Subgroup Analysis: "What Works Best for Whom and Why?"

16

Ferdinand Keller

Introduction

The chapter starts with an introductory example using main results from two large randomized trials to evaluate substance use prevention programs. Basic questions are explored such as: Is the program equally effective for boys and girls, or is it effective for baseline users of alcohol although no overall beneficial effect could be confirmed? The next section looks at how subgroups can be defined and introduces the distinction between manifest (= directly observable) and latent (= not directly observable) variables. Then, statistical approaches for conducting subgroup analyses are presented. The focus will be on mainly newer methods taking into account the multilevel structure of data, mediation and moderation approaches, and the testing of the interaction effect as gold standard in biostatistics. Special emphasis is given to models using latent variables such as latent class analysis (LCA) and growth mixture models (GMM). Exploratory subgroup analysis has been enhanced considerably by applying these so-called mixture models (LCA and GMM are just two specific methods of the family of mixture models). They help in identifying potential differences in outcome that might exist in a population, and to estimate treatment effects for previously unknown subgroups.

Despite this pool of advantageous new methods, some basic (intrinsic) risks in subgroup analyses remain. Two major issues for the appraisal of subgroup findings are introduced: (a) is there an overall significant effect in the trial, and (b) is the subgroup analysis preplanned (= confirmatory analysis) or use primarily for exploratory purposes. These subjects set the framework for a proper interpretation of subgroup results. In particular, the problem of finding false-positive results arises, but, conversely, it may also falsely be concluded that an intervention is not effective in a subgroup (false-negative result). Some examples from the literature are given to illustrate potential pitfalls. Finally, strategies for dealing with the risks and limitations of subgroup analysis are discussed (i.e., meta-analysis, statistical adjustment of error rates, and some recent methods), and some agreed-upon recommendations for reporting of results are provided.

Why Subgroup Analysis?

Subgroup analysis can help in detecting differential response to an intervention and is often used to evaluate the effectiveness for specific subgroups. Consider as an illustrating example the results of two large randomized trials that were designed to evaluate the effectiveness of

F. Keller (✉)
Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital Ulm, Ulm, Germany
e-mail: Ferdinand.keller@uniklinik-ulm.de

Z. Sloboda et al. (eds.), *Prevention of Substance Use*, Advances in Prevention Science,
https://doi.org/10.1007/978-3-030-00627-3_16

universal school-based substance abuse prevention programs with comparable preventive interventions applied to same-aged populations. One is the U.S. Adolescent Substance Abuse Prevention Study (ASAPS) (Sloboda et al., 2009) and the other the EU-DAP study (EUropean Drug Addiction Prevention trial) (Faggiano et al., 2010). The overall findings of these two interventions varied across programs. A full summary of the results is beyond the scope of this chapter, but there were differences with regard to alcohol use that may serve as an initial focus for the present topic. In the 18-month follow-up of EU-DAP, persisting beneficial program effects were found for episodes of drunkenness (Faggiano et al., 2010). In ASAPS follow-up, no beneficial effects on alcohol use were found (Sloboda et al., 2009). Several questions arise consequently for further analyses: Is there a beneficial effect for a specific subgroup within the ASAPS sample (despite the missing overall effect), e.g., for baseline users of alcohol? For EU-DAP: Is the (overall significant) intervention also effective in specific subgroups, e.g., in male and female students alike?

More generally, Bloom and Michalopoulos (2013) propose three types of research questions that may motivate subgroup analyses:

– how widespread are the effects of an intervention?
– is the intervention effective for a specific subgroup?
– is the intervention effective for any subgroup?

Definition and Types of Subgroups

Subgroup analysis is usually defined as an analysis in which the intervention effect is evaluated in a defined subset of the participants in a trial, or in complementary subsets, such as by sex or in age categories. Subgroups can be characterized by manifest (= directly observable) or latent (= not directly observable) variables.

In application to prevention research, subgroups can be defined in many different ways and Bloom and Michalopoulos (2013) suggest defining subgroups in terms of several characteristics:

- Demographic variables (age, gender, educational background, etc.)
- Risk factors (past smoking, drinking, drug abuse, etc.)
- Current health status or severity of a problem/disease which is to be treated by the intervention

In larger studies, subgroups may also be built according to geographic location or site (county, state; hospital, school). More recently, new kinds of variables are available for statistical analyses, in particular genetic and epigenetic predictors (Latendresse, Musci, & Maher, 2018). It should be emphasized that subgroup analyses should not be based on all variables that are available in the data set, but should be motivated by the underlying theory of change of the intervention program. The theory should also provide guidance to determine factors that explain variation in responsiveness to the intervention as well as moderators and mediators of impact.

Characteristics like those listed above are considered directly observable and they are called manifest variables in statistical terminology. Many characteristics are, however, not directly observable, but are inferred from indicators such as items of questionnaires or by other types of assessment instruments. Examples are ample in the social sciences, e.g., personality factors or intelligence components are considered to be latent constructs. Examples in prevention science are that not everyone involved in a targeted intervention responds equally to the intervention due to a (unknown) combination of variables (Nylund-Gibson & Hart, 2014), or a persons' attitude towards alcohol or drug use. Such variables are termed latent variables. Both manifest and latent variables are often used to model heterogeneity, i.e., to explain quantitative or qualitative differences in a population. Understanding the heterogeneity among individuals within a targeted population, or, vice versa, uncovering the way individuals are similar, ultimately provides the opportunity to understand

outcomes and to design better treatment measures and intervention efforts (Nylund-Gibson & Hart, 2014).

Latent subgroups may also be defined longitudinally, i.e., by the responsiveness to an intervention or by trajectories in outcome across the observation period. Examples are the course of aggressive behavior across school grades (Petras, Masyn, & Ialongo, 2011) or the degree of delinquent behavior during adolescence (Jones & Nagin, 2007). These "definitions," however, are based on probabilistic assignment of individuals to their most likely class and emerge only during the study. Since group membership is not known at baseline and, therefore, stratified randomization of treatment assignment to the subgroups is not possible, this type of subgroup is usually not included in "pure" subgroup analysis recommendations. Nonetheless, heterogeneity in the developmental course and subgroup differences can be hypothesized and used for confirmatory analyses of the trial.

Statistical Approaches for Conducting Subgroup Analysis

Subgroup Analysis with Manifest Variables

For the analysis of subgroups defined by manifest variables, several statistical approaches have been proposed. In a simplifying manner, two main approaches could be distinguished: (1) hierarchical (or multilevel) linear models for longitudinal designs and (2) the mediation and moderation approach. Both model families are discussed only briefly below, since they cover a wide range of potential models and an extensive introduction is beyond the scope of this chapter. Furthermore, mediational models are addressed in a special chapter in this book (O'Rourke and MacKinnon). Finally, (3) the addition of interaction terms to the statistical model in question as the recommended method in biostatistics is introduced and discussed.

1. Hierarchical (or multilevel) linear models are often applied in the social sciences. They correct for clustering (e.g., students nested in classes, classes nested in schools, or, in the longitudinal case, observations within persons and with explaining covariates added) and provide correct *p*-values for this type of nested data. They also overcome some limitations of "classical," well-known techniques such as repeated measures ANOVA, in allowing for missing data and unequal time spaces between observations (Hox, 2010; Singer & Willett, 2003; Verbeke & Molenberghs, 2000).
2. Another well-known and applied approach is mediation and moderation analysis. Fairchild and MacKinnon (2014) in their introduction to these methods target the same question as the title of this chapter when they discuss these models "with the ultimate goals of identifying the active ingredients of these programs and to address the question what works for whom under what conditions" (p. 538). Advantages of the mediation-moderation approach are its potential to inform about the effectiveness of program components and thus to refine curriculum development and implementation strategies. Fairchild and MacKinnon (2014) provide a comprehensive introduction into the mediation model and the moderation model, and also their combination. For example, they found in the evaluation of a worksite wellness program that outcome was moderated by part-time versus full-time work status. A mediation model was then used to explain this difference, and it could be shown that full-time workers were getting more exposure to program-related social norms at the work place, contributing to their larger program effect. If mediators are also measured repeatedly during a trial, they can be incorporated in various types of longitudinal structural equation mediation models to determine the active components of a program. Goldsmith et al. (2017) provide a tutorial how to fit and interpret various longitudinal mediation models, based on a trial of rehabili-

tative treatments for chronic fatigue syndrome as a motivating example. Wang and Ware (2013) also show the opportunities of moderator analyses in detecting subgroup effects. Schochet, Puma, and Deke (2014) provide a formal introduction into subgroup analysis within the regression context and Cordova et al. (2014) give a conceptual overview over statistical models that aim to identify those pathways through which prevention interventions work.

3. In biostatistics, there is agreement that the appropriate way to examine whether a treatment effect differs between subgroups is to test for an interaction effect between treatment and subgroup (Brookes et al., 2004; Rothwell, 2005; Schulz, Altman, Moher, & CONSORT Group, 2010). (In the social sciences, the question of interest whether the treatment effect varies among the levels of a baseline factor is often referred to as moderator analysis). Separate analyses of the treatment effect within each subgroup are not recommended since such multiple comparisons increase the risk of obtaining false-positive results. Conversely, subgroup-specific comparisons result in smaller data sets and thus reduced power to detect a true treatment effect (false-negative finding).
   The test of the interaction effect revealed to be quite reliable; simulation studies have shown that the interaction test performed well (Brookes et al., 2001). When there was no true overall treatment effect, the percentage of false-positive overall tests remained at 5%; in the presence of a true overall effect, the percentage of tests that were (correctly) significant reflected the power of the data set (Brookes et al., 2004). These authors also show how power goes down in subgroup analyses. Regarding power of the interaction test, a trial with 80% power for the overall effect had only 29% power to detect an interaction effect of the same magnitude. For interactions of this size to be detected with the same power as the overall effect, sample sizes need to be inflated fourfold (Brookes et al., 2004). Given this lack of power for the interaction test in the analysis of a trial (that is usually powered only for the main effect), failure to find a significant interaction does not show that the treatment effect seen overall applies to all individuals (Wang & Ware, 2013).

Subgroup Analysis with Latent Variables

If one is interested in detecting unknown subpopulations defined by a *set* of indicators within the study sample who respond differently to the intervention, identification of subpopulations based on mixture models is well suited. The basic idea behind mixture modeling lies in assuming that the observed values of variables (e.g., means, frequencies in cross-tables, regression coefficients, trajectories) are not the same for all persons in the sample, but are different for subgroups within the sample. In other words, and narrowed down to the case of latent class analysis (LCA), one assumes that the overall population heterogeneity with respect to a set of manifest (categorical) variables results from the existence of two or more distinct homogeneous subgroups, or latent classes, of individuals (Masyn, 2013). Over the last two decades, several variations of mixture modeling have been developed, and the models can be grouped according to whether the latent variable is considered categorical or continuous, and whether analysis of a cross-sectional or a longitudinal design is intended (c.f. Muthén, 2002; Nylund-Gibson & Hart, 2014).

Most applications of these mixture models in prevention science seem to use a categorical latent variable to describe population heterogeneity. An example of LCA is provided by Lanza and Rhoades (2013, see below in Section "Recent Strategies"). Conventional regression analysis can be made more flexible by regression mixture analysis where latent classes in the data can be identified and regression parameter estimates can vary between latent classes. Van Horn et al.

(2009) use regression mixture analysis to capture differential effects of family resources on children's academic outcomes and Ding (2006) provides a worked-through example of this method where differential relationships between children's math achievement, children's math self-concept, and teacher's rating are analyzed.

LCA can be extended to the longitudinal case, called latent transition analysis (LTA—e.g., Collins & Lanza, 2010). In longitudinal studies with continuous outcome variables, especially with more than three assessment points, it is favorable to identify latent classes with the latent class growth model (LCGM) proposed by Nagin (Jones & Nagin, 2007; Nagin, 1999) or in a more general form, the so-called growth mixture models (GMM—Muthén and Muthén, 2000; Pickles & Croudace, 2010). GMM are conducted to estimate the number of latent classes with the same trajectory, the size of the latent classes, and to attribute individuals to these trajectory classes which are characterized by different courses over time. For example, Petras et al. (2011) examined the impact of two universal preventive interventions in first grade on the growth of aggressive/disruptive behavior in grades 1–3 and 6–12. They modeled growth trajectories for each of the two time periods separately, and then associated the latent trajectory classes of aggressive/disruptive behavior across the two time periods using a latent transition model. Subsequently, it was tested whether the interventions had direct effects on trajectory class membership in the two time periods and whether the interventions affected the transition between periods. One of the findings was that males in the intervention condition were significantly more likely than control males to transition from the high trajectory class in grades 1–3 to a low class in grades 6–12.

A challenge of these methods lies in the problem that the number of latent classes is unknown and must be estimated by comparing various statistical criteria such as goodness of fit and information criteria (Petras & Masyn, 2010; Wright & Hallquist, 2014; Muthén, 2003). The trajectory groups cannot be prespecified (and are therefore not known at baseline), but it is usually attempted to relate the latent classes that emerge in the GMM to baseline characteristics or consequences of change, e.g., relate the course of aggressive behavior trajectories in school to records of violent and criminal behavior as young adults (cf. Petras & Masyn, 2010). An excellent introduction with applications in Mplus syntax (Muthén and Muthén, 1998–2012) is given in Jung and Wickrama (2008).

The mixture model approach is mostly used in an exploratory manner and seems especially promising in prevention science since most subgroup analyses are conducted for universal intervention programs. It helps to gain more information on heterogeneity in the sample and to transfer and integrate the findings into substantive theories. The cost for making use of these very flexible methods is that they are (primarily) data-driven and hypotheses based on the findings should be subjected to further testing. There has also been extended discussion about how to find the "correct" number of latent classes and whether the classes represent "real" entities or more statistical artifacts (see Masyn, 2013; Muthén, 2003). Unfortunately, some of these issues cannot be solved by means of replication since a new sample will give a similar distribution with similar ambiguities about the characteristics of the population distribution (Petras & Masyn, 2010).

In principle, approaches like hierarchical (or multilevel) linear models and especially moderator/mediator models deal with relations (covariance) between *variables* and are called variable-oriented, while LCA/GMM deal with *individuals*, called person-oriented approach. Both look at the same data matrix (one on the "columns," the other on the "rows") and are equivalent, but have their advantages depending on the research question (Masyn, 2013; Muthén & Muthén, 2000). Advantage of the person-oriented approach is the identification of previously unknown groups of persons (latent classes) which is usually not possible in the variable-oriented approach (the distinguishing combination(s) of moderator variables had to be known).

Risks and Limitations of Subgroup Analysis

The second part of this chapter details some risks and problems that arise when applying and interpreting subgroup analysis. Let us refer back to the questions from the introductory example, e.g., it was asked whether the intervention in EU-DAP was effective for boys and for girls. Indeed, the effectiveness of the program was examined according to sex, and a significant association between the program and a lower prevalence of all behavioral outcomes was found among boys, but not among girls (Vigna-Taglianti et al., 2009). The researchers state as a limitation that there was not enough power in the study for subgroup analyses, which had an impact on the precision of the estimates. Thus, it may be likely that no significant effect was found for a specific subgroup (here: females), because there was not sufficient statistical power to detect the effect, and it is falsely assumed that this subgroup received no benefit from the intervention. This type of error is called false-negative or (in statistics) type II error. On the other hand, testing for subgroup differences in the ASAPS study might reveal a significant effect for a specific subgroup, but it may be a statistical artifact caused by performing many statistical tests and thus increasing the chance of finding a (spurious) significant effect. This type of error is called false-positive or type I error. Furthermore, many statisticians would question the validity of such post-hoc subgroup differences in the absence of an overall significant effect (here: no significant overall effect on alcohol use in ASAPS).

More generally, proper interpretation of subgroup differences demands consideration of various prerequisites, in particular the number of statistical tests performed, whether they are testing preplanned hypotheses or are exploratory, and whether the intervention effect is significant in the full sample of the trial.[1]

In case of a positive overall effect in a study, further subgroup analysis is justified and can be used to detect differential response to an intervention. The general research question then is "Do the treatment effects vary among the levels of a baseline factor?" (Wang, Lagakos, Ware, Hunter, & Drazen, 2007, p. 2189), e.g., for males and females, for different ethnicities, or for varying levels of illness at baseline. However, as indicated above, in these applications of subgroup analysis there is the risk of false-negative results.

In the case where no overall effect is found in a study, the situation gets more complicated. Since usually much time, effort, and money have been invested in conducting large prevention program studies with randomized control groups or quasi-experimental designs, the question arises whether the tested program is effective for specific subgroups within the study population (although there is no significant effect on the overall study population). In general, statisticians would reject these further analyses (except for conducting exploratory analyses that have to be confirmed in future studies) and would call this approach as "rescuing a failed trial" or "exercises in pure data dredging." Applied scientists, on the other hand, may argue that a difference in effectiveness for subgroups is valid if there are good reasons to explain the difference. Prevention scientists/practitioners may argue as well, based on their experience while planning and conducting the prevention programs, that a subgroup difference may be valid. Unfortunately, almost all subgroup differences seem explainable post-hoc, and there are numerous examples where these effects turned out later to be false-positive (see the example from biotech research below).

Besides the question whether there is a significant overall effect in the trial, another distinction is important for statistical analysis and interpretation of subgroup findings: were the analyses

[1] A special situation arises in some universal prevention trials where it is not expected to find an overall effect, but only for a specific subgroup. For technical and/or ethical reasons, however, it is not possible to apply targeted prevention to this subgroup. For example, Petras et al. (2011) evaluated the program Good Behavior Game in school classes and expected that the impact on aggressive behavior was concentrated among high aggressive boys. Usually, though, overall effects are reported in universal prevention, and the effect sizes of the full trial are included in meta-analysis.

exploratory or confirmatory? Confirmatory analyses provide an appropriate basis to assess how strongly the study's prespecified central hypotheses are supported by the data. Exploratory analyses, on the other hand, examine relationships within the data to identify outcomes or subgroups for which impacts may exist. The goal of these exploratory analyses is to generate hypotheses that could be subject to more rigorous future examination. Overall, the strength of evidence based on confirmatory findings is higher than that based on exploratory findings, and this difference should be made clear to one's reader (Bloom & Michalopoulos, 2013).

Biostatisticians have especially criticized that exploratory analyses testing many subgroup differences increase the risk of false-positive results and may produce spurious findings. This problem is known under different names, e.g., alpha-error inflation, multiple testing problem, or as multiplicity in biomedical guidelines. Most statistical textbooks provide a formal treatment of the problem of multiple testing. The following excurse is based on Schochet (2008).

For example, a difference between two treatment groups is to be explored, and a *t*-test is applied for testing the significance of the difference. Suppose that the null hypothesis is true for each test and that the tests are independent. Then, the chance of finding at least one spurious impact is $1 - (1 - \text{alpha}^N)$, where alpha is the percentage of type I errors and N is the number of tests, e.g., if several outcomes or, equivalently, subgroups are tested. If the alpha error is set at 5%, the probability of making at least one type I error is 10% if two tests are conducted, and 23% if five tests and 40% if ten tests are conducted.

Thus, the more subgroup analyses are performed the higher the chance to find significant subgroup differences. Therefore, guidelines have been developed for statistical analyses in pharmacological trials as well as recommendations for interpreting and reporting estimates of intervention effects for subgroups of a study sample. These guidelines have become very strict and it is unlikely that any conclusion of treatment efficacy based solely on exploratory subgroup analyses would be accepted in the absence of a significant overall effect (EMA—ICH E9, 2006). However, there is also the risk of false-negative results in subgroup analysis, i.e., the finding that a particular subgroup does not benefit from an intervention program or gets even worse. Such findings may also be chance findings or a consequence of low power to detect true effects.

The examples presented in the next section show some false-positive as well as false-negative findings that were from minor up to major importance. Because no good examples from substance use research seem available, they come from medical science. Furthermore, problems with post-hoc findings in subgroups have been recognized much earlier in medical science, in particular in pharmacological treatment studies, than in prevention research. Therefore, exploratory findings in, e.g., cardiology have meanwhile been subject to replications, and it could be determined whether reproducibility could be achieved. Several elaborated reviews of these results have been compiled, biostatisticians have developed consensus on the process and requirements of statistical analysis, and finally guidelines have been published for planning and presenting the results of investigations (see below).

Example: subgroups with false-positive finding

Differentiation according to the severity of illness is a common practice in doing exploratory analyses of trials (especially if there is no overall significant effect), e.g., one is interested in whether the intervention is effective at an early stage of the disease or at an advanced stage or in both. Major erroneous findings seem not to exist in prevention science, at least they are not referenced in respective articles. Therefore, a striking example from biotech research where personal and financial consequences have been dramatic may illustrate the potential danger of a post-hoc subgroup interpretation that was prematurely communicated as a scientific result and turned out later to be false-positive. The following summary is based on an article by David Brown in the *Washington Post (September 23,* 2013*);* c.f. also Hodgson (2016).

The biotech company InterMune sought approval to market its drug for a more common

ailment, idiopathic pulmonary fibrosis (IPF). In all, 330 patients were randomly assigned to get either interferon gamma-1b or placebo injections. Disease progression or death occurred in 46 percent of those on the drug and 52 percent of those on placebo. That was not a significant difference ($p = 0.08$). However, when looking into subgroups it turned out that people with mild to moderate cases of the disease had a dramatic difference in survival: only 5% of those taking the drug died, compared with 16% of those on placebo. The p-value was 0.004.

The company announced in a press release that the drug *"Reduces Mortality by 70% in Patients with Mild to Moderate Disease."* This statement had severe consequences for the CEO (6 months of home confinement and partial exclusion from working).

InterMune run another trial (planned sample: 826 patients at 81 hospitals) in order to maximize the chance of getting clear-cut results. It enrolled only people with mild to moderate lung damage. And it failed. A little more than a year into the study, more people on the drug had died (15%) than people on placebo (13%).

Besides the personal consequences for the CEO, the more interesting thing for science is that the findings of exploratory subgroup analyses (i.e., a positive treatment effect in mild/moderate illness) should be clearly distinguished from confirmed results. The example also underscores the importance of replication studies.

Examples: subgroups with no or negative finding

Rothwell (2005) warns that we must also be cautious in focusing on subgroups with an apparent neutral or negative trend. As mentioned above, the correct statistical analysis is not to test the significance of the treatment effect in every subgroup, but whether the effect differs between the subgroups, i.e., the interaction effect treatment × subgroup has to be examined.

The following examples taken from Rothwell (2005) illustrate complications on various levels of interpretability of the findings:

1. In a trial on the treatment of severe stenosis, carotid endarterectomy was significantly beneficial. A subgroup analysis according to day of birth revealed that there was no significant effect for patients born on the weekend and on Tuesday and Thursday. Significant effects emerge for Monday, Wednesday, and Friday. These differences in effectiveness were due to chance; there was no subgroup x treatment effect interaction ($p = 0.83$).
2. In a large trial on the effectiveness of Aspirin vs. Placebo in acute myocardial infarction, the study result was highly significant in favor of Aspirin ($p < 0.0001$). In subsequent subgroup analyses, the zodiac signs of the patients were considered and Aspirin was ineffective in patients born under zodiac signs of Libra and Gemini, but was beneficial in all other zodiac signs. The subgroup treatment effect interaction seems $p = 0.01$ (estimated by Rothwell), but there is no explanation of this result (Libra and Gemini are not adjacent on the Zodiac) and Rothwell concludes that a more appropriate test of the interaction effect would "undoubtedly be nonsignificant" (Rothwell, 2005, p. 182).
3. However, Rothwell provides further examples where highly significant interaction effects occur by chance indicating that some subgroups have no benefit. One comes from the stenosis trial explained above, where different benefits were observed according to month of birth of the patient (interaction $p < 0.001$), but the differences could not be explained by any other plausible variable.
4. While these examples are more or less curious and had no practical consequences for treatment decisions, others were more damaging. Rothwell (2005) reports the observation in a large Canadian study in the 1970s that aspirin was effective in preventing stroke and death in men but not in women (interaction $p = 0.003$). Thus, women were considered not to benefit from aspirin and were undertreated for at least a decade, until subsequent studies and meta-analyses showed effectiveness in both groups.

These examples have shown that some of the differential results can easily be falsified if the correct statistical test (= test of interaction

effect) is applied (example 1). Others are more difficult to reject, but finally will be rejected, usually because there is no rational explanation for a subgroup finding (example 2), and even others like the gender difference in the effectiveness of aspirin (example 4) can only be overcome by replication in subsequent trials and by combining their outcomes in meta-analyses. Thus, the best test of the validity of subgroup-specific effects is reproducibility in other trials, since interaction effects may yield spurious results because of alpha error (examples 3 and 4).

Risk-Benefit Considerations

Beyond the methodological and statistical problems in determining the effectiveness of a program, a risk not to be neglected is the potential harm of prevention programs. For example, Sloboda et al. (2009) found moderate iatrogenic effects for the subgroup of baseline nonusers of alcohol in the ASAPS study.

Usually, prevention interventions are not considered to be harmful, at least in the context of universal prevention programs (in selected intervention programs, there is the risk of labeling and stigmatization). However, there are hints that iatrogenic effects emerge in universal substance prevention programs. Another example for negative consequences caused by a prevention program is the evaluation of the National Youth Anti-Drug Media Campaign (1998–2004) in the USA (Hornik, Jacobsohn, Orwin, Piesse, & Kalton, 2008). The campaign followed three large, nationally representative cohorts of adolescents over four time-points. The evaluation results revealed that the campaign had no overall effect on marijuana use or other outcome variables. Furthermore, there were hints for pro-marijuana effects in time-lagged analyses, i.e., unfavorable lagged exposure effects. Based on these results and further analyses of the campaign, Burkhart and Simon (2015) discuss the important ethical concern that an increasing intention to use cannabis (and even actual use) occurred in some subgroups that previously had little interest in the drug. The analysis found evidence that these effects were due to an increase in the perceived popularity and prevalence of marijuana use through the campaign. Mass media campaigns may have iatrogenic effects—by increasing normative beliefs, resulting in higher intentions to use (Burkhart & Simon, 2015).

In addition to the problem of actual harm, there is the general problem that use of an ineffective treatment can be highly detrimental if this prevents the use of a more effective alternative (Rothwell, 2005). Faggiano, Giannotta, and Allara (2014) provide further examples of unexpected or counterintuitive effects in prevention research and some possible explanations.

Strategies against Chance Findings

Replication and Meta-Analysis

There is general agreement that the best test of validity of subgroup-treatment effect interactions is not significance but reproducibility in other trials (Rothwell, 2005; or, more generally, Cohen, 1994). In prevention science, replication studies to confirm findings are also considered an important scientific principle for improving our knowledge. In the first "standards of evidence" in prevention science provided by Flay and colleagues in 2005 it was recognized that exact replication in which the same intervention is tested on a new sample from the same population, delivered in the same way to the same kinds of people with the same training as in the original study, is rare (Gottfredson et al., 2015, p. 908). However, almost a contradiction, replication studies are much more likely to be for the purpose of testing *variations* in the intervention or of generalizing results to *different* settings or populations than for ruling out chance findings (Gottfredson et al., 2015).

If a sufficient number of studies on a topic are available, meta-analysis is a promising way to see patterns of effects for subpopulations across trials. Borenstein and Higgins (2013) recommend the use of meta-analysis because it allows the researcher to compare the treatment effect in

different subgroups, even if these subgroups appear in separate studies. They also discuss several statistical issues related to this procedure (e.g., selection of a statistical model, statistical power for the comparison). Concerning the field of cardiovascular disease prevention and treatment, Rao et al. (2017) made a recent statement on the methodological standards for meta-analyses. Their paper also outlines some emerging methods, specifically network analysis (i.e.: test and relate several treatment conditions which have not been tested in the same trial) or Bayes methods which permit the incorporation of evidence from a variety of sources and prior knowledge.

Other statistical methods for pooling results have been proposed as well. Brown et al. (2013) present three data-sharing strategies for combining information across trials. Besides the standard meta-analysis with no sharing of data, they discuss the integrative data analysis for moderator effects where (in contrast to traditional meta-analysis) all the individual level data are combined into one dataset. The third strategy uses parallel data analysis where each of the respective trial research groups conduct analysis on their own data, following standardized analysis protocols. Results of these analyses done in parallel are then combined into a synthesis. Brown et al. (2013) conclude that the last two methods, integrative data analysis and parallel data analyses, share advantages over traditional methods available in meta-analysis.

Finally, suffice to say, results of this accumulation of empirical knowledge by these data analytic strategies should be viewed in parallel with substantive theory development and theoretically grounded research questions to move those results to a confirmatory framework and to design subsequent studies accordingly.

Statistical Techniques

In a specific trial or study, however, interpretation has to be based on currently available empirical results. Several statistical solutions have been proposed to protect against false-positive subgroup findings. Probably the most popular approach is Bonferroni correction where the level of significance is adjusted to the number of tests conducted. However, this approach yields conservative bounds on type I error and, hence, has low power (Schochet, 2008). This author (based on meetings by a 13-member Expert Advisory Panel) offers an overview of some modified and sometimes more powerful versions of the Bonferroni method and discusses advantages and limitations (c.f. also Bloom & Michalopoulos, 2013; Wang & Ware, 2013). In particular, strategies for dealing with multiplicity must strike a reasonable balance between testing rigor, i.e., to adjust downward the alpha level, and statistical power, i.e., the chance of finding truly effective interventions in subgroups (Schochet, 2008).

In addition to computing such formal adjustments, there may be cases where the overall picture seems straightforward. In the study on the effects of an antidrug media campaign on adolescents, Hornik et al. (2008) performed 80 subgroup analyses in the final set of analyses, and they found 20 significant effects, with 19 of those in a pro-marijuana direction. Thus, they conclude that there is "an overriding pattern of unfavorable lagged exposure effects" (p. 2232). In contrast, only three of 80 (= 3.7%) subgroup analyses revealed significant effects for contemporaneous associations and they were therefore considered as chance findings.

More generally, Bloom and Michalopoulos (2013) propose four main approaches to minimize the risk of revealing spuriously significant results due to multiple hypothesis testing:

1. Distinguish between confirmatory and explanatory findings
2. Minimize the number of confirmatory hypothesis tests
3. Create an omnibus hypothesis test
4. Make adjustments to multiple tests

Recent Strategies

Other strategies beyond "simple subgroup testing" have been proposed and used as well. In

many medical publications, variables that are identified in previous research or in hypothesis-generating analyses are combined into a composite index. Patients are categorized according to a "risk score" based on their profile considering multiple prognostic or predictive characteristics.

In psychometrics, it is well known that unidimensionality of scores must be confirmed, e.g., by confirmatory factor analysis (CFA) or even by testing the strict assumptions of the Rasch model in the item response theory (IRT) context. In addition, there might be higher-order interactions in variables used for subgrouping which are not captured by these analyses. Therefore, it seems preferable to make less demanding assumptions for establishing sum scores and use qualitative differences between groups of persons. A well-elaborated approach to find previously unknown classes of persons on the basis of several categorical characteristics and combinations thereof is latent class analysis (LCA—see Nylund-Gibson and Hart (2014) for a comprehensive introduction into LCA in prevention science, and Masyn (2013) for a general overview).

The LCA strategy to reduce the risk of many tests was proposed and applied by Lanza and Rhoades (2013) in a prevention context. They used six variables with binary coding each (e.g., household poverty, single-parent status, peer alcohol use) and applied LCA to identify a small set of underlying subgroups characterized by multiple dimensions, which may differ in their response to treatment. The LCA revealed five latent subgroups that represent key patterns: Low Risk, Peer Risk, Economic Risk, Household and Peer Risk, and Multi-Contextual Risk. A comparison of these five subgroups concerning outcome is feasible, while a combination of the six variables would have led to $2^6 = 64$ different subgroups. A similar approach was taken by Bühler, Seemüller, and Läge (2014) where initial illness severity was not taken as a sum score but LCA was conducted to identify different types of depression on the symptom level and treat them as separate groups in the longitudinal analysis. Instead of reducing the number of response patterns by latent variables, the identification of "types" (and "anti-types") has also been proposed on the manifest level by means of configuration frequency analysis (c.f. Stemmler, 2014).

It should be added that many have commented on the dangers of subgroup analysis (Foster, Taylor, & Ruberg, 2011), but there has been little serious investigation of methodologies for proper identification of subgroups other than the above-mentioned statistical adjustments for alpha error. Foster et al. propose a method, referred to as "virtual twins," that involves predicting response probabilities for treatment and control "twins" for each subject. The difference in these probabilities is then used as the outcome in a classification or regression tree, which can potentially include any set of the covariates. Another recent proposition is to use a Bayesian approach for identifying patient subgroups within the subgroup of patients that showed positive treatment effects (Schnell, Tang, Offen, & Carlin, 2016). The authors propose a *credible subgroup* method to identify two bounding subgroups for the benefiting subgroup: one for which it is likely that all members simultaneously have a treatment effect exceeding a specified threshold, and another for which it is likely that no members do.

Finally, yet importantly, it should be emphasized that drawing valid conclusions regarding subgroups is an issue to be addressed at the planning stage. Stratified randomization of treatment assignment might be considered to ensure sufficient representation in the subgroups of interest (Wang & Ware, 2013).

Recommendations for Reporting Subgroup Findings

In general, incomplete reporting of the interventions tested and the methods used for conducting a trial has often been a problem in scientific reporting, and therefore, numerous guidelines across different fields have been proposed. One of the best known for reporting parallel group randomized trials is the Consolidated Standards of Reporting Trials (CONSORT—Schulz et al., 2010). The CONSORT guideline was developed

by biomedical researchers and is therefore not broad enough to cover all aspects relevant for reporting in prevention science (Gottfredson et al., 2015). A new CONSORT extension for randomized controlled trials in social and psychological research (CONSORT—SPI) has been announced, but has not yet been released.

Independently from these extensions, standards for reporting are quite comparable in their main requests. CONSORT (Schulz et al., 2010, Table 1) demand as information concerning ancillary analyses when reporting a randomized trial: "Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory." Gottfredson et al. (2015, p. 909) follow CONSORT in stating: "…should include the elements identified in …CONSORT… or extension of these guidelines" (p. 908). In addition, results must be reported for every targeted outcome that has been measured in an efficacy study, regardless of whether they are positive, nonsignificant, or negative.

Specifically for "subgroup issues," recommendations are analogous and follow the same conventions. Rothwell (2005, p. 177) proposes that "all subgroup analyses that were done should be reported—i.e., not only the number of subgroup variables but also the number of different outcomes analysed by subgroup, different lengths of follow-up etc." Wang et al. (2007, p. 2193) recommend (among other points) the following:

- present subgroup results in the abstract only if the subgroup analyses were based on a primary study outcome, if they were prespecified, and if they were interpreted in light of the totality of prespecified subgroup analyses undertaken.
- avoid overinterpretation of subgroup differences. Be properly cautious in appraising their credibility, acknowledge the limitations, and provide supporting or contradictory data from other studies, if any.

With regard to prevention science, nonetheless, there are still challenges around reporting and interpreting subgroup findings, and there was no consensus around a number of critical issues in the expert meeting (Supplee, Kelly, MacKinnon, & Yoches Barofsky, 2013).

Conclusions

This chapter intended to give a broad conceptual introduction into the current status of subgroup analysis. It aimed at presenting the many opportunities provided by recently developed statistical approaches for subgroup analysis, be it confirmatory or exploratory, but also presents the potential risks of subgroup analysis.

The scientific background for this chapter is guided by placing an emphasis on methodological principles and the consequences of increasing regulatory constraints demanded by federal agencies like the Food and Drug Administration in the United States or the European Medicines Agency, in reaction to publication bias concerning study results, and in-transparent and selective reporting of significant outcome differences. These requirements are helpful for the evaluation of effectiveness and efficacy within a regulatory framework.

On the other hand, statistical concerns about mining the data may have been overemphasized and may present barriers to progress in understanding the effects of interventions. Furthermore, the prominence of adhering to the *p*-value as the definite criterion for decision-making seems sometimes too arbitrary or overly rigid (besides the widely observed misunderstanding and misuse of statistical inference). That issue was criticized not only by social scientists (e.g., Cohen, 1994) but also by statisticians themselves over the past few decades (see the statement of the American Statistical Association (Wasserstein & Lazar, 2016)).

In conclusion, many advanced statistical techniques are available. However as emphasized often in this chapter, there is a need for the development of strong theories in prevention science that would guide subgroup analyses that need to be considered during any study's planning phase. Thus, confirmatory tests are not conducted enough during exploratory research. However, it is recommended that all the new methods be used

in an exploratory way to increase knowledge, but their findings should be distinguished clearly from confirmatory results and ALL exploratory findings should be reported, in order to bring them finally (via pooling of results with meta-analysis or integrated data analysis) to a confirmatory framework. Proper inference requires full reporting and transparency (Wasserstein & Lazar, 2016). In a single trial, the limitations of subgroup analysis should be acknowledged.

Acknowledgements I gratefully acknowledge thoughtful comments and suggestions provided by Hanno Petras, Zili Sloboda and Anke de Haan on an earlier version of this chapter.

References

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

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science, 14*, 134–143.

Brookes, S. T., Whitley, E., Egger, M., Davey Smith, G., Mulheran, P. A., & Peters, T. J. (2004). Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology, 57*, 229–236.

Brookes, S. T., Whitley, E., Peters, T. J., Mulheran, P. A., Egger, M., & Davey Smith, G. (2001). Subgroup analyses in randomised controlled trials: Quantifying the risks of false-positives and false-negatives. *Health Technology Assessment, 5*, 1–56.

Brown, C. H., Sloboda, Z., Faggiano, F., Teasdale, B., Keller, F., Burkhart, G., … the Prevention Science and Methodology Group. (2013). Methods for synthesizing findings on moderation effects across multiple randomized trials. *Prevention Science, 14*, 144–156.

Brown, D. (2013, September 23). The press-release conviction of a biotech CEO and its impact on scientific research. *Washington Post*.

Bühler, J., Seemüller, F., & Läge, D. (2014). The predictive power of subgroups: An empirical approach to identify depressive symptom patterns that predict response to treatment. *Journal of Affective Disorders, 163*, 81–87.

Burkhart, G., & Simon, R. (2015). Prevention strategies and basics. In N. el-Guebaly et al. (Eds.), *Textbook of addiction treatment: International perspectives* (pp. 115–141). Milan: Springer.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.

Cordova, D., Estrada, Y., Malcolm, S. N., Huang, S., Brown, C. H., Pantin, H., & Prado, G. (2014). Prevention science: An epidemiological approach. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science* (pp. 1–23). New York, NY: Springer.

Ding, C.S. (2006). Using regression mixture analysis in educational research. *Practical Assessment, Research & Evaluation*, *11*(11). Retrieved February 2, 2018, from http://pareonline.net/getvn.asp?v=11&n=11

European Medicines Agency. (2006). ICH Topic E 9 Statistical Principles for Clinical Trials. Retrieved February 1, 2018, from http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf

Faggiano, F., Giannotta, F., & Allara, E. (2014). Strengthening prevention science to ensure effectiveness of intervention in practice: Setting up an international agenda. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science* (pp. 597–613). New York, NY: Springer.

Faggiano, F., Vigna-Taglianti, F., Burkhart, G., Bohrn, K., Cuomo, L., Gregori, D., …, Galanti, M.R. & the EU-Dap Study Group. (2010). The effectiveness of a school-based substance abuse prevention program: 18-month follow-up of the EU-dap cluster randomized controlled trial. Drug and Alcohol Dependence, 108, 56–64.

Fairchild, A. J., & MacKinnon, D. P. (2014). Using mediation and moderation analysis to enhance prevention research. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science* (pp. 537–555). New York, NY: Springer.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine, 30*, 2867–2880.

Goldsmith, K. A., MacKinnon, D. P., Chalder, T., White, P. D., Sharpe, M., & Pickles, A. (2017). Tutorial: The practical application of longitudinal structural equation mediation models in clinical trials. *Psychological Methods, 23*, 191–207.

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science, 16*, 893–926.

Hodgson, J. (2016). When biotech goes bad. *Nature Biotechnology, 14*, 284–291.

Hornik, R., Jacobsohn, L., Orwin, R., Piesse, A., & Kalton, G. (2008). Effects of the national youth anti-drug media campaign on youths. *American Journal of Public Health, 98*, 2229–2236.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Jones, B. L., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and a SAS procedure for

estimating them. *Sociological Methods Research, 35*, 542–571.

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317.

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

Latendresse, S. J., Musci, R., & Maher, B. S. (2018). Critical issues in the inclusion of genetic and epigenetic information in prevention and intervention trials. *Prevention Science, 19*, 58–67.

Masyn, K. (2013). Latent class analysis and finite mixture modeling. In T. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Statistical analysis) (Vol. 2, pp. 551–611). New York, NY: Oxford University Press.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81–117.

Muthén, B. O. (2003). Statistical and substantive checking in growth mixture modeling. *Psychological Methods, 8*, 369–377.

Muthén, B. O., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*, 882–891.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods, 4*, 139–157.

Nylund-Gibson, K., & Hart, S. H. (2014). Latent class analysis in prevention science. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science* (pp. 493–511). New York, NY: Springer.

Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 69–100). New York, NY: Springer.

Petras, H., Masyn, K., & Ialongo, N. (2011). The developmental impact of two first grade preventive interventions on aggressive/disruptive behavior in childhood and adolescence: An application of Latent Transition Growth Mixture Modeling. *Prevention Science, 12*, 300–313.

Pickles, A., & Croudace, T. (2010). Latent mixture models for multivariate and longitudinal outcomes. *Statistical Methods in Medical Research, 19*, 271–289.

Rao, G., Lopez-Jimenez, F., Boyd, J., D'Amico, F., Durant, N. H., Hlatky, M. A., … Wessel, J. (2017). Methodological standards for meta-analyses and qualitative systematic reviews of cardiac prevention and treatment studies: A scientific statement from the American Heart Association. *Circulation, 136*, e172–e194.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *Lancet, 365*, 176–186.

Schnell, P. M., Tang, Q., Offen, W. W., & Carlin, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics, 72*, 1026–1036.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved February 2, 2018, from http://ncee.ed.gov

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved February 1, 2018, from http://ies.ed.gov/ncee/edlabs

Schulz, K.F., Altman, D.G., Moher, D., & CONSORT Group. (2010). *CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials*. Retrieved February 1, 2018, from http://www.consort-statement.org/downloads/consort-statement

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York, NY: Oxford University Press.

Sloboda, Z., Stephens, R. C., Stephens, P. C., Grey, S. F., Teasdale, B., Hawthorne, R. D., … Marquette, J. F. (2009). The adolescent substance abuse prevention study: A randomized field trial of a universal substance abuse prevention program. *Drug and Alcohol Dependence, 102*, 1–10.

Stemmler, M. (2014). *Person-centered methods: Configural frequency analysis (CFA) and other methods for the analysis of contingency tables*. Heidelberg: Springer.

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

Van Horn, M. L., Jaki, T., Masyn, K., Ramey, S. L., Smith, J. A., & Antaramian, S. (2009). Assessing differential effects: Applying regression mixture models to identify variations in the influence of family resources on academic achievement. *Developmental Psychology, 45*(5), 1298–1313.

Verbeke, G., & Molenberghs, M. (2000). *Linear mixed models for longitudinal data* (2nd ed.). New York: Springer.

Vigna-Taglianti, F., Vadrucci, S., Faggiano, F., Burkhart, G., Siliquini, R., Galanti, M. R., & EU-Dap Study Group. (2009). Is universal prevention against youths' substance misuse really universal? Gender specific effects in the EU-Dap school-based prevention trial.

Journal of Epidemiology and Community Health, 63, 722–728.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine: Reporting of subgroup analyses in clinical trials. *New England Journal of Medicine, 357*, 2189–2194.

Wang, R., & Ware, J. H. (2013). Detecting moderator effects using subgroup analysis. *Prevention Science, 14*, 111–120.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*, 129–133.

Wright, A. G. C., & Hallquist, M. N. (2014). Mixture modeling methods for the assessment of normal and abnormal personality, part II: Longitudinal models. *Journal of Personality Assessment, 96*, 269–282.