Methods, Challenges, and Best Practices for Conducting Subgroup Analysis

OPRE REPORT #2021-17
A. BRECK AND B. WAKAR
JANUARY 2021

Policy and program evaluation research often seeks to estimate the effects of interventions across broad study populations. However, focusing on average treatment effects may obscure important variation across subgroups of the treated population. Researchers may use subgroup analysis to determine if the impact of a policy or program varies by groups and to develop a better understanding of which interventions are most effective for particular groups of participants under specific circumstances. This approach enables policymakers to efficiently allocate resources (Haegerich & Massetti, 2013; Supplee et al., 2013).

Researchers who do not conduct subgroup analysis properly may overstate or misreport differences in treatment effects across subgroups. Reviews of peer-reviewed literature that include subgroup analyses suggest there is a "credibility gap" in how subgroup analyses are often conducted and presented, and this realm of scientific literature is of "uneven" quality (Fan et al., 2019; Inglis et al., 2018; Wang & Ware, 2013). Researchers of all disciplines must consider several important points when designing, implementing, and presenting subgroup analyses to avoid these problems.

Effective and credible subgroup analysis follows several important recommendations, in addition to the aspects covered in Inglis et al. (2018) (see text box Key Features of a Well-Designed Subgroup Analysis). These best practices are relevant to researchers across social science disciplines and policy subject areas.

The goal of this brief is to provide a concise resource on the features of well-designed and implemented subgroup analysis and a set of summary recommendations. We build on the work presented at the Office of Planning, Research, and Evaluation's (OPRE) 2009 methods meeting on subgroup analysis and corresponding publications in a special issue of Prevention Science (MacKinnon et al., 2013). While this brief focuses on use of a multiple regression framework, we also provide an overview of alternative approaches to conducting subgroup analyses. In appendix A we provide examples of peer-reviewed publications that implement various approaches to conducting subgroup analysis.

RESEARCH DESIGN ELEMENTS FOR CONFIRMATORY SUBGROUP ANALYSIS

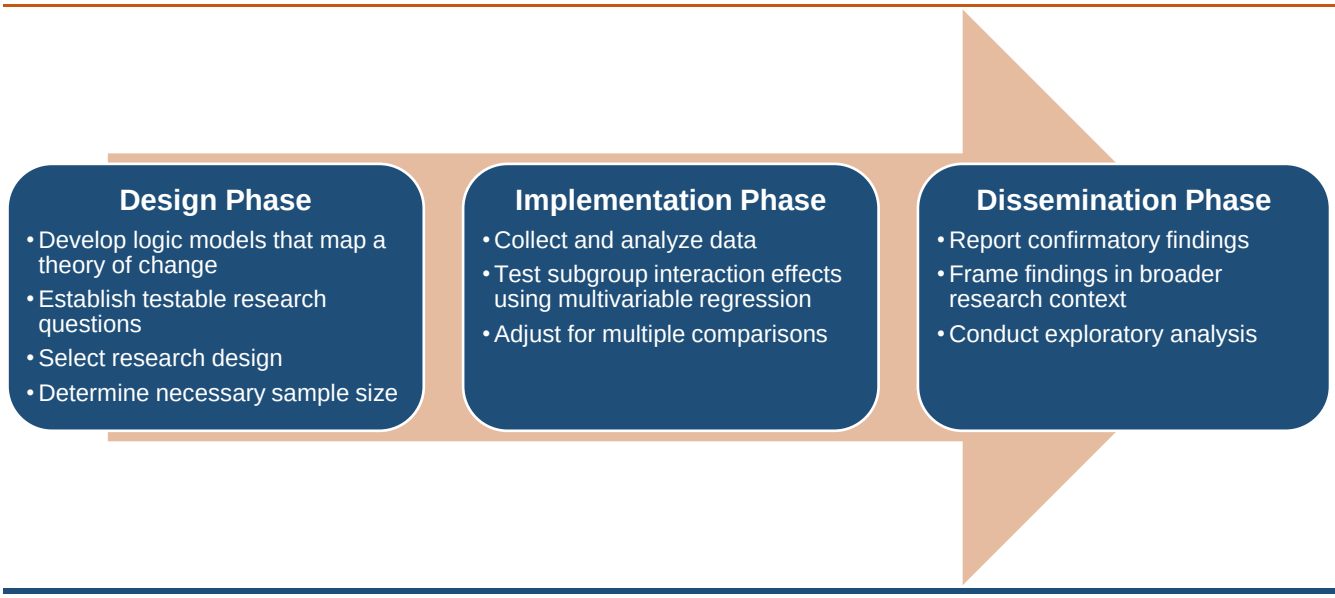
Subgroup analysis may be either *confirmatory*, such that hypotheses regarding differences are defined in advance and

tested, or *exploratory*, such that analysis proceeds without clear theory-driven hypotheses in mind (Bloom & Michalopoulos, 2013). The focus of this brief is on **confirmatory subgroup analysis**. Confirmatory analysis requires that researchers plan a research strategy to limit bias from confounding factors and enhance statistical power to estimate treatment effects. Confirmatory analysis follows from a predetermined research design and is consistent with existing theory. Results from **exploratory subgroup analysis** do not support inference about the causal effects of treatments and are not necessarily included in the research plan. Often exploratory analysis is considered and conducted only after data are collected. To the extent exploratory results are included in research dissemination products, they should be accompanied by clear caveats about the strength of their evidence (Bloom & Michalopoulos, 2013).

Figure 1 provides an overview of the important considerations during each phase in the research process.

The sections that follow describe key elements of subgroup analysis research design in further detail.

Figure 1. Important Phases of a Research Study with Subgroup Analysis



Important Research Plan Components

This section presents several critical elements researchers should consider when planning, implementing, or reviewing research plans that incorporate subgroup analysis.

Theory of change for subgroup analysis. Before developing a research plan to evaluate a policy or program, researchers should develop a logic model that maps the overall theory of change. The benefit of this process is a clear rationale for designing a research plan to examine how a treatment might differentially affect subgroups (Farrell

et al., 2013). A logic model can identify characteristics that are potential candidates for subgroup analysis by drawing attention to prespecified, measurable, moderating variables that interact with the intervention. Well-developed logic models that map the theory of change can also help inform the choice of focusing on testing of treatment effects for a subgroup versus testing for differences in treatment effects across subgroups. For example, if the primary research concern is whether a treatment is effective for a given subgroup, then the researchers would test for a zero effect among that specific group; if the primary research concern is whether a particular subgroup's effect is greater than another subgroup's, then the researchers would test for a difference between groups.

Logic models also illustrate which potential subgroups are predetermined and uncorrelated with treatment (Wang & Ware, 2013). When membership in subgroups is not predetermined or is correlated with exposure to an intervention, estimates of treatment effects may be biased. While subgroups will often be determined prior to treatment (e.g., age, race, ethnicity), it is important that membership in a **subgroup be conditionally independent of exposure to treatment**. In other words, each treated unit, regardless of membership in any given subgroup, should have equal chance of being treated. Having identified and motivated plausible subgroups, researchers may then develop research questions for confirmatory subgroup analysis.

Research questions. The research design process begins in earnest once researchers have established testable research questions. These **research questions should be clear, include a testable hypothesis, and be motivated by a theoretical rationale**. Research questions should be made explicit in the beginning of any research plan and restated early in any project reports or publications.

Research questions and hypotheses for subgroup analysis should state whether the important treatment effect comparisons are between groups and/or if there are hypothesized treatment effects for a particular subgroup. Research questions should be explicit in identifying the comparisons of interest. For example, researchers may wish to draw conclusions that compare two groups (such as whether an intervention has different effects on elementary school children than on middle school children) or multiple subgroups (such as people who live in urban, suburban, or rural settings). In either case, they should frame hypotheses in terms of the difference between those subgroups. If researchers instead wish to draw conclusions about one subgroup in isolation (e.g., to examine whether there are treatment effects specifically for elementary school children or specifically for people who live in rural areas), research questions should be framed in terms of that subgroup.

Experimental or quasi-experimental research design. Many research design considerations outlined in this memo are relevant for both experimental and quasi-experimental research. However, implementation of subgroup analysis under each design may differ. For example, in both cases, subgroups should be specified prior to conducting any analysis, and subgroup membership should be measured prior to

exposure to treatment in both designs. In a **randomized controlled trial (RCT)**, in which researchers randomly assign study participants to either a treatment group or a control group, the possibility that confounding factors bias estimates of treatment effects is low. An RCT that includes plans for conducting subgroup analyses should stratify participants to treatment by target subgroups to minimize differences in subgroup representation in the treatment and comparison groups. In a **quasi-experimental research design**, researchers are unable to randomly manipulate treatment assignment. In the absence of true randomization to treatment, researchers may choose among several analytic approaches that seek to mimic random assignment to treatment (Campbell et al., 2002). Other specific considerations unique to quasi-experimental designs appear below.

Sample selection and statistical power. Researchers should plan for a **sampling strategy that will ensure sufficient sample sizes**. In particular, small sample sizes may have too little power for subgroup analysis, which can lead to false negative results (Burke et al., 2015). As a general rule, for a grouping variable with two levels (e.g., middle school versus elementary school aged children), to detect heterogeneous subgroup effects with the same power as an overall main effect, a study needs a sample size at least four times as large (Bloom & Michalopoulos, 2013; Klerman, 2009; Wang & Ware, 2013). Statistical power to detect an effect in a subgroup varies across subgroups (Lanza & Rhoades, 2013), depending on subgroup sample size, exposure to treatment, within-group variation in treatment effects, and interclass correlation of sampled clusters. As in any research, sample size in subgroup analysis is closely tied to significance and confidence interval width, with larger samples more likely to yield significant findings resulting from narrower confidence intervals.

In either prospective or retrospective studies, researchers should pay attention to overall sample size and sample size within each subgroup. Power is enhanced for a given sample size when subgroups are equally sized (Brown et al., 2013). To this end, stratified sampling with levels of the subgroup variable as strata may be preferable to simple random sampling, especially if levels of the subgroup variable are not equally distributed in the population. For instance, a subgroup analysis of a medical intervention by baseline risk factor, where risk factor is classified as high or low, may sacrifice power if simple random sampling is used and high-risk patients represent a small proportion of the population. Stratifying the sample by risk score could help mitigate this problem. While stratified sampling is optimal for increasing power, it may not always be feasible. Other options for sample selection include convenience sampling or cluster sampling, but what these sampling methods gain in practicality they may sacrifice in the generalizability of analysis results (Groves et al., 2009).

When conducting quasi-experiments and natural experiments, the study sample size is often out of the researchers' control. Rather than conduct a power analysis to determine the required sample size for a prospective study, researchers may instead calculate the **minimum detectable effect**. This estimate indicates the smallest identifiable treatment

effect given the study sample size, treatment exposure, and variation in outcome measures across subgroups. Calculating the minimum detectable effect when the study sample size is fixed can set expectations about the limits of the study to find meaningful treatment effects. Gurgand and Rathelot (2012) provide helpful guidance on calculating minimum detectable effects. The analysis plan for the Pathways for Advancing Careers and Education project provides another helpful example (Abt Associates, 2015).

Adjustments for multiple comparisons. When determining a sampling strategy and a required sample size, it is important to accommodate the changes to statistical power caused by multiple comparisons because the number of tests affects the overall power of the study. Researchers can maximize power by distinguishing between exploratory and confirmatory findings, minimizing the number of confirmatory tests, and testing an omnibus hypothesis that considers all outcome measures and subgroups together (Bloom & Michalopoulos, 2013; Schochet, 2008). Examples of limiting the number of comparisons can be found in the analysis plan for Health Profession Opportunity Grants Impact Study (Harvill et al., 2015) and the Building Strong Families Project (Moore et al., 2012).

Analytic Approach Considerations

This section presents a range of considerations for researchers developing and implementing analytic plans for subgroup analyses.

Key Features of a Well-Designed Subgroup Analysis

- Subgroup analysis motivated with theoretical rationale
- Hypothesis stated prior to analysis
- Subgroup variable(s) measured at baseline
- Randomization to treatment stratified by subgroup
- Interaction effects tested
- Limited number of subgroup analyses
- Accounted for multiple hypothesis testing

Source: Inglis et al., 2018

Multiple regression. The standard approach to evaluating heterogeneous treatment effects across randomly assigned subgroups is to test the **interaction between a treatment and subgroup indicator** (Wang & Ware, 2013). In practice, this can be accomplished using a multiple regression framework. With this approach, each outcome is regressed on indicators for receipt of treatment, the subgroup of interest, an interaction of the treatment and subgroup variables, and any additional control variables. As an illustration, a model for testing for subgroup treatment effects on a continuous outcome may take the form of the following linear regression model:

$$Y = B_0 + B_1x + B_2t + B_3x*t + e$$

where B_3 is the coefficient of interest on the interaction of treatment and treated unit characteristic x, and e is an error term with standard properties. Drawing an example of a subgroup analysis from the OPRE evaluation of the Health Profession Opportunity Grants (Harvill et al., 2015), consider that in this case, variable x is a binary indicator for employment status, which takes the value of 0 for unemployed individuals and value of 1

for employed individuals; then B2 represents the treatment effect for unemployed individuals, and B2 + B3 represents the treatment effect for employed individuals. The test of the null hypothesis H0: B3 = 0 evaluates whether there is a difference in treatment effects between employed and unemployed people who received treatment.

Regression modeling options. The precise model for testing interaction effects can vary, depending on the form of the dependent variable. For example, for continuous outcomes, such as weight or body mass index, a linear regression model is often appropriate; for binary variables, such as an indicator for graduating high school, a logistic regression is appropriate; for time-to-event outcomes, such as with survival models, a Cox proportional hazard model is usually appropriate; and for count variables, such as the number of primary care appointments in a year, a Poisson regression is often appropriate (Mihaylova et al., 2011; Wooldridge, 2002).

Appropriate Regression Approaches for Different Types of Outcome Measures

Outcome Variable	Regression Approach
Continuous	Linear regression
Binary	Logistic regression
Time-to-event	Cox proportional hazard model
Count	Poisson regression

When examining effects among subgroups defined by categorical variables, the coefficient on the interaction term indicates the magnitude of the difference in treatment effects from a baseline group. If there are multiple subgroups to compare (e.g., across multiple racial groups), the researcher must complete additional significance testing to identify the magnitude and statistical significance of differences between groups (see section on multiple comparisons above). For example, OPRE's Supporting Healthy Marriage Evaluation examined subgroup effects across married couples of different races (both individuals were Hispanic, both African American, both White, or other or multiracial) (Hsueh et al., 2012). In this study, authors estimated the effect of program participation both within group (e.g., whether there were program effects for Hispanic couples) and across groups (e.g., whether program effects were different for Hispanic couples compared with African-American couples).

When examining subgroup effects across multiple subgroups, the estimating regression should include an indicator for each subgroup (excluding one baseline group) and separate interaction terms for each subgroup and a treatment indicator. As in the example provided above for a model with two subgroups, the coefficient on each interaction term indicates the difference in effects for the subgroup relative to the baseline group. Additional statistical tests can test the differences between groups.

If subgroups are defined based on levels of a continuous variable, differences in treatment effects can be estimated using results from the regression that include interactions between treatment exposure and continuous versions of the grouping variable. Many statistical software packages make these comparisons easy to implement by either comparing outcomes at different points of the distribution of the continuous grouping variable or by transforming the continuous

variable into discrete groups. In either case, the guidelines for multiple hypothesis testing described above still apply (Abadie & Cattaneo, 2018; Wang & Ware, 2013).

Experimental vs. Quasi-Experimental Research Designs

- Experimental research reduces bias by minimizing differences in characteristics of the treatment and control group samples
- Quasi-experimental research designs can account for selection bias using statistical tools or plausibly random exposure to treatment

Estimation approaches for quasi-experimental data. Unlike analysis of RCT data, the identification of causal treatment effects with quasi-experimental data requires estimation strategies that account for the potential confounder bias. For example, propensity score matching attempts to account for selection to treatment by creating a comparison group of study participants matched on observable characteristics to participants in the treatment group. Because selection bias at baseline may be mistaken for heterogeneous effects, researchers should perform robustness checks when using observational data (Breen et al., 2015; Robins, 1999; Sharkey & Elwert, 2011).

Examples of quasi-experimental designs that can identify causal effects when certain assumptions are met include difference-in-differences, regression discontinuity, propensity score matching, and instrumental variable regression. Numerous resources provide detailed review of these methods, including Abadie and Cattaneo (2018) and Angrist and Pischke (2009).

INTERPRETATION OF SUBGROUP ANALYSIS RESULTS

The results of a subgroup analysis must be interpreted with care. Researchers must decide whether they are interested in **differences in treatment effects between subgroups or effects for each subgroup**; that is, whether there is an expectation of differences in treatment effects across the study population or there is a subset of the population for whom the treatment effect is significant. In the former, a single statistical test (such as a t-test or one-way ANOVA) must be conducted to compare the treatment effect in each group to the effect in every other group if researchers wish to make inference that compares subgroups. In contrast, when looking for a treatment effect within a subset of the population, the treatment effect in each subgroup must be compared to zero (or no effect), and the number of tests to conduct will match the number of subgroups.

If significant treatment effects are found in one subgroup but not another, it may be tempting to conclude the subgroups are significantly different, but that would not necessarily be correct. The hypothetical example in table 1, in which the sample size for middle-school-aged children is small, shows that although the risk difference is equal for middle and elementary-school-aged children, the treatment effect is only significant for elementary-school-aged children. This example demonstrates that conclusions about differences between groups should be based only on tests for those differences.

Table 1. Hypothetical Example Demonstrating Differences in Statistical Significance for Two Subgroups of Different Sample Sizes

	Elementary-School-Aged Child Absenteeism	Middle-School-Aged Child Absenteeism
Treatment	32/40	4/5
Control	16/40	2/5
	Risk Difference = 0.4; $p < 0.001$	Risk Difference = 0.4; $p = 0.52$

Source: Wang & Ware, 2013

Researchers should also consider policy or clinical importance as well as statistical significance (Wasserstein et al., 2019). Policy or clinical importance represents a difference (between subgroup impacts, or between outcomes for treatment vs. control group) that is large enough to inform practice; for instance, a hospital-based intervention showing a savings in cost per patient large enough to motivate proliferation of the intervention technique. Ideally, policy-relevant results should be supported by statistically significant tests. However, statistical significance alone may not be enough to warrant widespread interest in intervention results. Especially when researchers have the good fortune of large sample size, small differences between groups may be statistically significant. That is, not every statistically significant finding should result in a policy-relevant recommendation.

PRESENTATION OF SUBGROUP ANALYSIS RESULTS

When presenting subgroup analyses, researchers should ground their presentation in a theoretical rationale that explains why treatment effects may vary across groups. Reports of results should include direction and magnitude of treatment effects for each subgroup, and/or differences in treatment effects across subgroups, based on hypotheses. Confidence intervals should also be shown around the subgroup treatment effects (Farrell et al., 2013).

Presentation of Subgroup Analysis Should Include:

- Direction and magnitude of treatment effects for each subgroup, and/or
- Differences in treatment effects across subgroups, when across group comparisons were pre-specified in research questions

Conclusions about heterogeneity of effects between subgroups depend on the metric used (Wang & Ware, 2013). For example, there may be homogeneity with respect to relative risk but heterogeneity when measuring absolute risk: for instance, a treatment that reduces absenteeism by half for both elementary and middle-school-aged children. However, if absenteeism is rare for one group—elementary-school-aged children, in this hypothetical example—a reduction by half may not represent a substantial absolute difference (say, from 1 percent to 0.5 percent). However, for middle-school-aged children, where the baseline absenteeism rate is higher, a reduction by half could be a substantial absolute reduction. In this example, there is evidence to support implementing this treatment for middle-school-aged children but not for elementary-school-aged children. This finding would have been masked if only relative risk had been examined. Researchers must choose metrics carefully based on the research question at hand.

CHALLENGES AND OTHER CONSIDERATIONS FOR SUBGROUP ANALYSIS

Researchers should consider several challenges when planning and implementing subgroup analysis.

Multiple hypothesis testing. Chief among the challenges of conducting subgroup analysis is multiple hypothesis testing. The probability of making a false claim for at least one of the tests conducted increases as the number of tests increases (Hill et al., 2009). If 100 tests are carried out independently and a *p*-value of 0.05 is used as a cut point for determining statistical significance, it is anticipated on average that 5 of the tests will lead to a false positive result. While running that many tests may seem unlikely at first consideration, it can easily be accomplished if multiple variables (especially variables with many levels) are used to create subgroups and subgroups are compared pairwise.

Primary Limitations of Subgroup Analysis

- Multiple hypothesis testing → False positives
- Insufficient statistical power → False negatives
- Correlated moderators → Imprecise or biased treatment effect estimates

Source: Burke et al., 2015

An easy solution to the multiple testing problem is to apply corrections to *p*-values after calculating them. The Bonferroni correction is perhaps the most well-known option. Bonferroni and similar corrections control the **family-wise error rate**, such that the probability of at least one test resulting in a false positive is set to a desired value, usually 0.05. This type of correction has the advantage of being easy to apply. However, this type of adjustment is often too conservative because it can lead to a family-wise error rate below the desired value. This is because the underlying family-wise error rate correction assumes the tests are independent, which is not the case in practice (Hill et al., 2009).

As an alternative to these types of post-calculation corrections to *p*-values, **false discovery rate** looks at the expected proportion of false positive results. If the number of significant findings is larger than the number of expected false positive results, researchers need to investigate further, but determining which test(s) to trust as true positive(s) can be difficult (Hill et al., 2009; Wang & Ware, 2013). Recommendations vary, with Wang and Ware (2013) advocating for a family-wise error rate approach for confirmatory subgroup analyses and a false discovery rate approach for exploratory subgroup analyses.

Researchers should be judicious when choosing subgroups to compare in the confirmatory analyses. The design should ensure adequate sample size and power for planned comparisons. Accounting for multiple comparisons by controlling the family-wise error rate will reduce the power of tests but is important for protecting against false positives.

Research context. Another challenge of subgroup analysis is interpreting results in a greater research context. Subgroup analysis results can have both **internal and external consistency**. That is, subgroup analysis results should corroborate other findings from

the research at hand (internal consistency) and should also be consistent with external theory and empirical findings. Any subgroup analysis that contradicts previous beliefs or findings should be viewed cautiously (Bloom & Michalopoulos, 2013).

Post hoc subgroups. As previously established, subgroup analyses should be specified as part of the analysis plan. However, this may not always be possible; for example, when subgroups arise during study implementation or the analysis phase of a project. As a case study, consider a classroom intervention in which the treatment curriculum was created in English, but some of the students received instruction in Spanish (Price & Olsen, 2013). Researchers, upon learning about language differences among classrooms, asked whether the language of instruction would alter the effect of the intervention. While this is an interesting and relevant question, because this subgroup was not anticipated in the study planning phase, results should be interpreted as exploratory.

Exploratory subgroup analysis. Researchers may use exploratory analyses to create post hoc subgroups that will serve as confirmatory subgroups in a future study. In this case, data visualization may guide the creation of the subgroups. For instance, researchers could examine histograms or boxplots of an outcome variable by potential grouping variables to see if the distribution of the outcome varies by subgroup. Scatter plots for each proposed subgroup between independent variables and regression residuals may also be useful; if the slope of the relationship between these variables is different for different levels of the grouping variable, this indicates potentially interesting heterogeneous relationships (Morgenthaler, 2009).

Preregistration. Researchers may consider preregistering their study plans before beginning to conduct subgroup analyses. Preregistering in a public repository commits researchers to a specific design, hypothesis, and/or data analysis plan. This precommitment provides incentive for researchers to implement their original data collection and analysis decisions and increases transparency and rigor by documenting discrepancies between planned and actual study decisions. The precommitment also reduces any tendency for researchers to falsify or unwittingly manipulate their research to support a specific hypothesis.

For additional information on preregistration processes, see materials from OPRE's 2019 methods meeting (Corker, 2019).

OTHER APPROACHES AND RECENT METHODOLOGICAL DEVELOPMENTS IN SUBGROUP ANALYSIS

Several alternative approaches have been employed to address limitations of a traditional multivariable regression approach to identify heterogeneous treatment effects across subgroups. A brief overview of a few of these options follows.

Meta-analysis. Using results from multiple studies, in a meta-analysis framework, is one approach that addresses limitations of limited

power and limited generalizability from a single study. In meta-analysis, weighted averages of treatment effects from multiple similar studies yield more precise treatment effect estimates. Brown et al. (2013) provide a thorough review of this approach. Lipsey (2003) provides a helpful review of constraints to the approach.

Bayesian methods. One approach to reduce the likelihood of obtaining false positives as a result of multiple hypothesis testing is to limit the number of subgroups examined in any given study. Authors of a recent paper provide **guidelines for selecting subgroups based on prior probabilities** of finding any subgroup effects (Burke et al., 2015). Using prior probabilities and Bayes rule, the authors recommend limiting subgroup analysis to no more than one or two subgroups, each of which should have at least 20 percent, or preferably at least 50 percent, prior probability of having subgroup effects. Burke et al. also recommend that even in studies with large sample sizes and a corresponding high degree of power, all subgroup analyses should be justified *a priori*.

For additional information on Bayesian methods, see materials from OPRE's 2017 methods meeting (OPRE, 2019b).

Another approach to reducing the likelihood of false positive results when conducting subgroup analysis is to use a **hierarchical shrinkage estimator**. The standard (e.g., frequentist) approach to analyzing subgroup treatment effects involves separate significance tests for treatment effects of each subgroup. In contrast, the hierarchical shrinkage estimator incorporates information about likely treatment effects, based on an assumption about the prior distribution of mean average treatment effects and a standard deviation of treatment effects across subgroups. This approach is implemented in a multilevel model framework, which allows point estimates to vary by group without increasing uncertainty relative to the overall treatment effect estimates. Jennifer Hill's presentation at the OPRE 2009 meeting and articles by Henderson et al. (2016), Hill et al. (2009), and Pennello and Rothmann (2019) provide additional detail on implementation of this approach for subgroup analysis.

Machine learning. Recent developments in machine learning research provide methods of estimating treatment effects across subgroups. These approaches are particularly useful in application of *post hoc* subgroup analysis. These data-based methods for identifying subgroups are not subject to the usual concerns about data mining or multiple hypothesis testing and can identify subgroups with maximum treatment effect heterogeneity.

Examination of possible effects among subgroups not explicitly identified during the research design phase of a project increases the likelihood of identifying false positive results. Best practice is to avoid presentation of any *post hoc* analysis results as confirmatory analysis. However, researchers building on methods of **supervised machine learning** have proposed methods for *post hoc* subgroup analysis of randomized control study data that are not subject to concerns related to multiple hypothesis testing (Athey & Imbens, 2015).

Another machine learning application, called "**causal forests,**" is an effective tool for estimating subgroup treatment effects (Knittel & Stolper, 2019). The primary advantage of the causal forest approach is that it classifies subgroups by maximizing differences in estimated treatment effects between each of the groups. A researcher can pursue this approach without an *a priori* expectation or theory to motivate the subgroup analysis. Like the approach proposed by Athey and Imbens (2015), researchers can use a causal forest to estimate heterogeneous treatment effects without concerns for multiple hypothesis testing.

SUMMARY RECOMMENDATIONS

Subgroup analyses to identify heterogeneous or subgroup specific treatment effects are important efforts useful for informing policy creation and program design to efficiently allocate resources. Effective and credible subgroup analysis follows several important recommendations:

- Be explicit about whether subgroup analyses are confirmatory versus exploratory.
- Motivate subgroup analysis with theoretical rationale and prior research.
- Determine subgroups and subgroup comparisons early in the design process.
- State research questions and hypotheses prior to collecting any data or conducting any analysis.
- Calculate minimum necessary sample size or minimum detectable effect during the design stage.
- Measure subgroup variable at baseline.
- Stratify randomization to treatment by subgroup.
- Test interaction effects of subgroups and treatment.
- Limit the number of subgroup analyses and adjust for multiple comparisons.

See appendix A for additional resources.

REFERENCES

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*(1), 465–503.

Abt Associates. (2015). *Pathways for Advancing Careers and Education (PACE)*. Technical supplement to the evaluation design report: Impact analysis plan, 1–43.

Angrist, J., & Pischke, S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Athey, S., & Imbens. G. (2015). *Machine learning methods for estimating heterogeneous causal effects*. https://pdfs.semanticscholar.org/86ce/004214845a1683d59b64c4363a067d342cac.pdf

Bloom, H. S., & Michalopoulos, C. (2013). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects on subgroups. *Prevention Science 14*, 179–188.

Breen, R., Choi, S., & Holm, A. (2015). Heterogeneous causal effects and sample selection bias. *Sociological Science 2*, 351–369.

Brown, C. H., Sloboda, Z., Faggiano, F., Teasdale, B., Keller, F., Burkhart, G., Vigna-Taglianti, F., Howe, G., Masyn, K., Wang, W., Muthén, B., Stephens, P., Grey, S., & Perrino, T. (2013). Methods for synthesizing findings on moderation effects across multiple randomized trials. *Prevention Science 14*(2), 144–156.

Burke, J. F., Sussman, J. B., Kent, D. M., & Hayward, R. A. (2015). Three simple rules to ensure reasonably credible subgroup analyses. *BMJ 351*.

Campbell, D. T., Cook, T. D., & Shadish, W. R. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin, and Company.

Corker, K. (2019). *Pre-registration: What & why*. Grand Valley State University. https://opremethodsmeeting.org/wp-content/uploads/2019/10/Pre-registration_Corker.pdf

Fan, J., Song, F., & Bachmann, M. O. (2019). Justification and reporting of subgroup analyses were lacking or inadequate in randomized controlled trials. *Journal of Clinical Epidemiology 108*, 17–25.

Farrell, A. D., Henry, D. B., & Bettencourt, A. (2013). Methodological challenges examining subgroup differences: Examples from universal school-based youth violence prevention trials. *Prevention Science 14*(2), 121–133.

Groves, R. M., Fowler, Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. 2nd ed. Wiley.

Gurgand, M., & Rathelot, R. (2012). *Power calculation*. J-PAL Advanced course.

Haegerich, T. M., & Massetti, G. M. (2013). Commentary on subgroup analysis in intervention research: Opportunities for the public health approach to violence prevention. *Prevention Science 14*(2), 193–198.

Harvill, E., L., Moulton, S., & Peck, L. R. (2015). *Health Profession Opportunity Grants impact study technical supplement to the evaluation design report: Impact analysis plan*. Technical supplement. OPRE Report No. 2015- 80. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Henderson, N. C., Louis, T. A., Wang, C., & Varadhan, R. (2016). Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Services and Outcomes Research Methodology 16*(4), 213–233.

Hill, J., Gelman, A., & Yajima, M. (2009). Why we (usually) don't need to worry about multiple comparisons. *Interagency Meeting on Subgroup Analysis*.

Hsueh, J., Alderson, D. P., Lundquist, E., Michaloopoulos, C., Gubits, D., Fein, D., & Knox, V. (2012). *The Suppporting Healthy Marriage evaluation: Early impacts on low-income families*.

Inglis, G., Archibald, D., Doi, L., Laird, Y., Malden, S., Marryat, L., McAteer, J., Pringle, J., & Frank, J. (2018). Credibility of subgroup analyses by socioeconomic status in public health intervention evaluations: An underappreciated problem? *SSM Population Health 6*, 245–251.

Klerman, J. (2009). Subgroups analysis: A view from the trenches. *Interagency Meeting on Subgroup Analysis*.

Knittel, C. R., & Stolper, S. (2019). *Using machine learning to target treatment: The case of household energy use* (NBER Working Paper 26531). National Bureau of Economic Research, Inc.

Lanza, S. T., & Rhoades, B. L. (2013). Latet class analysis: An alternative on subgroup analysis in prevention and treatment. *Prevention Science 14*(2), 157–168.

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *Ann. Am. Acad. Pol. Soc. Sci. 587*(1), 69–81. http://journals.sagepub.com/doi/10.1177/0002716202250791

MacKinnon, D., Supplee, L. H., Kelly, B. C., & Barofsky, M. Y. (2013). Special issue: Subgroup analysis in prevention and intervention research. *Prevention Science 14*, 107–110. http://link.springer.com/journal/11121/14/2

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics 8*, 897–916.

Moore, Q., Wood, R. G., Clarkwest, A., Killewald, A., & Monahan, S. (2012). *The long-term effects of building strong families: A relationship skills education program for unmarried parents.* Technical Supplement. Mathematica Policy Research.

Morgenthaler, S. (2009). Exploratory data analysis. *WIREs Computational Statistics 1*, 33–43.

OPRE (Office of Planning, Research, and Evaluation). (2019a). *Interagency subgroup analysis meeting*. https://opremethodsmeeting.org/meetings/2009/

OPRE. (2019b). *Bayesian methods for social policy research and evaluation*. https://opremethodsmeeting.org/meetings/2017/

Pennello, G., & Rothmann, M. (2019). Bayesian subgroup analysis with hierarchical models. *Biopharmaceutical applied statistics symposium: Vol. 2, biostatistical analysis of clinical trials.* Springer.

Price, C., & Olsen, R. (2013). Post-hoc subgroups. *Interagency Meeting on Subgroup Analysis*.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese 121*(1/2), 151–79.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations of educational interventions* (NCEE 2008-4018). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Sharkey, P., & Elwert, F. (2011). The legacy of disadvantage: Multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology 116*(6), 1934–1981.

Supplee, L. H., Kelly, B. C., MacKinnon, D. M., & Barofsky, M. Y. (2013). Introduction to the special issue: Subgroup analysis in prevention and intervention research. *Prevention Science 14*(2), 107–110.

Wang, R., & Ware, J. H. (2013). Detecting moderator effects using subgroup analyses. *Prevention Science 14*(2), 111–20.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond 'p < 0.05.' *American Statistician 73*(sup1), 1–19.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. 2nd ed. The MIT Press.

This brief was prepared by Insight Policy Research (1901 North Moore Street, Suite 1100, Arlington, VA 22209) under Contract Number HHSP233201500109I. The Administration for Children and Families' Contracting Officer's Representatives are Emily Ball Jabbour and Kriti Jain. The Insight Project Director is Rachel Holzwart, and the Deputy Project Director is Hilary Wagner.

APPENDIX A. RELEVANT RESOURCES

Table A.1. Select Open Access Resources Available From Peer-Reviewed Journals and OPRE

Title	First Author	Year
Pathways for Advancing Careers and Education (PACE), technical supplement to the evaluation design report: Impact analysis plan	Abt Associates	2015
Subgroups analysis when treatment and moderators are time-varying	Almirall	2013
Time-varying effect moderation using the structural nested mean model: Estimation using inverse-weighted regression with residuals	Almirall	2014
Methods for synthesizing findings on moderation effects across multiple randomized trials	Brown	2013
Three simple rules to ensure reasonably credible subgroup analyses	Burke	2015
Reporting of heterogeneity of treatment effect in cohort studies: A review of the literature	Dahan	2018
Justification and reporting of subgroup analyses were lacking or inadequate in randomized controlled trials	Fan	2019
Do the effects of a relationship education program vary for different types of couples? Exploratory subgroup analysis in the Supporting Healthy Marriage evaluation	Gubits	2014
Health Profession Opportunity Grants (HPOG) impact study technical supplement to the evaluation design report: Impact analysis plan	Harvill	2015
Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research	Henderson	2016
Credibility of subgroup analyses by socioeconomic status in public health intervention evaluations: An underappreciated problem?	Inglis	2018
Working toward wellness: Telephone care management for Medicaid recipients with depression, eighteen months after random assignment	Kim	2010
Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment	Lanza	2013
The long- term effects of building strong families: A relationship skills education program for unmarried parents	Moore	2012
Introduction to the special issue: Subgroup analysis in prevention and intervention research	Supplee	2013
Subgroup analyses in confirmatory clinical trials: Time to be specific about their purposes	Tanniou	2016
Detecting moderator effects using subgroup analyses	Wang	2013
Estimating moderated causal effects with time-varying treatments and time-varying moderators: Structural nested mean models and regression with residuals	Wodtke	2017

Note: This table includes a selection of available open-access resources.