DOWNLOAD PDF SUBGROUP COMPARISONS

Chapter 1 : How to work with a subgroup analysis

Subgroup comparisons are done by: a) dividing cases into the appropriate subgroups, describing each subgroup in terms of a given variable, and comparing those descriptions across subgroups. b) dividing, describing, and comparing subgroups on independent and dependent variables. c) analyzing the.

Abstract In the analysis of prevention and intervention studies, it is often important to investigate whether treatment effects vary among subgroups of patients defined by individual characteristics. However, subgroup analyses can be misleading if they test data-driven hypotheses, employ inappropriate statistical methods, or fail to account for multiple testing. These problems have led to a general suspicion of findings from subgroup analyses. This article discusses sound methods for conducting subgroup analyses to detect moderators. Multiple authors have argued that, to assess whether a treatment effect varies across subgroups defined by patient characteristics, analyses should be based on tests for interaction rather than treatment comparisons within the subgroups. We discuss the concept of heterogeneity and its dependence on the metric used to describe treatment effects. We discuss issues of multiple comparisons related to subgroup analyses and the importance of considering multiplicity in the interpretation of results. We also discuss the types of questions that would lead to subgroup analyses and how different scientific goals may affect the study at the design stage. Finally, we discuss subgroup analyses based on post-baseline factors and the complexity associated with this type of subgroup analysis. Moderator, Subgroup analysis, Heterogeneity, Interaction, Subset Introduction Subgroup analyses are often performed as part of the analysis of prevention or intervention studies to assess whether treatment effects vary across subpopulations. For example, Sacks et al. Subgroup analyses are needed to refine guidance for patient management. This paper describes methods for conducting subgroup analysis in randomized studies. Definition Subgroup analysis refers to any comparison of patient outcomes between treatment groups across subsets of patients defined by patient characteristics. We will focus on subgroup analyses for subgroups defined by baseline factors. In such cases, the usual question of interest is whether the treatment effect varies among the levels of the baseline factor Kraemer This type of analysis is referred to as a moderator analysis in the social sciences. A variable is a moderator if it satisfies both eligibility and analytic criteria Kraemer et al. The eligibility criterion requires that the variable under consideration precede treatment in time and be uncorrelated with treatment. In a randomized clinical trial RCT , baseline characteristics satisfy the eligibility criterion by study design. The analytic criterion calls for demonstration of treatment effect heterogeneity across levels of the grouping variable. VanderWeele and Robins define moderators in the causal inference potential outcome framework and discuss four types of effect modification based on directed acyclic graphs. For specificity, we assume that a study includes an investigational treatment and a control, but the ideas carry over to any comparison between treatment groups. If the study endpoint is a continuous outcome, for example, weight gain or increase in blood pressure, the treatment effect may be measured by the difference between means. If the study endpoint is a binary outcome, for example, whether a patient with Hepatitis C had a sustained virologic response 6 months after randomization, the treatment effect may be measured by the difference in proportion of responders, the rate ratio, or the odds ratio of the response rates. Sometimes the study endpoint is a time-to-event such as time from randomization to death. In this case, the treatment effect may be measured by the treatment versus control hazard ratio or the arithmetic difference in survival rates at a specific follow-up time. Though our primary focus is RCTs, these measures for comparing treatment groups are also employed in observational studies, for example, to characterize the effect of a new policy on health services, or the effect of behavioral interventions on prevention. Type of Subgroup Analysis and Study Design The questions asked in subgroup analyses have various levels of specificity. In the most general formulation, one can ask whether the treatment effect seen overall is consistent across categories of patients defined by levels of a baseline characteristic. In this case, there is usually no specific hypothesis about the type of heterogeneity that might be observed. Rather, the analyses are motivated by the recognition that the treatment effect might depend on patient characteristics and the desire to assess whether and how such variation occurs. For example, Jackson et al. This type of subgroup analysis can be regarded as hypothesis-generating and

important findings should be validated in future studies Kraemer et al. In contrast, a hypothesis-testing subgroup analysis can occur when one is interested in learning how effects of a new treatment vary according to a baseline factor identified a priori, perhaps motivated by a previous study. One may ask whether the treatment effect increases with the level of a specific ordinal or continuous baseline factor as in Sacks et al. In addition to the overall question, whether treatment with statin drugs after initial MI reduces risk of additional cardiac events in patients with LDL cholesterol levels in the normal range, the authors were interested in assessing whether the magnitude of this benefit depends on the baseline LDL cholesterol level. Plans for subgroup analyses should be considered during the design of a study. If investigators are interested in testing hypotheses about specific variables or specific relations, these hypotheses should be included in the primary or secondary objectives. It would be advisable to consider stratified randomization of treatment assignments to ensure sufficient representation in the subgroups of interest so that the study has adequate power to detect the moderation effect. Methods for Conducting Subgroup Analysis Subgroup analysis usually starts with a test for interaction; that is, a test to determine whether the relative effects of study treatments vary significantly among subgroups of patients. Various interaction tests have been proposed for detecting treatment effect heterogeneity Byar ; Byar and Green ; Halperin et al. Typically, one chooses a model based on the type of response variable and then includes an interaction term s in the model. Aiken and West discussed how to test and interpret interactions using multiple regression. In what follows, we use three examples to illustrate how subgroup analyses can be conducted when the response variable is continuous, binary, or a time-to-event. Continuous Endpoint Tolan et al. SAFEChildren II included an additional program consisting of session multiple family groups, paired with a reading club with access to age-appropriate books. Let Y denote academic achievement, measured by standardized scores on reading skills tests used by the schools. A formal test for moderation is a test of the null hypothesis H0: Dichotomous Outcome In Gardner et al. In such cases, a logistic regression model is commonly used. A formal test for moderation can again be performed by testing the null hypothesis H0: Time-to-event Outcome If the outcome, Y, is a time-to-event outcome, a Cox proportional hazard model is commonly used. One outcome measure was the time to new-onset hypertension. A formal test of the null hypothesis H0: If the test for interaction yields statistically significant results, then the data suggest that the treatment effect differs among subgroups. In this case, the overall treatment effect, whether statistically significant or not, may not be relevant and can be misleading for patient management. Therefore, when a significant interaction is identified, the treatment effect within each subgroup should be presented. In the linear models discussed above, we considered dichotomous baseline factors and assessed treatment heterogeneity across the two resulting subgroups. These methods extend readily to ordinal or continuous baseline factors. When the baseline factor is continuous, it may be advantageous to define clinically meaningful categories to facilitate interpretation of subgroup analyses. However, several authors have pointed out the negative consequences of categorization of the candidate variable. We recommend that categorization employed for ease of interpretation, if necessary, be introduced only after a significant interaction between treatment and the continuous candidate variable has been demonstrated. If the linearity assumptions for the effects of the variables on the outcome are likely to be violated, generalized additive models GAMs can be utilized to allow non-linear effects Hastie and Tibshirani ; Marra and Radice Specifically, a GAM is an additive regression model of the form: They can be either non-parametric smoothers or regression splines, and can be determined using the data, thereby allowing the linearity assumptions to be relaxed. Out of concern for violation of model assumptions and loss of interpretability of the effect sizes associated with linear models, Kraemer proposed a non-parametric approach for detection of binary candidate moderators and outlined ideas for extending this approach to non-binary variables. This approach is based on the area under the ROC curve AUC , which estimates the probability that a randomly selected patient in the treatment arm has a clinically more desirable response than a randomly selected patient in the control arm. This nonparametric approach does not require the assumptions associated with a linear model, but yields consistent results when these assumptions hold. Qualitative interactions, the interactions that result in directional changes of treatment effects in different subgroups of patients Peto , are especially important since they often have implications for patient management. Several tests have been proposed for assessing whether

there are qualitative interactions across I disjoint patient subsets in the setting where two treatments are compared, including a likelihood ratio test Gail and Simon , a range test Piantadosi and Gail , and a test based on simultaneous confidence intervals Pan and Wolfe Piantadosi and Gail found that, if the new treatment is harmful in a few subsets, the range test is more powerful than the likelihood ratio test; otherwise, the likelihood ratio test is more powerful. Silvapulle obtained an exact null distribution for the Gail-Simon test statistic and proposed tests that are robust against outliers. Li and Chan extended the range test by performing the usual range test on the extreme values of all the subgroups first and subsequently on all subgroups of the subsets in a stepwise manner. One limitation of these tests is the necessity of grouping subjects into disjoint subsets using pre-specified criteria. Several graphical methods have been proposed recently. Song and Pepe proposed the selection impact SI curve, which can be used to choose a treatment strategy based on whether the value of a single biomarker exceeds a threshold. Bonetti and Gelber , proposed the subpopulation treatment effect pattern plot STEPP method, which provides a display of treatment effect estimates for different but potentially overlapping subsets of patients. Although the inference procedure for the STEPP method allows patients subsets to be defined according to more than one covariate, it is challenging to develop grouping criteria in this case. Motivated by the notion that a treatment may work best for the sickest patients, Follmann and Proschan examined treatment interaction along a single severity index defined by a linear combination of baseline covariates. This method uses estimates of individual-level treatment differences to create an index for clustering subjects, and then makes inferences about average treatment differences in each cluster of subjects. Classification and regression tree CART analysis is another useful tool for investigating interactions among baseline factors without imposing parametric assumptions on the relationship between the outcome and candidate variables Breiman et al. To illustrate, consider the scenario described in Table 1 , where Rt denotes the mortality risk in the treatment group, Rc denotes mortality risk in the control group, RR denotes the relative risk and Rc-Rt denotes the risk difference. We note that risk increases with baseline performance status in each treatment group. Table 1 A hypothetical illustrative example for the dependence of heterogeneity of treatment effects on metric Baseline.

Chapter 2 : PROSPECT HEIGHTS SD 23 | District Snapshot

CASOAâ„¢ Subgroup Comparisons. CASOAâ„¢ Subgroup Comparisons â€¢ â€¢ â€¢ â€¢ â€¢ NRC is a charter member of the AAPOR Transparency Initiative, providing clear disclosure of our sound and ethical.

First of all, subgroup analyses may demonstrate consistent results over various complementary subpopulations, e. This would indicate stability of a treatment effect over a broad study population. Subgroup analysis could also identify patient subsets with a particular treatment effect, either positive or negative. This might be of interest if high rates of side effects call for selective use of a new therapy. Finally, in trials with an overall negative result, subgroup analyses might identify patient subsets with a significant treatment effect. While the first two approaches are performed frequently, the third is considered inappropriate by the scientific community. What do statisticians think about debates on subgroup data? One major concern is that few trials have sufficient statistical power to estimate a treatment effect reliably in multiple subgroups. So, statisticians would caution us that these analyses are hypothesis generating at best, and cannot be regarded as evidence. Is that just the perfect line never to be missed in the limitations section of a manuscript, or should we be more cautious beyond this? Data on the elderly are of common interest. Treatment effects could be quite different and side effects more frequent in older patients. However, the question is what is elderly? Is it all patients over 65 years of age? Or is it 75, 80, or perhaps 85 years? A trial database is easily analysed for all these different cut-off values. Such data sets may be very consistent. However, they could also be quite divergent, with significant differences at one particular cut-off, perhaps non-significant trends at others, or results contradicting the main results. Sometimes these results would fit current knowledge or a hypothesis, but sometimes it may be quite difficult to understand differences that emerge. They could indicate a differential treatment effect based on age, but could also be play of chance. Which data will end up in a publication? Most probably those that fit the hypothesis, or provide the most evidence. That leads us back to Peter Sleight. He once commented on subgroup analyses of ISIS-2, which had indicated a beneficial effect of aspirin after myocardial infarction in patients born under all astrological signs except for Gemini and Libra. However, when presented with other less ridiculous subgroup analyses they are likely to believe the results, â€¦, particularly if the result can be justified by some pet theory. There are further issues, relating to adjustments for covariates, for multiple comparisons, the decrease of statistical power, etc. So, how can we identify data which are the result of play of chance, but do not make us burst out laughing? How can we be certain that the data presented were not taken out of the context of an array of multiple analyses performed, chosen in the end because they best fit a hypothesis? How can we identify more reliable subgroup analyses? Subgroup analyses which have been pre-specified before data are available would eliminate data selection, but not play of chance. Was the subgroup variable a stratification factor at randomization? Was the subgroup hypothesis specified a priori? Was the significant interaction effect independent, if there were multiple significant interactions? Was the direction of the subgroup effect correctly pre-specified? Was the subgroup effect consistent with evidence from previous related studies? Was the subgroup effect consistent across related outcomes? Was there any indirect evidence to support the apparent subgroup effectâ€”for example, biological rationale, laboratory tests, animal studies? Was the subgroup variable a baseline characteristic?

Chapter 3 : subgroup comparison - SAS Support Communities

The limitations of subgroup analyses are well establishedâ€"false positives due to multiple comparisons, false negatives due to inadequate power, and limited ability to inform individual treatment decisions because patients have multiple characteristics that vary simultaneously.

Open in a separate window The horizontal arrow indicates a within-subgroup test. The results of this test are called a subgroup effect. The vertical arrow indicates a between-subgroup interaction test. The results of this test are called an interaction. It is not performed in our example. The purpose of this article is to consider criteria for sound subgroup analyses in RCTs, assuming good underlying methodological quality of the main trial i. Clinical scenario A year-old woman keeps returning to your practice with recurrent anterior dislocations of her shoulder. Since her initial dislocation more than 3 years ago, she has had 3 recurrent dislocations and several subluxations of her shoulder. On the second dislocation, you tried a different method of reduction and immobilized her shoulder for a longer time, but unfortunately this did not prevent another redislocation. You noticed that some patients in your practice with dislocated shoulders did not have any recurrences, despite receiving exactly the same treatment. Age of the patient 4 â€" 6 and duration of immobilization 7 , 8 might explain the difference in recurrence, but the data remain largely inconclusive. You recall a colleague discussing immobilization of the shoulder in external rotation ER rather than the usual internal rotation IR as a great method to reduce recurrences. You decide to expand your search to identify the best available evidence on shoulder position. Literature search To find out if internal rotation immobilization has ever been compared with another immobilization method, you search the available literature. You perform a comprehensive search 9 using the following search terms: You retrieve the article for further evaluation while consulting guidelines to assess surgical RCTs. The authors hypothesized that immobilization in ER would decrease the recurrence rate. From a total of patients with a mean age of 37 years, 94 patients were randomly assigned to immobilization up to 3 days after reduction in IR and to immobilization in ER for 3 weeks. The primary outcome assessed was a recurrent dislocation or subluxation of the shoulder, and the minimum follow-up period was 2 years. Significance of subgroup tests was set at the 0. No significant differences were found in the other age groups. Although the example is a nonoperative one, similar guidelines can be applied to RCTs on operative interventions. Was the subgroup analysis predefined or was it carried out post hoc? Was the subgroup analysis one of a small number? Did the power calculation account for between-subgroup treatment effects? Were subgroup definitions based on prerandomization patient characteristics? Was randomization stratified for important subgroup variables? Analysis Were interaction tests used for assessing subgroup treatment effect interactions? Were the significances of treatment effect interactions adjusted for multiplicity? Were the subgroups checked for comparability of prognostic factors? Reporting Are all performed subgroup analyses reported? Are subgroup analyses reported as relative risk reductions? Does the emphasis of the discussion and conclusion remain on the overall treatment effect? Applicability Is the subgroup difference consistent across other studies? Is the subgroup effect or interaction clinically important? Are the patients in the subgroup comparable to my patients? Is the between-subgroup treatment effect clinically important? Design Was the subgroup analysis based on a rational indication? The first step in evaluating a subgroup analysis is to determine its logical sense. In line with the main RCT analysis i. When undertaken without any clinical explanatory background, no subgroup analysis will attain practical utility, no matter how significant the results are. Patient groups that are expected to have different treatment effects compared with the general trend may be analyzed as separate groups, but only if explained by differences in risk of attaining a certain outcome or by differences in pathophysiology. Especially in surgical trials in which mortality is a frequently measured outcome, it is of great importance that patients with a high mortality risk will be recognized and analyzed separately. As for subgroups based on pathophysiology, differences in underlying disease mechanisms could induce heterogeneity in treatment effect and therefore justify a subgroup analysis. The report by Itoi and colleagues 10 explains that the subgroup of patients younger than 30 years was chosen because of previously demonstrated increased risk for redislocation in this group. However, it does not provide

a rationale for analyzing this subgroup for treatmentâ€”control differences in secondary outcomes compliance, sports participation, shoulder stiffness. Results would have made more sense if the authors had expressed evidence-based hypotheses that secondary outcomes would particularly decrease in this age group. Similarly, they did not justify their subgroup analysis based on the delay between dislocation and immobilization. Subgroup analyses that are performed to test hypotheses generated before the study has started should be clearly distinguished from those identified after the main trial analyses are performed. Such analyses are generated by the trial data rather than the data being tested, and they should be regarded as unreliable unless they can be replicated by other studies. Besides predefining the subgroup variables, the expected direction the same or the opposite direction as the overall treatment effect and the magnitude of the subgroup effects should be reported at the beginning of the trial. For example, the results of a certain subgroup analysis can become statistically significant when an alternative definition is used. However, in the description of the statistical analysis in the methods section, they state that they analyzed a subgroup of patients aged 30 years or younger and that they categorized age in 4 subgroups, but they do not define the subgroups in that section. In the discussion section, the authors report the results for a subgroup of patients aged 30 years or younger, but they analyzed 2 different groups under the age of 30 individually, of which the results were significant in the subgroup of patients aged 21â€”30 years. This illustrates the importance of exact predefined definitions of the subgroup categories. The chance of falsely obtaining significant subgroup effects and interactions i. The number of subgroup analyses is the product of the number of subgroups and the number of outcomes analyzed. Therefore, the outcome measures used to compare subgroups should be limited to the primary outcome of the main trial and secondary outcomes that are unique to specific subgroups. In addition, the number of subgroups should be limited. This may prevent subgroups from becoming too small, thereby reducing the chance of false-negative results. In addition to the methodological setbacks, conducting too many subgroup analyses will result in confusion for both readers and authors. Exhausting subgroup analyses distract readers from the key message concerning the observed overall effect. Additionally, it is hard for authors to discuss their results in a well-organized and clear manner. Itoi and colleagues 10 basically repeat their main effect analyses on their subgroup of patients aged 30 years and younger. Adding to the complexity, they further subdivide this subgroup based on the delay between dislocation and immobilization. In fact, this is a double subgroup analysis, which should certainly be interpreted cautiously. The power of a trial is the ability to detect a difference between 2 groups if one truly exists and is positively correlated with the magnitude of the treatment effect and the sample size of the study. It is very unlikely for these trials to detect subgroup effects or interactions because subgroups always include fewer patients than the main treatment groups. Consequently, subgroup analyses are frequently underpowered, which means there is a greater probability of false-negative results. For detection of interactions of the same size and with the same power as the overall effect, the sample sizes should be inflated 4-fold. This means that even larger sample sizes are needed, and these are very unlikely to achieve in practice. Therefore, it would be more reliable to look at the overall results of a study than the apparent effect observed within a subgroup. Because the authors applied the same statistical test for each within-subgroup analysis, they would have needed the same number of patients in each subgroup as the number calculated for the overall treatment groups 42 patients for each group to reach a similar power for each subgroup analysis. As the sample size needed for a certain power is also dependent on the estimated effect size, the subgroups should have contained even more patients to detect a smaller effect than the overall effect. Except for the immobilization with IR and ER on day 1 subgroups, none of the other subgroups contained a sufficient number of patients, which means that the probability of false-negative nonsignificant results was large for all these subgroups. The results from the subgroup analyses are therefore probably not valid, and the conclusion about the absence of treatment effect should be questioned. A variable by which a subgroup is defined should not be affected by treatment response. Thus, only disease characteristics obtained before randomization and independent patient characteristics e. If subgroups are based on outcome-dependent data, an observed interaction may be simply the result of one subgroup that had a better prognosis rather than being truly caused by the treatment. For example, comparing compliant to noncompliant patients is invalid since compliance is related to prognosis. Comparing treatment effects for patients who did

and did not crossover would cause misleading results. After all, crossover patients have significantly different prognostic characteristics than patients who receive the treatment to which they were randomly assigned. Additionally, patients who crossed over to nonoperative care were older, had less pain and experienced less disability. For a subgroup analysis to be clinically applicable, surgeons need to know what types of patients are going to benefit from a type of treatment before they decide on a treatment option. Itoi and colleagues 10 specified their subgroups by patient age and by the day immobilization was started. These variables cannot have been influenced by the effect of immobilization in any way. Was randomization stratified by important subgroup variables? In the design of a trial with predefined subgroups, stratification of randomization by important subgroup variables should be considered. Because subgroups usually are of limited size, it cannot be assumed that prognosis at baseline is similar among subgroups unless randomization was stratified. When stratification of randomization is based on subgroup variables, it is more likely that treatment assignments within subgroups are balanced, making each subgroup a small trial. Because randomization makes it likely for the subgroups to be similar in all aspects except treatment, valid inferences about treatment efficacy within subgroups are likely to be drawn. They compared the mean age and other patient characteristics for statistical significance between the 2 treatment groups postrandomization. However, they did not compare the groups with regard to the age categories and the day immobilization was started. In trials comparing 2 operative interventions, surgical skill may be of major importance in determining treatment results. However, this difference may have been due, by chance, to stentless operations having been performed by more skilled surgeons than stented procedures. Analysis Were interaction tests used for assessing between-subgroup treatment effect interactions? One should not question whether a treatment is efficacious in subgroup 1 and subgroup 2 separately both subgroup effects , but if treatment efficacy differs between subgroup 1 and 2 an interaction; Table 1 The former is investigated by simple tests as used for the main analysis e. The most frequently used formal interaction tests include the Mantelâ€"Haenszel technique and regression models. The vertical line indicates similar risks of dislocation recurrence between the ER and IR groups.

Chapter 4 : MCQS Question: Sampling

Subgroup comparisons are done by dividing cases into the appropriate subgroups, describing each subgroup in terms of a given variable, and comparing those descriptions across subgroups. Professor Wilton decided to test whether blondes have more fun.

Chapter 5 : ABINGDON-AVON CUSD | District Snapshot

As Michael notes, when comparing a subgroup to an overall group, researchers typically compare the subgroup to the subset of the overall group that does not include the subgroup.

Chapter 6 : Alumni Survey: Subgroup Comparisons â€“ Office of Institutional Research and Planning

comparisons of subgroup means or item difficulty parameters merely identify group differences, 4 but provide no substantive description of the nature of the differences.

Chapter 7 : Detecting Moderator Effects Using Subgroup Analyses

The Problem of Too Many Statistical Tests: Subgroup Analyses in a Study Comparing the Effectiveness of Online and Live Lectures Abstract The more statistical analyses performed in the analysis of research data, the more likely it is that one or more.