Published in final edited form as:
Prev Sci. 2013 April ; 14(2): 111–120. doi:10.1007/s11121-011-0221-x.

Detecting Moderator Effects Using Subgroup Analyses

Rui Wang and **James H. Ware**
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, SPH2, 4th Floor, Boston, MA 02115, USA

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. These "subgroup analyses" can provide information about how best to use a new prevention or intervention program. 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.

Keywords

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. (1996) reported results from a randomized clinical trial which demonstrated that pravastatin reduces the risk of a fatal coronary event or a nonfatal myocardial infarction in patients with low-density lipoprotein (LDL) cholesterol levels of 115 to 174 mg per deciliter. However, subgroup analysis revealed that the beneficial effect was not seen in the subgroup of patients with LDL levels below 125 mg/dL. 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

rwang@hsph.harvard.edu.

interest is whether the treatment effect varies among the levels of the baseline factor (Kraemer 2006). 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. 2008). 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. Such heterogeneity is called "effect modification" by epidemiologists. VanderWeele and Robins (2007) define moderators in the causal inference potential outcome framework and discuss four types of effect modification based on directed acyclic graphs. Here we consider both categorical and continuous baseline factors and use the term "treatment effect" to refer to some quantitative measure of differences in the distribution of response between two treatment groups. 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. (2006) examined whether the effect of calcium and vitamin D on prevention of colorectal cancer varied across subgroups defined by 16 different patient characteristics. This type of subgroup analysis can be regarded as hypothesis-generating and important findings should be validated in future studies (Kraemer et al. 2002).

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. (1996). 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 1985; Byar and Green 1980; Halperin et al. 1977; Keppel and Wickens 2004; Patel and Hoel 1973; Peto 1982; Schemper 1988; Shuster and van Eys 1983; Sleeper and Harrington 1990). 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 (1991) 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. (2009) assessed whether a booster intervention (SAFEChildren II) achieved larger positive effects than an initial intervention (SAFEChildren I) on growth and level of school achievement and bonding, child behavior and social competence, and parenting practices, family relationships, and parental involvement in school. SAFEChildren I (Schools and Families Educating Children) included a 22-session family intervention to help inner-city parents manage their children's transition into first grade, as well as 20 sessions of academic tutoring for children. SAFEChildren II included an additional program consisting of 20-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. Let X=0, or 1 denote the initial intervention group and the booster intervention group respectively. Let Z=0 or 1 denote male or female, respectively. To test whether treatment effects on academic achievement vary according to child's gender, we can consider the following linear regression model:

$$\mathrm{Y}=\alpha+\beta\mathrm{X}+\gamma\mathrm{Z}+\delta\mathrm{X}*\mathrm{Z}+\varepsilon$$

where ε follows a normal distribution with mean 0 and variance σ^2 and X*Z denotes the product of X and Z. A formal test for moderation is a test of the null hypothesis H_0: δ=0. Note that β denotes the treatment effect in the male group (Z=0) and β + δ denotes the treatment effect in the female group (Z=1). Therefore, δ measures the difference in treatment effects between two subgroups defined by gender. Here γ denotes the difference in mean outcome between males and females in the initial intervention group. If we reject the null hypothesis that δ=0, we can conclude that the treatment effect differs between the two subgroups and gender is a moderator of the treatment effect. We can examine a point estimate and confidence interval for δ to assess the scientific significance of this interaction.

Dichotomous Outcome

In Gardner et al. (2009), families with a 2-year-old were randomly assigned to an empirically supported family-centered intervention (the Family Check-Up) for problem behavior in early childhood (X=1)or a no-intervention control group (X=0). A primary outcome for the trial was whether the behavior was a problem for the parent, as reported by caregivers (Y=1 or 0). It was of interest to assess whether treatment effects differed according to single parenthood, defined as having no partner living in the household (Z=1) or not (Z=0). In such cases, a logistic regression model is commonly used.

$$\operatorname{logit}(p) = \alpha + \beta X + \gamma Z + \delta X * Z,$$

where p = P(Y = 1|X,Z), and logit(p) = log[p/(1 – p)]. In this model, β represents the log-odds ratio of reporting that the behavior is a problem for the parent in the intervention group as compared to the no-intervention group among those *without* single parenthood, and β + δ represents the same log-odds ratio among those *with* single parenthood. Here γ represents the log-odds ratio comparing single parenthood to without single parenthood among those in the no-intervention group. As in the previous example, δ measures treatment effect heterogeneity across the status of single parenthood. A formal test for moderation can again be performed by testing the null hypothesis H_0: δ=0.

Time-to-event Outcome

If the outcome, Y, is a time-to-event outcome, a Cox proportional hazard model is commonly used. Julius et al. (2006) assessed feasibility of treating prehypertension with candesartan. One outcome measure was the time to new-onset hypertension. Let *X*=0 or 1 denote the placebo group or the candesartan group and *Z*=0 or 1 denote male or female, respectively. A Cox PH model would take the form

$$h(t) = h_0(t) \exp(\alpha + \beta X + \gamma Z + \delta X * Z),$$

where h(t) is the hazard function, representing the incidence of hypertension at time t. In this model, β represents the log-hazard ratio of development of clinical hypertension comparing the treatment group to the placebo group among males, γ represents the log-hazard ratio comparing females to males in the placebo group, and β+δ represents the log-hazard ratio of development of clinical hypertension comparing the treatment group to the placebo group among females. Once again, δ measures treatment effect heterogeneity across gender subgroups. A formal test of the null hypothesis H_0: δ=0 can be performed to examine whether treatment effects vary across gender. 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. MacCallum et al. (2002) provide an excellent review. 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 1990; Marra and Radice 2010). Specifically, a GAM is an additive regression model of the form:

$$g[E(Y)] = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p\left(X_p\right)$$

where g(•) is a smooth monotonic link function and X_i's,i=1,…,p, are candidate variables, including possible interaction terms. The f_i's,i=1,…,p, are functions of the corresponding variables. 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 (2008) 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 1982), 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 1985), a range test (Piantadosi and Gail 1993), and a test based on simultaneous confidence intervals (Pan and Wolfe 1997). Piantadosi and Gail (1993) 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 (2001) obtained an exact null distribution for the Gail-Simon test statistic and proposed tests that are robust against outliers. Li and Chan (2006) 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 (2004) 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 (2000, 2004) 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. This method is implemented in an add-on R package “STEPP.” The discussions in Bonetti and Gelber focused on defining patient subsets based on a continuous covariate. 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 (1999) examined treatment interaction along a single severity index defined by a linear combination of baseline covariates. Cai et al. (2010) proposed a method for using multiple baseline characteristics to estimate subject-level treatment effects that can guide patient management and treatment selection. 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. 1998). Lemon et al. (2003) offer an excellent review of the use of CART analysis in public health and behavioral research to identify easily defined, mutually exclusive subgroups that may differentially benefit from the behavioral strategy of interest. CART analysis is implemented in two add-on R packages “rpart” and “tree.”

It is important to note that assessment of the heterogeneity of treatment effects is dependent on the metric used to quantify effects. To illustrate, consider the scenario described in Table 1, where R_t denotes the mortality risk in the treatment group, R_c denotes mortality risk in the control group, RR denotes the relative risk and R_c-R_t denotes the risk difference. We note that risk increases with baseline performance status in each treatment group. In this example, there are no interactions on the `relative risk' scale, but interactions are present on the `absolute risk' scale.

A natural question arises as to which metric to choose. One commonly used method, examining the interaction term in the logistic regression model, assesses interactions on the odds ratio scale. However, the odds ratio is hard to interpret unless it approximates the risk ratio (Sackett 1996; Schwartz et al. 1999). The odds ratio also suffers from a paradoxical effect for chained outcome probabilities (Newcombe 2006). Furthermore, Kraemer (2004) pointed out several problems with the rationales used to justify the wide use of the odds ratio. Ideally, one would pick a metric that is most relevant to the question of interest. If a treatment has serious side effects, then perhaps it is not worth using in patients where the absolute benefit is small. In the illustrative example above, we might conclude that a change from 7% to 3.5% is a substantial risk reduction, whereas a change from 1% to 0.5%, again corresponding to a relative risk of 2, might not be worthwhile, especially when the active treatment has serious side effects. In other situations, the intervention is expensive. For example, for the Building Nebraska Families (BNF) program, an intensive pre-employment, educational program that targets hard-to-employ clients in rural areas and aims to help them develop the basic skills required for employment: the average cost of serving a BNF participant was $7,383 (Meckstroth et al. 2008, p. 42). It is important to consider whether the benefits gained from the intervention outweigh the costs of the intervention. When evaluating the benefits of an intervention relative to other factors such as side effects or costs, the risk difference is often a more relevant metric. Another advantage of the risk difference and its inverse, “the number needed to treat,” is that these measures convey useful clinical information for guiding patient treatment (Altman and Andersen 1999;Cook and Sackett 1995; Wen et al. 2005). Rothman (1986) defines three indices: (1) the relative excess risk due to interaction (RERI), (2) the proportion of disease among those with both exposures that is attributable to their interaction (AP), and (3) the synergy index (S). These indices are useful for assessment of interaction on the additive scale in epidemiologic studies. Rothman notes that, in the absence of additive interaction, both RERI and AP are 0 and S is 1. Hosmer and Lemeshow (1992) show how to obtain confidence intervals for these indices using routine output from logistic regression software based on the delta method estimate of the variance. Assmann et al. (1996) compare and contrast several methods for obtaining confidence intervals for these measures based on the delta method and the bootstrap method and find that the bootstrap percentile confidence interval method has the best coverage behavior for the settings they examined through simulation studies. Li and Chambless (2007) extend the concepts of RERI, AP and S to the setting of a time-to-event endpoint when a Cox proportional hazard model is used, describe confidence interval estimation techniques for these measures based on the delta method and the bootstrap method, and provide SAS program codes for implementation in their Appendix.

A Mistake to Avoid

A mistake in subgroup analyses is to base a claim of heterogeneity on separate tests of the treatment effect within each subgroup. For example, suppose that we want to assess whether a treatment effect varies across gender and we conduct statistical tests separately within males and females. Suppose that we obtain a *p* value <0.001 favoring treatment within males and a *p* value=.52 within females. Can we conclude that treatment is better than control in males, but not in females? Consider the hypothetical situation described in Table 2. These data generate exactly those *p* values, yet the risk difference (RD) is 0.4 within each gender. It is only the smaller sample size that leads to the "nonsignificant" result in females. If we test for interaction, we will not find a significant interaction based on the risk difference because the risk difference estimates are identical for both men and women.

Such circumstances can occur in practice. For example, Bombardier et al. (2000) reported a significantly higher rate of MI in patients receiving rofecoxib than in those receiving naproxen. However, in the subgroup of patients for whom aspirin was not indicated, the risks were not significantly different. Does this result provide sufficient evidence to conclude that rofecoxib is safe for patients when aspirin is not indicated? A closer look at the data (Curfman et al. 2005) revealed that the MI event rate was very low in the "aspirin not indicated" group. Even though the rate of MIs in the rofecoxib group was about twice that of the naproxen group, the difference did not reach statistical significance. The possibly mistaken conclusion that rofecoxib is safe in the "aspirin not indicated" group can potentially impact the majority of the population as most patients belong to this group.

The Multiplicity Issue

Another concern regarding subgroup analyses is multiplicity; that is, the fact that the probability of a false positive finding increases as the number of subgroup analyses increases. Figure 1 in Lagakos (2006) depicts the probability that multiple subgroup analyses will yield at least one, two, or three false positive findings as a function of the number of subgroup analyses conducted. As an illustration, when 20 independent subgroup analyses are performed, each at the 5% level, the probability of at least one false positive is 0.64, the probability of at least two false positives is 0.26, and the probability of at least three false positives is 0.08.

One simple, commonly used method for addressing multiplicity is the Bonferroni method. Let s denote the total number of tests performed. This method sets the allowable error rate for each test to be α/s. This adjustment ensures that the family-wise error rate (FWER), the probability of rejecting one or more of the hypotheses erroneously when performing multiple hypothesis tests (Shaffer 1995), is less than or equal to α. The Bonferroni method is valid regardless of the correlation among tests, but it is conservative because only hypotheses with associated p values $\leq \alpha/s$ are rejected. This results in a reduction in power if some of the hypotheses are actually false. Several sequential methods have been proposed to increase efficiency. Holm (1979) proposed ordering the p values for the s hypotheses from smallest to largest, $p(1) \leq p(2) \leq \ldots \leq p(s)$, and then proceeding sequentially: (1) If $p(1) > \alpha/s$, accept all the s hypotheses; (2) if $p(1) \leq \alpha/s$, reject the hypothesis corresponding to $p(1)$ and then consider $p(2)$; (3) continue successive testing of the hypotheses until the first hypothesis such that $p(l) > \alpha/(s - l + 1)$. A further improvement is provided by Simes-Hochberg (Simes 1986; Hochberg 1988). The modified sequential procedure starts from the largest p value $p(s)$: (1) If $p(s) \leq \alpha$, then all hypotheses are rejected; (2) Otherwise, proceed to $p(s-1)$. If $p(s-1) \leq \alpha/2$, then all hypotheses associated with p values $\leq p(s-1)$ are rejected; (3) Continue until the first hypothesis such that $p(s-l) \leq \alpha/(l+1)$, then all hypotheses associated with p values smaller than or equal to $p(s-l)$ are rejected. Hommel's method is more powerful than the Simes-Hochberg approach but is a similar procedure, except that the ordered p values are compared to values calculated based on the maximum of the p values from the remaining hypotheses (Hommel 1988).

Instead of controlling the FWER, some have suggested approaches that control the false discovery rate (FDR), defined as the expected proportion of falsely rejected hypotheses. It can be shown that FDR ≤ FWER. To control the FDR at f, we first order the p values (p_l, l = 1, 2, …,s) from the s tests in decreasing order. For $l = s$, $s - 1$, …, 1, calculate the critical significance level d_l as $(l/s)\cdot f$, and compare p_l to d_l. If $p_l \leq d_l$, the remaining l tests are rejected (Benjamini and Hochberg 1995). Storey and Tibshirani (2003) illustrate how to use the concept of the FDR to measure statistical significance by using q-values. Tests with q-values smaller than f are called significant, which results in a FDR of f among the significant tests. Dabney and Storey have developed a software called “QVALUE” which takes a list of p-values and estimates their q-values (http://genomics.princeton.edu/storeylab/qvalue/).

For testing pre-specified subgroup hypotheses, we recommend use of methods that control for FWER. Koch and Gansky (1996) discussed statistical considerations for addressing multiplicity associated with subgroup analyses in confirmatory studies. For subgroup analyses intended to be exploratory, methods that control for FDR are, in general, more appropriate.

When multiple potential moderators are identified in hypothesis-generating analyses, a reviewer of this manuscript suggested that it would be advantageous to consider whether these variables can be integrated into a single composite index, which can then be evaluated in future hypothesis-testing studies. For example, if age, gender, cholesterol levels, and blood pressure are found to be moderators, instead of evaluating these moderators separately, we may consider using the Framingham risk index to summarize information contained in these variables for dimension reduction.

As an alternative to formal adjustment for multiple comparisons, one can specify how many tests of true null hypotheses would be expected to be significant depending on the number of total tests performed. This can be a useful way to put the statistical tests reported in a paper into proper perspective. This gives a global assessment of whether the overall subgroup findings could be explained by chance, rather than a corrected p-value for each test. For example, Wactawski-Wende et al. (2006) reported that “37 subgroup comparisons were tested, with 19 reported … accordingly, the results of two tests would be expected to be significant at the 0.05 level by chance” (p. 686). If the number of significant subgroup findings is smaller than or equal to the number expected by chance, then we may suspect that these findings are in fact due to chance. If we observe more significant subgroup findings than would be expected, we are more confident that some of these findings are real. It would then be important to examine each of these findings more carefully, considering effect size, strength of evidence and scientific plausibility, to determine which hypotheses merit further investigation.

Lack of Power

Worthy of mention also is that failure to find significant interactions does not demonstrate definitively that the treatment effect seen overall applies to all subjects. Tests for interaction often have limited power. To illustrate, suppose that we have a clinical trial with two treatment arms, and a continuous outcome, such as weight loss after 6 months of a diet program. Let Y denote weight loss after 6 months, μ_1 and μ_2 the mean weight loss in the two diet programs, and σ the population standard deviation of weight loss after 6 months for either treatment group. A clinical trial is typically powered to detect the main effect $\mu_1 - \mu_2$, which is usually estimated by $\bar{Y}_1 - \bar{Y}_2$, the difference in mean weight loss between the two treatment groups. If we assume that there are n subjects in each treatment group, a simple calculation reveals that $\mathrm{var}\left(\bar{Y}_1 - \bar{Y}_2\right) = 2\sigma^2/n$. However, if we are interested in detecting treatment heterogeneity across gender, then the appropriate statistic to examine would be $\left(\bar{Y}_{1f}-\bar{Y}_{2f}\right)-\left(\bar{Y}_{1m}-\bar{Y}_{2m}\right)$. Here we use $\bar{Y}_{ij}$ to denote the sample average in each subgroup where i = 1,2, and j = f,m. If we further assume that the number of subjects in each of the four subgroups is $n/2$, we obtain $\mathrm{Var}\left[\left(\bar{Y}_{1f}-\bar{Y}_{2f}\right)-\left(\bar{Y}_{1m}-\bar{Y}_{2m}\right)\right]=8\sigma^2/n$. In this simple example, for the test of heterogeneity to have the same power as the test of the main effect, we would need a sample size four times as large as the original sample size n. In addition, the effect sizes for interactions may be smaller than those for main effects (Collins et al. 2009), which leads to even less power for detecting interactions. Furthermore, Aiken and West (1991) show that measurement error associated with the covariates can attenuate the effect size of the interaction term, with substantial reduction of the power to detect interaction. To increase the power of multiple regression to detect moderator effects, Aguinis and Gottfredson (2010) consider several design, measurement, and analysis issues, including minimizing truncation of variables and having similar number of subjects in each subgroups formed by the moderator of interest in the design stage, increasing reliability and minimizing scale coarseness in measurements, and avoiding categorization of continuous variables in the analysis.

Subgroup Analysis Reporting and Interpretation

Several papers have evaluated the quality of reporting of subgroup analyses. Assmann et al. (2000) reviewed 50 consecutive clinical-trial reports published during July to September 1997 in four major medical journals. Hernández et al. (2006) reviewed 63 cardiovascular RCTs published in 2002 and 2004. Wang et al. (2007) reviewed 97 RCTs published in *The New England Journal of Medicine* from July 2005 to June 2006. All concluded that the quality of reporting of subgroup analyses is uneven. Reporting was found to lack completeness and clarity. Papers provided few details about the number of subgroup analyses conducted and whether subgroup analyses reported were pre-specified or post-hoc. Inappropriate statistical methods were used in many cases and multiple comparison issues were generally not addressed. Information was not reported consistently across different subgroup analyses. To improve reporting, Altman et al. (2001) provided guidance to authors about subgroup analysis in the Consolidated Standards of Reporting Trials (CONSORT) statement. Leading medical journals (see, for example, Wang et al (2007)) have adopted reporting guidelines. These guidelines encourage clear and complete reporting of the number of subgroup analyses conducted and statistical methods used, and recommend consistent reporting of results from interaction tests, summary statistics, point estimates, and confidence intervals for comparisons within subgroups.

In randomized studies, subgroup analysis measures how the effect of treatment varies across levels of a secondary factor. However, because the secondary factor is usually not determined by randomization, analyses intended to estimate the effect of intervening on the secondary variable (when that is possible) must control for confounding (VanderWeele and Knol in press).

Subgroup Analysis Based on Post-Baseline factors

This article focuses primarily on subgroup analyses based on baseline variables measured in randomized clinical trials. Subgroup analysis based on post-baseline factors is subject to all the considerations we have discussed for analyses based on baseline factors, but is more complicated. Because group membership can be affected by treatment, the advantage of randomization at baseline is lost. Subjects in the two treatment arms within a subgroup defined by a post-randomization variable may no longer be comparable so that differences in outcome can be confounded by measured or unmeasured differences in subject characteristics, just as comparisons of two groups in an observational study. As a result, these analyses are more susceptible to bias.

Post-baseline factors do not satisfy the temporal criterion for moderators in the MacArthur definition (Kraemer et al. 2008) and thus cannot be moderators. Here we restrict our attention to factors that can only be obtained after receiving treatment and may potentially be affected by treatment. Factors such as age or gender, even if they are obtained post-baseline, can still be moderators. Post-baseline factors may help explain how and why treatment affects outcome and be considered as mediators (Baron and Kenny 1986; Kraemer et al. 2002, 2008). Judd and Kenny (1981) propose mediation analysis methods based on a series of regression models. MacKinnon and Dwyer (1993) describe statistical methods to assess how prevention and intervention programs achieve their effects through mediators. Robins and Greenland (1992) and Pearl (2001) provide definitions for direct and indirect effects in the causal “counterfactual” framework for mediation analysis. These definitions can be used in the presence of treatment-mediator interactions and also help to clarify the confounding assumptions. Methods for causal mediation analysis are described in Jo (2008), Sobel (2008), VanderWeele and Vansteelandt (2009, 2010), and Imai et al. (2010). Worthy of mention also is the importance of controlling for mediator-outcome confounders (Cole and Hernan 2002; Judd and Kenny 1981; Pearl 2001; Robins and Greenland 1992). To address this issue, careful thought is needed in the design stage to plan for collection of information on potential mediator-outcome confounders. In the analysis stage, sensitivity analysis techniques (VanderWeele 2010)may be applied to assess the influence of unmeasured confounders on inferences about direct and indirect effects. MacKinnon (2008) provides a detailed overview of conceptual and statistical aspects of mediation analysis accompanied with many examples to illustrate how to conduct a mediation analysis.

Summary

Moderator analyses can be very informative. Properly conducted and reported moderator analyses allow us to target interventions to specific subpopulations. Moderator effects identified through hypothesis-generating moderator analysis require future studies for validation. Careful consideration should be given to hypothesis-testing moderator analysis in the design stage to control for type I and type II errors.

When conducting moderator analyses, it is important to pick the appropriate interaction test and metric for the question of interest. Adding a product term of treatment and the candidate covariate to a regression model and examining its coefficient is a common way to assess treatment heterogeneity. This approach assesses interaction on the additive scale for linear models but on the multiplicative scale for logistic or Cox proportional hazards models. Methods for assessing interactions on the additive scale should be utilized for binary or time-to-event outcomes if additive interactions are regarded as more meaningful scientifically. GAMs and non-parametric methods can be used when assumptions for linear models may be violated.

Investigators should report moderator analyses with completeness and clarity so that readers will have sufficient information for proper interpretation of the results. For hypothesis-testing moderator analysis, we recommend use of more conservative multiple comparison adjustment methods to control for type I error. For hypothesis-generating analyses, methods that control for FDR may be used. Finally, post-baseline factors may be mediators and methods for causal mediation analysis can be applied to estimate direct and indirect effects in the presence of treatment-mediation interactions.

Acknowledgments

We dedicate this paper to our friend and colleague, Dr. Stephen W. Lagakos, who inspired the work and provided valuable insights and discussions on many aspects of subgroup analyses. We are grateful to Drs. Robert J. McMahon, David P. Mackinnon, Tyler VanderWeele, and three reviewers for their comments, which have improved the paper. This work was in part supported by grant AI24643 from the National Institutes of Health.

References

Aguinis H, Gottfredson RK. Best-practice recommendations for estimating interaction effects using moderated multiple regression. Journal of Organizational Behavior. 2010; 31:776–786. doi:10.1002/job.719.

Aiken, LS.; West, SG. Multiple regression: testing and interpreting interactions. Sage; Newbury Park, CA: 1991.

Altman DG, Andersen K. Calculating the number needed to treat for trials where the outcome is time to an event. British Medical Journal. 1999; 319:1492–1495. Retrieved from http://www.bmj.com/. [PubMed: 10582940]

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Lang T. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Annals of Internal Medicine. 2001; 134:663–694. Retrieved from http://www.annals.org/. [PubMed: 11304107]

Assmann SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interactions. Epidemiology. 1996; 7:286–290. doi:10.1097/00001648-199605000-00012. [PubMed: 8728443]

Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000; 355:1064–1069. doi:10.1016/S0140-6736(00) 02039-0. [PubMed: 10744093]

Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology. 1986; 51:1173–1182. doi:a0020761/0022-3514.51.6.1173. [PubMed: 3806354]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289–300. Retrieved from http://www.wiley.com/bw/journal.asp?ref=1369-7412&site=1.

Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, Schnitzer TJ. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. New England Journal of medicine. 2000; 343:1520–1528. doi:10.1056/NEJM200011233432103. [PubMed: 11087881]

Bonetti M, Gelber RD. A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. Statistics in Medicine. 2000; 19:2595–2609. doi: 10.1002/1097-0258(20001015)19:19<2595::AIDSIM562>3.0.CO;2-M. [PubMed: 10986536]

Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics. 2004; 5:465–481. doi:10.1093/biostatistics/kxh002. [PubMed: 15208206]

Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Chapman & Hall/CRC; Boca Raton, FL: 1998.

Byar DP. Assessing apparent treatment-covariate interactions in randomized clinical trials. Statistics in Medicine. 1985; 4:255–263. doi:10.1002/sim.4780040304. [PubMed: 4059716]

Byar DP, Green S. The choice of treatment for cancer patients based on covariate information: Application to prostate cancer. Bulletin du Cancer. 1980; 67:477–490. Retrieved from http://www.john-libbey eurotext.fr/en/revues/medecine/bdc/sommaire.md. [PubMed: 7013866]

Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2010 Advance online publication. doi: 10.1093/biostatistics/kxq060.

Cole SR, Hernan MA. Fallibility in estimating direct effects. International Journal of Epidemiology. 2002; 31:163–165. doi:10.1093/ije/31.1.163. [PubMed: 11914314]

Collins LM, Dziak JJ, Li R. Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. Psychological Methods. 2009; 14:202–224. doi:a0020761/a0015826. [PubMed: 19719358]

Cook RJ, Sackett DL. The number needed to treat: A clinically useful measure of treatment effect. BMJ. 1995; 310:452–454. Retrieved from http://www.bmj.com/. [PubMed: 7873954]

Curfman GD, Morrissey S, Drazen JM. Expression of concern: Bombardier et al., Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. N Engl J Med. 2005; 343:1520–8. 2000. New England Journal of Medicine, 353, 2813-2814. doi: 10.1056/NEJMe058314.

Follmann D, Proschan M. A multivariate test for interaction for use in clinical trials. Biometrics. 1999; 55:1151–1155. doi:10.1111/j.0006-341X.1999.01151.x. [PubMed: 11315061]

Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985; 41:361–372. doi:10.2307/2530862. [PubMed: 4027319]

Gardner F, Connell A, Trentacosta CJ, Shaw DS, Dishion TJ, Wilson MN. Moderators of outcome in a brief family-centered intervention for preventing early problem behavior. Journal of Consulting and Clinical Psychology. 2009; 77:543–553. doi:a0020761/a0015622. [PubMed: 19485594]

Halperin M, Ware JH, Byar DP, Mantel N, Brown CC, Koziol J, Green SB. Testing for interaction in an I × J × K contingency table. Biometrika. 1977; 64:271–275. doi:10.2307/2335693.

Hastie, T.; Tibshirani, R. Generalised additive models. Chapman and Hall/CRC; Boca Raton, FL: 1990.

Hernández A, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? American Heart Journal. 2006; 151:257–264. doi:10.1016/j.ahj.2005.04.020. [PubMed: 16442886]

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988; 75:800–802. doi:10.1093/biomet/75.4.800.

Holm S. A simple sequential rejective multiple test procedure. Scandinavian Journal of Statistics. 1979; 6:65–70. Retrieved from http://www.blackwellpublishing.com/journal.asp?ref=0303-6898.

Hommel G. A stagewise rejective multiple test procedure on a modified Bonferroni test. Biometrika. 1988; 75:383–386. doi:10.1093/biomet/75.2.383.

Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. Epidemiology. 1992; 3:452–456. doi:10.1097/00001648-199209000-00012. [PubMed: 1391139]

Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychological Methods. 2010; 15:309–334. doi::a0020761/a0020761. [PubMed: 20954780]

Jackson RD, LaCroix AZ, Gass M, Wallace RB, Robbins J, Lewis CE, Barad D. Calcium plus vitamin D supplementation and the risk of fractures. New England Journal of Medicine. 2006; 354:669–683. doi:10.1056/NEJMoa055218 [Erratum, N Engl J Med 2006; 354:1102]. [PubMed: 16481635]

Jo B. Causal inference in randomized experiments with mediational processes. Psychological Methods. 2008; 13:314–336. doi:a0020761/a0014207. [PubMed: 19071997]

Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. Evaluation Review. 1981; 5:602–619. doi:10.1177/0193841X8100500502.

Julius S, Nesbitt SD, Egan BM, Weber MA, Michelson EL, Kaciroti N, Schork MA. Feasibility of treating prehypertension with an angiotension-receptor blocker. New England Journal of Medicine. 2006; 354:1685–1697. doi:10.1056/NEJMoa060838. [PubMed: 16537662]

Keppel, G.; Wickens, TD. Design and analysis: A researcher's handbook. Pearson/Prentice Hall; Upper Saddle River, NJ: 2004.

Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. Drug Information Journal. 1996; 30:523–533. Retrieved from http://www.diahome.org/DIAHome/Resources/FindPublications.aspx.

Kraemer HC. Reconsidering the odds ratio as a measure of 2×2 association in a population. Statistics in Medicine. 2004; 23:257–270. doi:10.1002/sim.1714. [PubMed: 14716727]

Kraemer HC. Moderators of treatment outcomes: Clinical, research, and policy importance. Journal of the American Medical Association. 2006; 296:1–4. doi:10.1001/jama.296.10.1286.

Kraemer HC. Toward non-parametric and clinically meaningful moderators and mediators. Statistics in Medicine. 2008; 27:1679–1692. doi:10.1002/sim.3149. [PubMed: 18008395]

Kraemer HC, Wilson T, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry. 2002; 59:877–883. doi:10.1001/archpsyc.59.10.877. [PubMed: 12365874]

Kraemer HC, Kiernan M, Essex M, Kupfer DJ. How and why criteria defining moderators and mediators differ between the Baron & Kenny and the MacArthur approaches. Health Psychology. 2008; 27:S101–S108. Retrieved from http://www. apa.org/pubs/journals/hea/. [PubMed: 18377151]

Lagakos SW. The challenge of subgroup analyses—reporting without distorting. New England Journal of Medicine. 2006; 354:1667–1669. doi:10.1056/NEJMp068070. [PubMed: 16625007]

Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. Annals of Behavioral Medicine. 2003; 26:172–181. doi:10.1207/S15324796ABM2603_02. [PubMed: 14644693]

Li J, Chan IS. Detecting qualitative interactions in clinical trials: An extension of range test. Journal of Biopharmaceutical Statistics. 2006; 16:831–841. doi:10.1080/10543400600801588. [PubMed: 17146982]

Li R, Chambless L. Test for additive interaction in proportional hazards models. Annals of Epidemiology. 2007; 17:227–236. doi:10.1016/j.annepidem.2006.10.009. [PubMed: 17320789]

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychological Methods. 2002; 7:19–40. doi:a0020761/1082-989X.7.1.19. [PubMed: 11928888]

MacKinnon, DP. Introduction to statistical mediation analysis. Taylor & Francis Group; New York, NY: 2008.

MacKinnon DP, Dwyer JH. Estimating mediated effects in prevention studies. Evaluation Review. 1993; 17:144–158. doi:10.1177/0193841X9301700202.

Marra G, Radice R. Penalised regression splines: Theory and application to medical research. Statistical Methods in Medical Research. 2010; 19:107–125. doi:10.1177/0962280208096688. [PubMed: 18815162]

Meckstroth, A.; Burwick, A.; Moore, Q.; Ponza, M.; Marsh, S.; McGuirk, A.; Zhao, Z. Teaching self-sufficiency: An impact and benefit-cost analysis of a home visitation and life skills education program. 2008. Retrieved from Mathematics Policy Research website: http://www.mathematica-mpr.com/publications/pdfs/teaching_self.pdf

Newcombe RG. A deficiency of the odds ratio as a measure of effect size. Statistics in Medicine. 2006; 25:4235–4240. doi:10.1002/sim.2683. [PubMed: 16927451]

Pan G, Wolfe DA. Test for qualitative interaction of clinical significance. Statistics in Medicine. 1997; 16:1645–1652. doi:10.1002/(SICI)1097-0258(19970730)16:14<1645::AIDSIM596>3.0.CO;2-G. [PubMed: 9257418]

Patel KM, Hoel DG. A nonparametric test for interaction in factorial experiments. Journal of the American Statistical Association. 1973; 68:615–620. doi:10.2307/2284788.

Pearl, J. Direct and indirect effects. Seventeenth Conference on Uncertainty and Artificial Intelligence; San Francisco, CA: Morgan Kaufmann; 2001. p. 411-20.

Peto, R. Statistical aspects of cancer trials. In: Halnan, KE., editor. Treatment of cancer. Chapman and Hall; London, UK: 1982. p. 867-871.

Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. Statistics in Medicine. 1993; 12:1239–1248. doi:10.1002/sim.4780121105. [PubMed: 8210823]

Robins JM, Greenland S. Identifiabilty and exchangeability for direct and indirect effects. Epidemiology. 1992; 3:143–155. doi:10.1097/00001648-199203000-00013. [PubMed: 1576220]

Rothman, KJ. Modern epidemiology. Little, Brown and Company; Boston, MA: 1986.

Sackett DL. Down with odds ratios! Evidence-Based Medicine. 1996; 1:164–166. doi:10.1629/09178.

Sacks FM, Pfeffer MA, Moye LA, Rouleau JL, Rutherford JD, Cole TG, Braunwald E. The effect of Pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. New England Journal of Medicine. 1996; 335:1001–1009. doi:10.1056/NEJM199610033351401. [PubMed: 8801446]

Schemper M. Non-parametric analysis of treatment-covariate interaction in the presence of censoring. Statistics in Medicine. 1988; 7:1257–1266. doi:10.1002/sim.4780071206. [PubMed: 3231949]

Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. New England Journal of Medicine. 1999; 341:279–283. doi:10.1056/NEJM199907223410411. [PubMed: 10413743]

Shaffer JP. Multiple hypothesis testing. Annual Review of Psychology. 1995; 46:561–584. doi: 10.1146/annurev.ps.46.020195.003021.

Shuster J, van Eys J. Interaction between prognostic factors and treatment. Controlled Clinical Trials. 1983; 4:209–214. doi:10.1016/0197-2456(83)90004-1. [PubMed: 6641234]

Silvapulle MJ. Tests against qualitative interaction: Exact critical values and robust tests. Biometrics. 2001; 57:1157–1165. doi:10.1111/j.0006-341X.2001.01157.x. [PubMed: 11764256]

Simes JR. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986; 73:751–754. doi:10.1093/biomet/73.3.751.

Sleeper LA, Harrington DP. Regression splines in the Cox model with application to covariate effects in liver disease. Journal of the American Statistical Association. 1990; 85:941–949. doi: 10.2307/2289591.

Sobel ME. Identification of causal parameters in randomized studies with mediating variables. Journal of Educational and Behavioral Statistics. 2008; 33:230–251. doi:10.3102/1076998607307239.

Song S, Pepe MS. Evaluating markers for selecting a patient's treatment. Biometrics. 2004; 60:874–883. doi:10.1111/j.0006-341X.2004.00242.x. [PubMed: 15606407]

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100:9440–9445. doi:10.1073/pnas.1530509100. [PubMed: 12883005]

Tolan PH, Gorman-Smith D, Henry D, Schoney M. The benefits of booster interventions: Evidence from a family-focused prevention program. Prevention Science. 2009; 10:287–297. doi:10.1007/s11121-009-0139-8. [PubMed: 19513845]

VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology. 2010; 21:540–551. doi:10.1097/EDE.0b013e3181df191c. [PubMed: 20479643]

VanderWeele TJ, Knol MJ. The interpretation of subgroup analyses in randomized trials: Heterogeneity versus secondary interventions. Annals of Internal Medicine. in press.

VanderWeele TJ, Robins JM. Four types of effect modification: A classification based on directed acyclic graphs. Epidemiology. 2007; 18:561–568. doi:10.1097/EDE.0b013e318127181b. [PubMed: 17700242]

VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. Statistics and Its Interface. 2009; 2:457–468. Retrieved from http://www.intlpress.com/SII/.

VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis with a dichotomous outcome. American Journal of Epidemiology. 2010; 172:1339–1348. doi:10.1093/aje/kwq332. [PubMed: 21036955]

Wactawski-Wende J, Kotchen JM, Anderson GL, Assaf AR, Brunner RL, O'Sullivan MJ, Manson E. Calcium plus vitamin D supplementation and the risk of colorectal cancer. New England Journal of Medicine. 2006; 354:684–696. doi:10.1056/ NEJMoa055222. [PubMed: 16481636]

Wang R, Lagakos SW, Ware H, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. New England Journal of Medicine. 2007; 357:2189–2194. doi:10.1056/NEJMsr077003. [PubMed: 18032770]

Wen L, Badgett R, Cornell J. Number needed to treat: A descriptor for weighing therapeutic options. American Journal of Health-System Pharmacology. 2005; 62:2031–2036. doi:10.2146/ajhp040558.

Table 1

A hypothetical illustrative example for the dependence of heterogeneity of treatment effects on metric

Baseline	Mortality risk			
Performance status	**Treatment**	**Control**	**RR**	**$R_c - R_t$**
1 (low risk)	.05	.1	0.5	.05
2 (median risk)	.2	.4	0.5	.2
3 (high risk)	.35	.7	0.5	.35

Table 2

A hypothetical example to illustrate that within-subgroup comparisons can lead to misleading results

	Males Cured	**Females Cured**
Treatment	32/40	4/5
Control	16/40	2/5
	RD=0.4, *p*<0.001	RD=0.4, *p*=.52