for patients irrespective of age, preoperative NIHSS score, or preoperative KPS. The fact that we observed a nonsignificant effect in two subgroups (NIHSS ≥2 and KPS ≤80) suggests that the analyses for these subgroups might have been underpowered. This may have led to a type II error (false negative finding due to insufficient power). A common mistake is to claim that there is treatment effect heterogeneity because the association (odds ratio, hazard ratio) between the treatment and the outcome is different within each of the levels of the baseline variable. For example, testing the association between awake craniotomy and complete resection in younger patients and then separately in older patients does not answer if age influences this association. Only the interaction test (e.g., awake*age) determines if the effect in younger patients is different than in older patients. The theoretical reasons for this have been explained in previous papers[30,71].

Interpreting post-hoc subgroup analyses can be a challenge. It is often helpful to evaluate if certain factors have been prespecified. This provides insight and may benefit the credibility of the subgroup findings. Examples of factors that can be prespecified are the rationale for the subgroup analysis (such as clinical or biological plausibility), the effect size and direction, covariate levels and cutoff values, the statistical methods that will be used, or the endpoint that will be studied. It is important to keep in mind that plausibility is not always reliable: most subgroups are comprised of well-known predictive or prognostic factors, and it is therefore very common for these subgroups to be considered "plausible". One way to address this is to prospectively discuss a range of factors that would be by their underlying mechanism the most probable factors, along with their direction of effect.

Multiplicity correction

Post-hoc analyses are often done with several subgroups. In an extreme form, this can lead to "data dredging" or "fishing expeditions". This means that a multiple subgroups is analyzed with the deliberate goal of finding a subgroup that has a significant outcome. This is problematic, because analyzing multiple of subgroups (multiple testing) makes these analyses prone to false-positive findings (type I error). The risk of false-positive findings increases with the number of tests that are performed (read: the number of hypotheses that are tested, the number of subgroups that are studied)[56,58]. This can be explained by the fact that typically, each analysis has an alpha of 5% associated with it (the risk to find a false-positive finding). If three analyses are performed with an alpha of 5%, the overall risk of a false-positive finding increases to 15%. This underlines the fact that for a correct interpretation of post-hoc analyses, it is vital to know how many tests have been performed for the subgroup analyses. The elevated risk of false-positive findings due to multiple testing can be mitigated by correcting for doing multiple tests (multiplicity correction). For multiplicity correction, three methods can be used: lowering the significance level α, increasing the

p-value, or widening the confidence interval of the individual tests. The goal of all three methods is to make it more difficult to reject the null hypothesis for an individual test (to get a significant result). This will decrease the risk of false-positive results. We will focus on the first method (lowering the significance level), because this is the most frequently used in daily practice. The approaches to do this are summarized in **Figure 2B.**

The most commonly used multiplicity correction method is the Bonferroni method. This method is relatively simple: it divides the significance level α for the individual tests by the total number of tests performed. For example, when five hypotheses are tested in a study with an overall significance level of 0.05, the Bonferroni method lowers the significance level for the individual tests to 0.01. This ensures that testing five hypotheses with a significance level of 0.01 or testing one hypothesis with a significance level of 0.05 will have the same probability of finding false-positive results. The Bonferroni method is a so-called *single-step* multiple testing procedure. This means that with one single adjustment to the significance level all the hypotheses of the study can be tested simultaneously (e.g., all hypotheses are significant when $p < 0.01$, instead of $p < 0.05$). The limitation of this method is that it decreases the power of the individual tests significantly as the number of tests increases. In other words, it can make the significance level α too conservative (e.g., 0.01 instead of 0.05 can be too strict statistically in some instances). Alternatives that carry greater power are *stepwise* procedures[57,58,72], which test the hypotheses in a particular order. This means that after each test, the significance level is adjusted based on the "data" of the first test, and with this new significance level, the next hypothesis is tested. Because the adjustment of the significance level of the test is based on the "data" of the previous test, these stepwise procedures are called *data-driven.* Examples of data-driven procedures are Holm's method, Hochberg's method, and Hommel's method[73–75]. The alternative to data-driven procedures is *pre-specified* procedures: the adjustment of the significance level is not based on the previous test but is pre-specified. Examples are the *fixed-sequence procedure,* the *fallback procedure,* and the *chain procedure*[57,58] **(Figure 2B).** The theory behind these methods falls outside of the scope of this paper.

These methods to correct for multiplicity (single-step or stepwise) are examples of *sequential* approaches. Their limitations are outlined in Figure 2B. An alternative to these methods is the *alpha spending function*, which offers a flexible approach by "spending" portions of the the overall type I error rate (alpha, usually 0.05) across multiple tests over time, such as the O'Brien-Fleming and Pocock boundaries.[76–79] This approach can be useful particularly when conducting interim analyses as part of an adaptive randomized trial. For example, rather than requiring a predetermined (fixed) plan for carrying out the interim analyses, it allows for flexibility in the timing and number of interim analyses. This makes it easier to deal with unplanned deviations in the trial or extra analyses without inflating the type I error. Furthermore, this approach can adapt to the trial data by adjusting the spending rate of alpha based on the results of previous analyses. For example, if an early analysis looks promising, a smaller portion of alpha can be spent early, leaving more room for later tests. This may result in higher power (lower type II error caused by an underpowered analysis), especially for early or unexpected interim analyses, by avoiding the sometimes overly conservative nature of traditional sequential approaches (such as the Bonferroni).

Application examples of subgroup analyses in neuro-oncology

The choice for a certain method of subgroup analysis is dependent on the hypothesis that the researchers aim to test or the research question that they try to answer. This will influence the choice to confirm, explore, or discover specific subgroups. As described earlier, two common reasons to use subgroup analyses are to study the association between treatment and outcome in an already known subgroup (etiological analysis), or to discover important subgroups.

We will give two practical examples that illustrate these reasons along with the pearls and pitfalls on their indications, analysis, and interpretation. These are summarized in **Figure 3.**

The study by Molinaro et al published in 2020 illustrates the first reason[16]. They investigated the value of minimizing the residual non-contrast-enhancing tumor volume in glioblastoma patients. The objective of this study was to assess the association of minimizing residual volume and survival while considering molecular and clinical factors. To this end, a variation of recursive partitioning analysis (RPA, a form of classification and regression trees (CART)) called partDSA was used to discover four distinct patient subgroups based on molecular and clinical data[16,80]. Each of these subgroups corresponded to different survival outcomes. These CART techniques are often displayed as the branches of a tree because they split the data into several smaller patient subgroups, hence the name "classification tree". The tree starts with a root node at the top of the tree. This root node includes all the available training data. From this root node, a split in the tree occurs and is based the variable which increases the homogeneity of outcome (here, survival) in the resulting two new nodes (in this case: temozolomide after surgery: yes or no). These nodes can either be terminal or nonterminal nodes. Terminal nodes are nodes after which that specific branch of the tree stops: there are no more splits after this node. These nodes represent a distinct subgroup in the data. In this study, the first terminal node occurred after the first split: patients who had not undergone temozolomide after surgery formed a separate subgroup. Nonterminal nodes are nodes after which the branching of the tree continues. In this practical example, patients who had undergone temozolomide after surgery were split again for IDH status, age at diagnosis, and residual non-contrast-enhancing (NCE) tumor after surgery. In the end, all the cases included in the learning set end up at one specific terminal node and the partition – the set of all terminal nodes – is completed. In the Molinaro study[16], this resulted in four distinct subgroups of patients, based on four different nonterminal nodes: patients who did not receive temozolomide after surgery (subgroup 1), patients who received temozolomide, had an IDH wildtype tumor, and were older than 65 years old (subgroup 2), patients who received temozolomide, had an IDH wildtype tumor, were 65 years or younger, and had a residual NCE tumor of more than 5.4 ml (subgroup 3), and patients who received

temozolomide and had an IDH mutated tumor OR those who received temozolomide, had an IDH wildtype tumor, were 65 years or younger, and had a residual NCE tumor of less than 5.4 ml (subgroup 4). For this latter subgroup, the overall survival was similar between the subset of IDHwt tumors and the IDHmt tumors. Note that the partDSA algorithm differs from CART in that it inherently combines subgroups to maximize the homogeneity within a terminal node, rather than separate subgroups with similar outcomes. This is illustrated in our example study by the fact that subgroup 4 was formed by two distinct types of patients: IDHmt patients, and younger IDHwt patients with supramaximal resection.

There are a few challenges with RPA[81]. The first is the risk of overfitting, i.e., fitting the data not only on the actual signal within the data but also on the noise. If the data are "overfit": they only fit on the training data and not on other data, such as testing data or validation data (often because the model captures noise rather than the underlying pattern, which may lead to poor generalizability). To counteract overfitting, it is best not to have too many branches (subgroups). Therefore, like a tree, it can be pruned: finding a subtree of the "first draft" of the tree that is the best at predicting the outcome and is relatively protected to overfitting. However, the important part is deciding when to stop. There are two ways to test this: using an independent test set, or cross-validation (preferred for smaller datasets). Both these tests work by testing different potential subtrees for their potential to reliably predict the subgroups in data other than the original training set that the tree was built on.

Random forest is a collection of individual trees which has excellent predictive abilities[82]. As the name implies, this method works by building a large number of trees using random subsets of the data (e.g., with bootstrapping or bagging methods[83–85]). The final prediction uses results from all the different trees (the individual trees should not be pruned like in RPA because otherwise , there will be a loss of information). Random forests are more stable than "single trees" as developed by RPA and are less susceptible to prediction errors. The choice between random forest versus "single-tree"

RPA depends on a number of factors[86]. One major advantage of single-tree RPA is the interpretability of the tree structure. Each branch represents a clear decision rule and the final leaves represent distinct subgroups of patients. This means that single-tree RPA is useful to identify a single set of subgroups, assuming that the structure of the subgroups is simple and interactions are limited. However, single-tree RPA is prone to overfitting the data, especially when the tree grows too large. Also, it uses a "greedy algorithm", which means that it splits each node without considering the overall structure. This can lead to suboptimal subgroup identification if the most informative splits are not chosen early in the tree. Third, it can be unstable, meaning that small changes in the data can lead to significant changes in the tree structure.[87] Random forests, however, can solve some of these problems of single-tree RPA. multiple possible splits and interactions between variables, which handle complex nonlinear relationships and interactions between subgroups, "rank" the variables based on importance, and can visualize the relationship between variables. Because random forests combine the decisions and predictions of a large number of trees, they are more stable, less prone to overfitting, and their results are more generalizable than single trees.[88] However, the downside of random forests is that they are more difficult to interpret than single trees (the overall model is a "black box"). Thus, the clear interpretability and identification of subgroups with single-tree RPA means that this method is ideal to clinical decision making. Random forests, on the other hand, are particularly useful when evaluating which variables are most important in predicting the outcome (especially when there is inadequate clinical knowledge to select the variables).

The advantages and disadvantages of CART techniques are shown in **Figure 3** along with their indications, organization, presentation, and interpretation**.** Other non-parametric methods exist to discover subgroups and have their own advantages and disadvantages. For example, some are better at handling continuous variables (MARS, multivariate adaptive regression splines[89]), more useful when the relationship between the predictor variable and outcome is not linear (GAM,

generalized additive models[90]), or instead of these supervised machine learning algorithms, are unsupervised (clustering[91]).

The second common reason to use subgroup analyses is the etiological analysis to study the association between a treatment and outcome. Examples of statistical methods that can be used for these analyses are linear regression models (for continuous outcomes), logistic regression models (for binary outcomes), and Cox proportional-hazards regression models (for censored outcomes). Rather than discovering subgroups with "data mining" methods such as CART, MARS, and GAM, subgroups that already have been defined by the researcher are analyzed.

It is vital that the homogeneity within and between subgroups is maximized before moving on to the actual data analysis. Homogeneity *within* the subgroup can be maximized by using stratification methods. In randomized controlled trials, this is called stratified randomization[92]. For example, Stummer et al randomly assigned patients to the 5-ALA and control arm, while considering the stratification factors age (≤55 years vs >55 years), KPS (70-80 vs >80), eloquence (non-eloquent vs eloquent), and study surgeon[3]. This means that the randomization method will make sure that the patients in both trial arms are evenly balanced for these factors. Observational studies can mimic the stratification randomization design by first stratifying patients based on predefined factors (e.g., age and functional status) before performing the analysis. An example can be observed in the GLIOMAP study and its supplementary analysis in which we stratified patients for age, preoperative neurological status, and preoperative functional status before moving on to the actual analysis[14,19].

Homogeneity *between* subgroups can be improved by using matching methods. The most commonly used method is propensity score matching[93]. This means that every patient gets assigned a propensity score. In the GLIOMAP study, this propensity score corresponded to the probability of having a certain exposure (awake craniotomy) based on their individual set of covariates (e.g., age,

functional status, etc)[14]. Patients from the group with the exposure (awake craniotomy) were matched with the patients from the group without the exposure (asleep resection) based on their propensity score (their individual set of covariates). This method acts as a countermeasure for confounding bias, a bias caused by the prognostic imbalance in the patient's covariates (e.g., if awake brain surgery is found to be associated with longer survival, but in particular younger patients undergo awake brain surgery: it is now unknown if the longer survival should be attributed to the intervention [awake brain surgery] or the confounding covariate [age]). It is good practice to show in appendix tables the outcome of the matching procedure, i.e., the propensity scores for the unmatched and matched cohorts. It should be noted that propensity score matching has two major drawbacks: it only addresses known confounders, and it affects the statistical power because the available sample size is decreased after matching.

After maximizing homogeneity within and between subgroups, the second step is regression modelling. Regression models can be used for predictive or etiological analyses. Predictive analyses look for potential predictors in the dataset, e.g., which baseline factors are predictive for a certain postoperative outcome. These analyses should be performed on unmatched cohorts to allow for optimum identification of the predictors in the "raw" data. Etiological analyses (also called inferential analyses) aim to determine causal relationships between covariates and outcomes, e.g., which baseline factors are associated with a certain postoperative outcome. Therefore, these analyses should be performed on the matched cohorts to adequately counteract confounding bias. Further adjustment might be needed if certain variables were unstable in the matching procedure: these can be included in the subsequent regression to adjust for them adequately (sensitivity analysis). For example, in the GLIOMAP study the molecular factors (MGMT methylation status, IDH mutation status) proved to be unstable in the matching procedure due to missing data[14]. We therefore included these factors in the regression analyses as a sensitivity analysis to adjust for them. It is considered good practice to present the regression analyses without and without these factors for reliable interpretation of the results.

Conclusions

The surgical and nonsurgical treatment options for neuro-oncological patients are rapidly expanding. The neuro-oncological community is working together on a number of promising studies to try to improve the care for individual brain tumor patients. In this current era of personalized neuro-oncology, subgroup analyses are necessary to determine the optimum surgical treatment strategy for individual patients, especially given the fact that truly randomized studies, even unblinded, are challenging to conduct when dealing with surgical questions. However, subgroup analyses must be used with consideration and care because they have important potential risks. We aimed to give the reader for the first time a complete overview of the practical and statistical considerations regarding these analyses in neuro-oncology studies. We examined the pearls and pitfalls and evaluated the special considerations that are unique to surgical neuro-oncological studies. This paper can be used for other medical fields as well, even though the examples given were pertinent to neuro-oncology. The goal of this comprehensive guide was to assist with practical examples how to design, use, and interpret subgroup analyses.

References

1. Senft C, Bink A, Franz K, Vatter H, Gasser T, Seifert V. Intraoperative MRI guidance and extent of resection in glioma surgery: a randomised, controlled trial. *The Lancet Oncology*. 2011;12:997-1003. doi:10.1016/S1470

2. Roder C, Stummer W, Coburger J, et al. Intraoperative MRI-Guided Resection Is Not Superior to 5-Aminolevulinic Acid Guidance in Newly Diagnosed Glioblastoma: A Prospective Controlled Multicenter Clinical Trial. *J Clin Oncol*. 2023;41(36):5512-5523. doi:10.1200/JCO.22.01862

3. Stummer W, Pichlmeier U, Meinel T, Wiestler D, Zanella F, Reulen HJ. Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase III trial. *The Lancet Oncology*. 2006;7:392-401. doi:10.1016/S1470-2045(06)

4. Herta J, Cho A, Roetzer-Pejrimovsky T, et al. Optimizing maximum resection of glioblastoma: Raman spectroscopy versus 5-aminolevulinic acid. *Journal of Neurosurgery*. 2023;139(2):334-343. doi:10.3171/2022.11.JNS22693

5. Pekmezci M, Morshed RA, Chunduru P, et al. Detection of glioma infiltration at the tumor margin using quantitative stimulated Raman scattering histology. *Sci Rep*. 2021;11(1):12162. doi:10.1038/s41598-021-91648-8

6. Incekara F, Smits M, Dirven L, et al. Intraoperative B-Mode Ultrasound Guided Surgery and the Extent of Glioblastoma Resection: A Randomized Controlled Trial. *Front Oncol*. 2021;11:649797. doi:10.3389/fonc.2021.649797

7. Mahboob S, McPhillips R, Qiu Z, et al. Intraoperative Ultrasound-Guided Resection of Gliomas: A Meta-Analysis and Review of the Literature. *World Neurosurgery*. 2016;92:255-263. doi:10.1016/j.wneu.2016.05.007

8. Krieg SM, Shiban E, Droese D, et al. Predictive Value and Safety of Intraoperative Neurophysiological Monitoring With Motor Evoked Potentials in Glioma Surgery. *Neurosurgery*. 2012;70(5):1060-1071. doi:10.1227/NEU.0b013e31823f5ade

9. Seidel K, Beck J, Stieglitz L, Schucht P, Raabe A. The warning-sign hierarchy between quantitative subcortical motor mapping and continuous motor evoked potential monitoring during resection of supratentorial brain tumors: Clinical article. *JNS*. 2013;118(2):287-296. doi:10.3171/2012.10.JNS12895

10. Gogos AJ, Young JS, Morshed RA, et al. Triple motor mapping: transcranial, bipolar, and monopolar mapping for supratentorial glioma resection adjacent to motor pathways. *J Neurosurg*. 2020;134(6):1728-1737. doi:10.3171/2020.3.JNS193434

11. Groot JFD, Kim AH, Prabhu S, et al. Efficacy of laser interstitial thermal therapy (LITT) for newly diagnosed and recurrent IDH wild-type glioblastoma. *Neuro-Oncology Advances*. 2022;4(1). doi:10.1093/noajnl/vdac040

12. Viozzi I, Overduin CG, Rijpma A, Rovers MM, Laan M ter. MR-guided LITT therapy in patients with primary irresectable glioblastoma: a prospective, controlled pilot study. *Journal of Neuro-Oncology*. 2023;164(2):405-412. doi:10.1007/s11060-023-04371-x

13. Raabe A, Beck J, Schucht P, Seidel K. Continuous dynamic mapping of the corticospinal tract during surgery of motor eloquent brain tumors: Evaluation of a new method: Clinical article. *Journal of Neurosurgery*. 2014;120(5):1015-1024. doi:10.3171/2014.1.JNS13909

14. Gerritsen JKW, Zwarthoed RH, Kilgallon JL, et al. Effect of awake craniotomy in glioblastoma in eloquent areas (GLIOMAP): a propensity score-matched analysis of an international, multicentre, cohort study. *The Lancet Oncology*. Published online May 2022. doi:10.1016/S1470-2045(22)00213-3

15. Karschnia P, Gerritsen JKW, Teske N, et al. The oncological role of resection in newly diagnosed diffuse adult-type glioma defined by the WHO 2021 classification: a Review by the RANO resect group. *The Lancet Oncology*. 2024;25(9):e404-e419. doi:10.1016/S1470-2045(24)00130-X

16. Molinaro AM, Hervey-Jumper S, Morshed RA, et al. Association of Maximal Extent of Resection of Contrast-Enhanced and Non-Contrast-Enhanced Tumor with Survival Within Molecular Subgroups of Patients with Newly Diagnosed Glioblastoma. *JAMA Oncology*. 2020;6(4):495-503. doi:10.1001/jamaoncol.2019.6143

17. Brown TJ, Brennan MC, Li M, et al. Association of the extent of resection with survival in glioblastoma: A systematic review and meta-analysis. *JAMA Oncology*. 2016;2(11):1460-1469. doi:10.1001/jamaoncol.2016.1373

18. Karschnia P, Young JS, Dono A, et al. Prognostic validation of a new classification system for extent of resection in glioblastoma: a report of the RANO resect group. *Neuro-Oncology*. Published online August 2022. doi:10.1093/neuonc/noac193

19. Gerritsen JKW, Zwarthoed RH, Kilgallon JL, et al. Impact of maximal extent of resection on postoperative deficits, patient functioning, and survival within clinically important glioblastoma subgroups. *Neuro-Oncology*. 2023;25(5):958-972. doi:10.1093/neuonc/noac255

20. Hervey-Jumper SL, Zhang Y, Phillips JJ, et al. Interactive Effects of Molecular, Therapeutic, and Patient Factors on Outcome of Diffuse Low-Grade Glioma. *Journal of Clinical Oncology*. 2023;41(11):2029-2042. doi:10.1200/JCO.21.02929

21. Wijnenga MMJ, French PJ, Dubbink HJ, et al. The impact of surgery in molecularly defined low-grade glioma: an integrated clinical, radiological, and molecular analysis. *Neuro-Oncology*. 2018;20(1):103-112. doi:10.1093/neuonc/nox176

22. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*. 2021;23(8):1231-1251. doi:10.1093/neuonc/noab106

23. Nassiri F, Liu J, Patil V, et al. A clinically applicable integrative molecular classification of meningiomas. *Nature*. 2021;597(7874):119-125. doi:10.1038/s41586-021-03850-3

24. Maas SLN, Stichel D, Hielscher T, et al. Integrated Molecular-Morphologic Meningioma Classification: A Multicenter Retrospective Analysis, Retrospectively and Prospectively Validated. *J Clin Oncol*. 2021;39:3839-3852. doi:10.1200/JCO.21

25. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet*. 1995;345(8965):1616-1619. doi:10.1016/s0140-6736(95)90120-5

26. Kent DM, Hayward RA. Limitations of Applying Summary Results of Clinical Trials to Individual Patients: The Need for Risk Stratification. *JAMA*. 2007;298(10):1209. doi:10.1001/jama.298.10.1209

27. Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. *BMJ*. Published online November 4, 2015:h5651. doi:10.1136/bmj.h5651

28. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in Medicine-Reporting of Subgroup Analyses in Clinical Trials. *NEJM*. 2007;357. www.nejm.org

29. Wang X, Piantadosi S, Le-Rademacher J, Mandrekar SJ. Statistical Considerations for Subgroup Analyses. *Journal of Thoracic Oncology*. 2021;16(3):375-380. doi:10.1016/j.jtho.2020.12.008

30. Groenwold RHH, Dekkers OM. Subgroup analyses in clinical research: too tempting? *European Journal of Endocrinology*. 2023;189(2):E1-E3. doi:10.1093/ejendo/lvad089

31. Freedman B. Equipoise and the Ethics of Clinical Research. *N Engl J Med*. 1987;317(3):141-145. doi:10.1056/NEJM198707163170304

32. Alderson P. Equipoise as a means of managing uncertainty: personal, communal and proxy. *J Med Ethics*. 1996;22(3):135-139. doi:10.1136/jme.22.3.135

33. Weijer C. For and against: Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial FOR AGAINST. *BMJ*. 2000;321(7263):756-758. doi:10.1136/bmj.321.7263.756

34. Picart T, Pallud J, Berthiller J, et al. Use of 5-ALA fluorescence–guided surgery versus white-light conventional microsurgery for the resection of newly diagnosed glioblastomas (RESECT study): a French multicenter randomized phase III study. *Journal of Neurosurgery*. 2024;140(4):987-1000. doi:10.3171/2023.7.JNS231170

35. Gerritsen JKW, Klimek M, Dirven CMF, et al. The SAFE-trial: Safe surgery for glioblastoma multiforme: Awake craniotomy versus surgery under general anesthesia. Study protocol for a multicenter prospective randomized controlled trial. *Contemporary Clinical Trials*. 2020;88. doi:10.1016/j.cct.2019.105876

36. Betensky RA. Don't Be Blinded by the Blinding. *NEJM Evidence*. 2022;1(3). doi:10.1056/EVIDe2100063

37. Vandenbroucke JP. When are observational studies as credible as randomised trials? *The Lancet*. 2004;363(9422):1728-1731. doi:10.1016/S0140-6736(04)16261-2

38. Takroni R, Sharma S, Reddy K, et al. Randomized controlled trials in neurosurgery. *Surgical Neurology International*. 2022;13:379. doi:10.25259/SNI_1032_2021

39. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet*. 2005;365(9453):82-93. doi:10.1016/S0140-6736(04)17670-8

40. Chavez-MacGregor M, Giordano SH. Randomized Clinical Trials and Observational Studies: Is There a Battle? *JCO*. 2016;34(8):772-773. doi:10.1200/JCO.2015.64.7487

41. Rothwell PM. Factors That Can Affect the External Validity of Randomised Controlled Trials. *PLOS Clin Trial*. 2006;1(1):e9. doi:10.1371/journal.pctr.0010009

42. Parsons NR, Kulikov Y, Girling A, Griffin D. A statistical framework for quantifying clinical equipoise for individual cases during randomized controlled surgical trials. *Trials*. 2011;12(1):258. doi:10.1186/1745-6215-12-258

43. Smith EJ, Naik A, Goel M, et al. Adult Neuro-Oncology Trials in the United States over Five Decades: Analysis of Trials Completion Rate to Guide the Path Forward. *Neuro-Oncology Advances*. Published online January 10, 2024:vdad169. doi:10.1093/noajnl/vdad169

44. Hutchins LF, Unger JM, Crowley JJ, Coltman CA, Albain KS. Underrepresentation of Patients 65 Years of Age or Older in Cancer-Treatment Trials. *N Engl J Med*. 1999;341(27):2061-2067. doi:10.1056/NEJM199912303412706

45. Steensma DP, Kantarjian HM. Impact of Cancer Research Bureaucracy on Innovation, Costs, and Patient Care. *JCO*. 2014;32(5):376-378. doi:10.1200/JCO.2013.54.2548

46. Concato J, Shah N, Horwitz RI. RANDOMIZED, CONTROLLED TRIALS, OBSERVATIONAL STUDIES, AND THE HIERARCHY OF RESEARCH DESIGNS. Published online 2006.

47. Benson K. A Comparison of Observational Studies and Randomized, Controlled Trials. *The New England Journal of Medicine*. Published online 2000.

48. Rothwell PM. Treating individuals 2: Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186. doi:10.1016/S0140-6736(05)17709-5

49. Ioannidis JPA, Lau J. The impact of high-risk patients on the results of clinical trials. *Journal of Clinical Epidemiology*. 1997;50(10):1089-1098. doi:10.1016/S0895-4356(97)00149-2

50. Almenawer SA, Badhiwala JH, Alhazzani W, et al. Biopsy versus partial versus gross total resection in older patients with high-grade glioma: a systematic review and meta-analysis. *Neuro-Oncology*. 2015;17(6):868-881. doi:10.1093/neuonc/nou349

51. Niare M, Desrousseaux J, Cavandoli C, et al. Outcome of glioblastoma resection in patients 80 years of age and older. *Acta Neurochirurgica*. 2022;164:373-383. doi:10.1007/s00701-021-04776-5/Published

52. Young JS, Al-Adli NN, Muster R, et al. Does waiting for surgery matter? How time from diagnostic MRI to resection affects outcomes in newly diagnosed glioblastoma. *Journal of Neurosurgery*. Published online June 2023:1-14. doi:10.3171/2023.5.jns23388

53. Krieg SM, Schnurbus L, Shiban E, et al. Surgery of highly eloquent gliomas primarily assessed as non-resectable: Risks and benefits in a cohort study. *BMC Cancer*. 2013;13. doi:10.1186/1471-2407-13-51

54. Southwell DG, Birk HS, Han SJ, Li J, Sall JW, Berger MS. Resection of gliomas deemed inoperable by neurosurgeons based on preoperative imaging studies. *Journal of Neurosurgery*. 2018;129(3):567-575. doi:10.3171/2017.5.JNS17166

55. 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. doi:10.1080/10543406.2015.1092033

56. Alosh M, Huque MF, Bretz F, D'Agostino RB. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Statistics in Medicine*. 2017;36(8):1334-1360. doi:10.1002/sim.7167

57. Dmitrienko A, D'Agostino R. Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*. 2013;32(29):5172-5218. doi:10.1002/sim.5990

58. Dmitrienko A, Millen B, Lipkovich I. Multiplicity considerations in subgroup analysis. *Statistics in Medicine*. 2017;36(28):4446-4454. doi:10.1002/sim.7416

59. Gerritsen JKW, Young JS, Chang SM, et al. SUPRAMAX-study: supramaximal resection versus maximal resection for glioblastoma patients: study protocol for an international multicentre prospective cohort study (ENCRAM 2201). *BMJ open*. 2024;14(4):e082274. doi:10.1136/bmjopen-2023-082274

60. Brannath W, Zuber E, Branson M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*. 2009;28(10):1445-1463. doi:10.1002/sim.3559

61. Bhatt DL, Mehta C. Adaptive Designs for Clinical Trials. Drazen JM, Harrington DP, McMurray JJV, Ware JH, Woodcock J, eds. *N Engl J Med*. 2016;375(1):65-74. doi:10.1056/NEJMra1510061

62. Chow SC, Chang M, Pong A. Statistical consideration of adaptive methods in clinical development. *J Biopharm Stat*. 2005;15(4):575-591. doi:10.1081/BIP-200062277

63. Pocock SJ. *Clinical Trials: A Practical Approach*. Wiley; 1983.

64. DeMets DL, Furberg C, Friedman LM. *Data Monitoring in Clinical Trials: A Case Studies Approach*. Springer; 2006.

65. Mukherjee A, Grayling MJ, Wason JMS. Adaptive Designs: Benefits and Cautions for Neurosurgery Trials. *World Neurosurgery*. 2022;161:316-322. doi:10.1016/j.wneu.2021.07.061

66. Rahman R, Trippa L, Lee EQ, et al. Inaugural Results of the Individualized Screening Trial of Innovative Glioblastoma Therapy: A Phase II Platform Trial for Newly Diagnosed Glioblastoma Using Bayesian Adaptive Randomization. *Journal of Clinical Oncology*. 2023;41(36):5524-5535. doi:10.1200/JCO.23.00493

67. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. *Statistics in Medicine*. 2009;28(1):3-23. doi:10.1002/sim.3461

68. Temple R. Enrichment of Clinical Study Populations. *Clin Pharmacol Ther*. 2010;88(6):774-778. doi:10.1038/clpt.2010.233

69. Lipkovich I, Dmitrienko A, B. R. Tutorial in biostatistics: data- driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*. 2017;36(1):136-196. doi:10.1002/sim.7064

70. Brankovic M, Kardys I, Steyerberg EW, et al. Understanding of interaction (subgroup) analysis in clinical trials. *Eur J Clin Investigation*. 2019;49(8):e13145. doi:10.1111/eci.13145

71. Barraclough H, Govindan R. Biostatistics Primer: What a Clinician Ought to Know: Subgroup Analyses. *Journal of Thoracic Oncology*. 2010;5(5):741-746. doi:10.1097/JTO.0b013e3181d9009e

72. Leon AC, Heo M. A Comparison of Multiplicity Adjustment Strategies for Correlated Binary Endpoints. *Journal of Biopharmaceutical Statistics*. 2005;15(5):839-855. doi:10.1081/BIP-200067922

73. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*. 1979;6(2):65-70.

74. Hochberg, Yosef. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75(4):800-802.

75. Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*. 1988;75(2):383-386.

76. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64(2):191-199.

77. Lan GK, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659-663.

78. Jennison C, Turnbull BW. Group-Sequential Analysis Incorporating Covariate Information. *J Am Statist Assoc*. 1997;92(440):1330-1341.

79. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35(3):549-556.

80. Lostritto K, Strawderman RL, Molinaro AM. A Partitioning Deletion/Substitution/Addition Algorithm for Creating Survival Risk Groups. *Biometrics*. 2012;68(4):1146-1156. doi:10.1111/j.1541-0420.2012.01756.x

81. Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: Advantages and disadvantages of recursive partitioning analysis. *Journal of Chronic Diseases*. 1984;37(9-10):721-731. doi:10.1016/0021-9681(84)90041-9

82. Rigatti SJ. Random Forest. *Journal of Insurance Medicine*. 2017;47(1):31-39. doi:10.17849/insm-47-01-31-39.1

83. Henderson AR. The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clinica Chimica Acta*. 2005;359(1-2):1-26. doi:10.1016/j.cccn.2005.04.002

84. Baser O, Crown WH, Pollicino C. Guidelines for selecting among different types of bootstraps. *Current Medical Research and Opinion*. 2006;22(4):799-808. doi:10.1185/030079906X100230

85. Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*. 2003;19(9):1090-1099. doi:10.1093/bioinformatics/btg038

86. Breiman, Leo. Random Forests. *Machine Learning*. 2001;45(1):5-32.

87. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup Analysis via Recursive Partitioning. *J Mach Learn Res*. Published online 2009. doi:10.2139/ssrn.1341380

88. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. 2012;99(6):323-329. doi:10.1016/j.ygeno.2012.04.003

89. Friedman JH, Roosen CB. An introduction to multivariate adaptive regression splines. *Stat Methods Med Res*. 1995;4(3):197-217. doi:10.1177/096228029500400303

90. Hastie T, Tibshirani R. Generalized additive models for medical research. *Stat Methods Med Res*. 1995;4(3):187-196. doi:10.1177/096228029500400302

91. Steinley Douglas. K- means clustering: A half- century synthesis. *Brit J Math & Statis*. 2006;59(1):1-34. doi:10.1348/000711005X48266

92. Green SB, Byar DP. The effect of stratified randomization on size and power of statistical tests in clinical trials. *Journal of Chronic Diseases*. 1978;31(6-7):445-454. doi:10.1016/0021-9681(78)90008-5

93. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786

Figure 1: Comparison of study designs in surgical neuro-oncology

Legend: #No equipoise which causes high heterogeneity in indication setting and procedure

NB: green color indicates favorable criteria (n = 11 for observational, n = 8 for RCT), while orange-red color indicates unfavorable criteria (n = 4 for observational, n = 7 for RCT).

Figure 2A: Subgroup analysis methods

Figure 2b: Multiplicity correction methods

Figure 3: Pearls and pitfalls for two common subgroup analysis methods in surgical neuro-oncology

Figure 4: Recommendations for subbroup analyses in surgical neuro-oncology studies

Table 1: Design and protocol of prospective key studies on resection for newly diagnosed gliomas.

Pubmed was searched for prospective cohorts of patients with newly diagnosed glioblastoma, astrocytoma grade 2-4, or oligodendroglioma grade 2-3. Only papers published after 2005 (following the introduction of the EORTC 26981/22981-protocol for concomitant radiochemotherapy in glioblastoma) with data on extent of resection were included. Database closure was January 1, 2024. Study names, design, key intervention, key findings, relevant subgroups analyzed, and pre-specified statistics are indicated. 95%-confidence intervals as measures of uncertainty are given within 'key results'.

Abbreviations: 5-ALA - 5-aminolevulinic acid, *CE* – contrast enhancing, *CI* – confidence interval, *EOR* - extent of resection, *FLAIR* - fluid-attenuated inversion recovery, *HR* - hazard ratio, *ioMRI* - intraoperative magnetic resonance imaging, *KPS* - Karnofsky Performance Score, *MGMT* - 06-methylguanine-DNA-methyltransferase, *NCE* - non-contrast enhancing, *OR* - odds ratio, *OS* - overall survival, *p* – p-value, *PFS* – progression-free survival, *RPA* - recursive partitioning analysis, *RTOG* - Radiation Therapy Oncology Group, *SD* – standard deviation, *STR* – subtotal resection, *WHO* – world health organization.

Study	Study design	Key intervention analyzed	Key surgical findings	Relevant subgroups analyzed	Pre-specified statistics
Glioblastoma					
NCT02379572 - Roder *et al.* in *J Clin Oncol*, 2023	Non-randomized, parallel cohort controlled trial (glioblastoma: *n* = 277)	ioMRI-guided resection *versus* 5-ALA-guided resection	• No difference in complete resection rate (81% *versus* 78%; OR 1·09, CI: 0·57-2·08; *p* = 0·79) or OS (HR 1·00, CI: 0·64-1·55; *p* = 0·99) between ioMRI *versus* 5-ALA	• OS and PFS stratified for different amount of residual CE tumor volume • OS stratified for residual tumor volumes in subgroups defined by MGMT promotor status	• Applied statistical tests pre-defined • Sample size calculation *a priori* done • No pre-defined protocol for subgroup analysis
RESECT (NCT01811121) - Picart *et al*. in *J Neurosurg*, 2023	Randomized, single-blinded, controlled phase III trial (glioblastoma: n=171)	5-ALA-guided *versus* conventional white-light guided resection	• Higher rate of complete resection of the contrast enhancement using 5-ALA: 79% vs. 48% (absolute difference 29%, CI: 17-40; *p* < 0·0001). • Complete resection was associated with higher OS (HR 0·65, 0·42-1·01; *p* = 0·05)	No subgroup analysis done	• Unclear whether statistical tests pre-defined • Sample size calculation *a priori* done • No pre-defined protocol for subgroup analysis
GGN (no NCT available) - Kreth *et al.* in *Ann Oncol*, 2013	Prospective longitudinal cohort study (glioblastoma: *n* = 273)	Complete *versus* subtotal resection *versus* biopsy	• Complete resection was associated with higher median OS: 17·1, CI: 12·6-21·5 *versus* 11·7, CI: 10·0-13·5 months (*p* = 0·001) • No differences in OS between subtotal resection and biopsy (*p* = 0·1)	Outcome stratified for extent of resection in following subgroups: • Treatment regimens • MGMT promotor status	• No pre-defined statistical protocol • No sample size calculation • No pre-defined protocol for subgroup analysis
NCT01394692 - Senft *et al*. in *Lancet Oncol,*	Randomized, open-label, controlled trial (glioblastoma: *n* = 46; anaplastic astrocytoma: *n* = 1, anaplastic oligodendroglioma: *n* =	ioMRI-guided *versus* conventional white-light guided resection	• Complete resection more frequent with ioMRI guided resection as compared to conventional resection: 96% *versus* 68%, *p* = 0·023 • Higher 6-month PFS with ioMRI guided resection as compared to conventional resection (67% *versus* 34%; OR 0·28, CI: 0·09-0·91; *p* = 0·046)	• Rate of complete resections compared in junior *versus* senior neurosurgeons • Rate of residual tumor depending on the use of neuronavigation • Outcome stratified for extent of resection in the overall cohort and in newly diagnosed grade IV tumors only	• Unclear whether statistical tests pre-defined • Sample size calculation *a priori* done • No pre-defined protocol for subgroup analysis

2011	1)				
NCT00241670 - Stummer *et al*. in *Neurosurgery*, 2008	*Post-hoc* analysis of the trial by Stummer *et al*. (glioblastoma: *n* = 243 from per-protocol cohort)	Complete *versus* subtotal resection	• Complete resection was associated with higher median OS: 16·7, CI: 13·4-19·0 *versus* 11.8, CI: 10·4-13·7 months ($p < 0.0001$; HR 1·75, CI: 1·26-2·44; $p = 0·0004$)	Outcome stratified for extent of resection in the following subgroups: • 5-ALA (*versus* white-light) • Eloquent (*versus* non-eloquent) • >60 years of age (*versus* <60 years)	• No pre-defined statistical protocol for *post-hoc* analysis • *Post-hoc* analysis: rested upon available cohort from NCT00241670 • No pre-defined protocol for *post-hoc* subgroup analysis
NCT00241670 - Pichlmeier *et al*. in *Neuro Oncol*, 2008	*Post-hoc* analysis of the trial by Stummer *et al*. (glioblastoma: *n* = 243 from per-protocol cohort)	Complete *versus* subtotal resection	• Overall cohort: complete resection was associated with higher median OS: 16·7, CI: 14·3-19·0 *versus* 11·8, CI: 10·4-13·7 months ($p < 0·0001$) • Subgroup analysis: complete resection associated with higher median OS in the RTOG-RPA class IV and V (IV: 17·7, CI: 14·3-22·5 *versus* 12·9, CI: 10·3-14·7; V: 13·7; CI: 8·3-17·6 *versus* 10·4, CI: 8·1-11·; $p = 0·0007$).	Outcome stratified for extent of resection in the subgroups determined per RTOG RPA (class III-V)	• No pre-defined statistical protocol for *post-hoc* analysis • *Post-hoc* analysis: rested upon available cohort from NCT00241670 • No pre-defined protocol for *post-hoc* subgroup analysis
NCT00241670 - Stummer *et al*. in *Lancet Oncol*, 2006	Randomized, controlled trial (grade III/IV: *n* = 270 in full-analysis cohort, including *n* = 137 glioblastomas)	5-ALA-guided *versus* conventional white-light guided resection	• Higher rate of complete resection of the contrast enhancement in the 5-ALA arm: 65% *versus* 36% (absolute difference 29%, CI: 17-40, $p < 0·0001$) • 6-months PFS higher in 5-ALA arm: 41·0%, CI: 32·8-49·2 *versus* 21.1%, CI: 14·0-28·2 with an absolute difference of 19·9%, CI: 9·1-30·7; $p = 0·0003$	Outcome as well as frequency of and time to re-resection stratified for type of surgery in the following subgroups: • Eloquent (*versus* non-eloquent) • >55 years of age (*versus* <50 years) • KPS >80 (*versus* 70-80) Outcome stratified per residual tumor volume	• Applied statistical tests pre-defined • Sample size calculation *a priori* done • Subgroup analyses based on factors used for randomisation
Astrocytoma and oligodendroglioma					
Shaw *et al.* in *J Neurosurg*, 2008	Prospective longitudinal cohort study (astrocytoma grade II: *n* = 61, oligodendroglioma grade 2: *n* = 50)	Complete *versus* subtotal resection	• Residual tumor volume ≥1 cm predictive for PFS (HR 3·54, CI: 1·83-6·84; $p = 0·0002$)	Outcome stratified per residual tumor volumes in the following subgroups: • Astrocytomas • Oligoastrocytomas • Oligodendrogliomas	• Unclear whether statistical tests pre-defined • No sample size calculation • Unclear whether pre-defined protocol for subgroup analysis

Figure 1: Comparison of study designs in surgical neuro-oncology

Design, implementation, and analysis	Randomized controlled trial	Observational study + propensity-score matching
Prospective design	Yes	Optional
Accounts for known confounders	Yes	Yes
Adequate when there is clinical or individual equipoise	Yes	Yes
Adequately powered	Yes	Yes
Stratification	Optional	Optional
Can identify causal relationships or efficacy	Yes	Yes
Accounts for unknown confounders, high internal validity	Yes	No
Loss of statistical power because of matching	No	Yes
Adequate when there are ethical concerns for randomization	No	Yes
Adequate when there is no clinical or individual equipoise#	No	Yes
Cost	High	Low
Potential problems with accrual	Yes	No
Risk of limited external validity	Yes	No
Episodic in design and techniques used / studied	Yes	No
Double-blinding or sham-controlled design possible	No	No

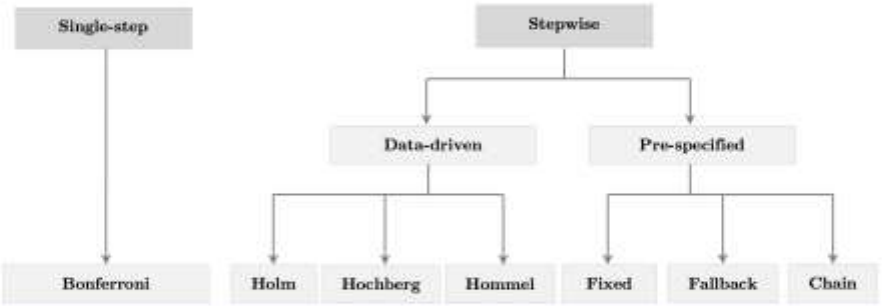
#No equipoise which causes high heterogeneity in indication setting and procedure

NB: green color indicates favorable criteria (n = 11 for observational, n = 8 for RCT), while orange-red color indicates unfavorable criteria (n = 4 for observational, n = 7 for RCT).

Figure 2A: Subgroup analysis methods

	Prospective confirmatory	Prospective exploratory	Post-hoc exploratory
Objective	Efficacy assessment	Consistency assessment	Hypothesis generating
Characteristics	Predefined subgroups Predefined interactions Predefined efficacy threshold Subgroups included in power analysis Always multiplicity correction	Predefined subgroups Predefined interactions Predefined consistency criterion Subgroups included in power analysis Multiplicity correction optional	No predefined subgroups or interactions Subgroups not included in power analysis Multiplicity correction optional
Pearls	Adaptive design to combine exploration and confirmation of treatment effect or threshold Enriched design to increase power of subgroup analysis Decision making framework	Forest plots Shrinking estimation Treatment-by-subgroup interaction or differential treatment effect test	Predefined algorithms for subgroup identification Assess reproducibility Positive findings must be confirmed in a confirmatory analysis
Pitfalls	Inadequate definition of enriched subgroup Inappropriate choice of multiplicity correction method Not using enrichment adjustment approach	Consistency criterion fails to take into account variability around point estimate Consistency criterion less useful when investigating treatments with large interactions	Cannot be used for inference High risk of type II errors Avoid large numbers of post-hoc subgroup analyses "Data mining" after failed study

Figure 2B: Multiplicity correction methods



	Bonferroni	Holm	Hochberg	Hommel	Fixed	Fallback	Chain
Pros	Easy to use Non-parametric	More power than Bonferroni Non-parametric (Holm)			Easy to use	Testing continues after a hypothesis is not rejected	Transfers error rate
Cons	Prone to be overly conservative	More advanced than Bonferroni Parametric (Hochberg, Hommel)			Testing stops once a hypothesis is not rejected	More advanced than fixed-sequence	

Figure 3: Pearls and pitfalls for two common subgroup analysis methods in surgical neuro-oncology

	Regression modelling in a propensity-score matched cohort design	Recursive partitioning analysis
Indications	Etiological analysis	Discovering new subgroups Data mining
Considerations Advantages	Prospective or post-hoc in design Matching minimizes confounding bias by creating balanced groups at baseline Matching is flexible: possibility to adjust settings such as caliper, matching ratio, and matching method Matching is transparent: covariate balance before and after matching can be presented Useful in observational studies when there are ethical concerns for performing a randomized controlled trial Able to handle heterogeneity in local practices with excellent external validity	Nonparametric method Simple, intuitive, easy-to-interpret Able to identify synergy among factors Able to identify nonlinear relationships between predictor and outcome Able for construct homogenous risk strata Able to handle missing values and outliers Able to handle large number of categories Able to identify predictors when little is known about the relationships between predictors and between predictor and outcome
Considerations Disadvantages	Parametric method Selection of appropriate matching covariates (known confounders) influences propensity scores: subject matter knowledge is necessary Matching decreases power and increases type II error rate because of excluding non-matched cases Patients from different groups with overlapping propensity scores may be similar to eachother but might not be representative for their own group Does not take into account unknown confounders	Post-hoc in design Not ideal with continuous or multilevel outcomes Tree building and structure can be unstable and may be prone to overfitting Often not possible to use for inference (random forests should be used for this) May miss additional predictive factors during the latter stages of the partitioning May increase multiplicity since it considers large number of possible partitions Less useful with low number of patients in dataset: unable to adequately partition
Analysis and presentation	Determine matching exposure, covariates, matching ratio and caliper settings: balance between too relaxed (residual confounding) or too strict (decreased external validity) Perform stratification if applicable Present the outcome of the matching procedure: covariate balance before and after matching Perform etiological regression analyses on matched cohorts Perform predictive regression analyses on unmatched cohorts Perform sensitivity analysis	Consider if RPA is the most appropriate machine learning method or if an alternative method should be used: e.g., another CART method (MARS, GAM) or an unsupervised algorithm such as clustering "Prune the tree" to counteract overfitting Know when to stop pruning: use an independent test set or use cross-validation (in case of smaller datasets) Use bootstrapping methods to illustrate the variability of single trees Use random forests with bootstrapping or bagging methods for inferential analyses

Figure 4: Recommendations for subgroup analyses in surgical neuro-oncology studies

Indication and planning

- Consider the objective of the subgroup analysis to select the appropriate method
- Consider the advantages and disadvantages of nonparametric and parametric methods for subgroup analyses (Figure 3)
- Consider post-hoc subgroup analyses to discover subgroups with desirable features such as improved benefit or reduced risks, side effects, or complications
- Consider adaptive trial designs to combine exploration and confirmation of a treatment effect or covariate threshold
- Consider targeted subgroup enrichment to increase the power of the subgroup, but one should be cautious to adequately define the subgroup and to anticipate delays in patient accrual
- Predefine the subgroups and their assumed direction of treatment effect: if possible, base them on important factors such as stratification factors or based on clinical plausibility of the beneficial effect. Avoid defining them based on outcome factors
- Predefine potential interactions between covariates
- If no subgroups are predefined, predefine the specific algorithms for systematic (disciplined) subgroups identification
- Include predefined subgroups in the power analysis and sample size calculation: failure to do so will inflate the type II error (false negatives) due to inadequate statistical power
- When planning a confirmatory subgroup analysis, predefine an efficacy threshold for the treatment effect that is expected to be achieved in the selected subgroup
- When planning a consistency assessment, predefine a threshold that the treatment effect of the least benefitted subgroup (or the complemented subgroup in a two-arm study) should meet (consistency criterion)
- When planning a post-hoc subgroup analysis as part of a "failed study", make sure that this is based on a solid clinical or methodological reason (e.g., inclusion criteria too relaxed to demonstrate overall effect).

Implementation and analysis

- Consider using stratification methods to make the subgroups more comparable (homogeneous) before performing the subgroup analysis
- Avoid performing large numbers of subgroup analyses: this will lead to inflation of the type I error (false positives) and subsequent over-interpretation of the results
- When performing confirmatory subgroup analyses, complete three steps: trial design and population selection, multiplicity correction, decision making
- Correct for multiplicity using one of three methods: lowering the significance level alpha, incressing the p-value, or widening the confidence interval of the individual tests
- The Bonferroni method is the easiest multiplicity correction method but when large numbers of subgroup analyses are performed, more advanced methods such as Holm's, Hochberg's, or Hommel's should be used because their increased power
- Use Bonferroni or Holm's method when the data warrants the use of a nonparametric method
- To visually present treatment effect heterogeneity, use forest plots
- To test for treatment effect heterogeneity, use treatment-by-subgroup interaction tests; however, in small subgroups, use tests of differential treatment effect instead of interaction tests
- If subgroup enrichment has been used, one should consider an enrichment adjustment approach to make the results applicable as if there was no enrichment design
- When using recursive partitioning analysis, prune the tree to counteract overfitting
- When using regression modelling, combine it with propensity-score matching to decrease confounding-by-indication
- When using regression modelling for etiological (inferential) analyses, use the matched subgroups
- When using regression modelling for predictive analyses, use the unmatched subgroups

Interpretation and reporting

- Discuss the methodological and statistical considerations in the Methods section
- Report all the subgroup analyses: not only the significant ones
- After performing confirmatory subgroup analyses, consider the tailored effect claim, broad effect claim, and the enhanced effect claim
- After post-hoc subgroup analyses, assess the reproducibility: clinical plausibility, effect size, and confirmation of the results
- Post-hoc subgroup analyses are over-interpreted easily due to their proneness to high type II error rates (false negatives) and cannot be used for inferential conclusions
- Positive findings in an exploratory subgroup (either prospective or post-hoc) should always be confirmed in subsequent confirmatory subgroup analyses