Chapter 3. Estimation and Reporting of Heterogeneity of Treatment Effects

Ravi Varadhan, Ph.D.
Johns Hopkins University School of Medicine, Baltimore, MD

John D. Seeger, Pharm.D., Dr.P.H.
Harvard Medical School and Brigham and Women's Hospital, Boston, MA

Abstract

Patient populations within a research study are heterogeneous. That is, they embody characteristics that vary between individuals, such as age, sex, disease etiology and severity, presence of comorbidities, concomitant exposures, and genetic variants. These varying patient characteristics can potentially modify the effect of a treatment on outcomes. Despite the presence of this heterogeneity, many studies estimate an average treatment effect (ATE) that implicitly assumes a similar treatment effect across heterogeneous patient characteristics. While this assumption may be warranted for some treatments, for others the treatment effect within subgroups may vary considerably from the ATE. This treatment effect heterogeneity may arise from an underlying causal mechanism or may be due to artifacts of measurements or methods (e.g., chance, bias, or confounding). Heterogeneity of treatment effect (HTE) is the nonrandom, explainable variability in the direction and magnitude of treatment effects for individuals within a population. The main goals of HTE analysis are to estimate treatment effects in clinically relevant subgroups and to predict whether an individual might benefit from a treatment. Subgroup analysis is the most common analytic approach for examining HTE. Selection of subgroups should be based on mechanism and plausibility (including clinical judgment), taking into account prior knowledge of treatment effect modifiers. This chapter focuses on defining and describing HTE and offers guidance on how to evaluate and report such heterogeneous effects using subgroup analysis. Understanding HTE is critical for decisions that are based on knowing how well a treatment is likely to work for an individual or group of similar individuals, and is relevant to most stakeholders, including patients, clinicians, and policymakers. The chapter concludes with a checklist of key considerations for discussion of HTE and for addressing planned subgroup analysis in an observational comparative effectiveness research (CER) protocol.

"If it were not for the great variability between individuals, medicine might as well be a science, not an art" (William Osler, 1892).

Introduction

Randomized controlled trials (RCTs) and observational studies of comparative effectiveness usually report an average treatment effect (ATE), even though experience suggests that the same treatment can have varying impacts in different people. The clinical experience and expectation that differences in patient prognostic characteristics will lead to heterogeneous responses to therapy is mainly why medicine is as much an art as it is science. Yet, studies tend to emphasize a single measure of the impact of treatment, the ATE, which is a summary of individual treatment effects (which cannot be examined directly without making untestable assumptions). Variation is often undesirable in studies and is reduced by excluding people with characteristics that are thought to cause variations in responses to treatment. This intentional restriction in patient heterogeneity within RCTs contributes to their limited generalizability. Determining whether a treatment works for people in a target population that differs from the study population requires additional information and methods.[1]

Heterogeneity of Treatment Effect

All studies have variability in the data. Random variability is generally not a concern because it is uncorrelated with explanatory variables and can be handled well with statistical approaches for quantifying uncertainty. We focus on the nonrandom variability in treatment effects that can be attributed to patient factors. We define HTE as nonrandom variability in the direction or magnitude of a treatment effect, in which the effect is measured using clinical outcomes (either a clinical event such as myocardial infarction or a change in a continuous clinical measure such as level of pain).[2]

Understanding HTE is critical for decisions that are based on knowing how well a treatment is likely to work for an individual or group of similar individuals, and is relevant to stakeholders including patients, clinicians, and policymakers. It also has implications for applicability to individual patients (personalized medicine) of findings from pragmatic trials and observational comparative effectiveness research (CER). Pragmatic trials are large and simple experiments on treatments, with broad eligibility criteria, from which evidence is expected to be generalizable. While these designs incorporate heterogeneity in the risk of outcome among the subjects, they may also lead to HTE for the treatments that are applied. These studies may be more likely to yield null ATE than efficacy trials, where stricter inclusion criteria produce relatively homogeneous study populations. Therefore, understanding major sources of variations in treatment response is essential. For a formal general definition of HTE, see Box 3.1.

There are numerous cases in which the effectiveness of specific therapies may be heterogeneous. For example, children may respond differently to therapy via different response to treatment or to aspects of dosing that are not realized. Older adults may have worse outcomes from surgeries and devices as well as more drug side effects or drug-drug interactions so that therapies may be less effective. Individuals with multiple conditions may be on several therapies that interfere with the new treatment (or each other), resulting in a different treatment effect in these patients. Genes may also influence response to therapy; since genetic differences (differences in allele frequencies) may cluster by race or ethnicity, these characteristics may represent proxies for genetic differences that are more difficult to measure directly.

Box 3.1. Formal definition of HTE

Let an individual or a targeted subgroup with specific levels of characteristics be denoted by i. Let z stand for treatment at two levels {1, 2}; for example, being given aspirin (z =1) or not (z =2). Let the potential outcomes corresponding to the two treatment levels be denoted as $\{Y_i(1), Y_i(2)\}$. The individual treatment effect can be defined as the contrast: $\theta_i = g(E[Y_i(1)]) - g(E[Y_i(0)])$. The potential outcomes Y_i can be continuous, categorical, or binary. When Y_i is binary, $E[Y_i(z)]$ denotes $prob(Y_i = 1)$ under treatment z. The function g(.) can be identity, log, or logit. For the absolute risk model, the individual treatment effect is $\theta_i = Prob(Y_i(2)=1) - Prob(Y_i(1)=1)$. For the relative risk model, $\theta_i = \log [prob(Y_i(2)=1)] - \log[prob(Y_i(1)=1)]$. Individual variability of treatment effect occurs if the variance $(\theta_i) > 0$. Group variability (HTE) occurs if the variance of individual treatment effect is nonrandom (i.e. correlated with explanatory variables) so that $\theta_{subgroup1}$ (average θ_i for a subgroup defined by level 1 of an explanatory variable) $\neq \theta_{subgroup2}$ (average θ_i for a subgroup defined by level 2 of an explanatory variable). When this variability encompasses treatment effects of different directions, i.e., both benefit and harm, this is sometimes called a qualitative treatment interaction, whereas differences in the magnitude of treatment effect in the same direction are called quantitative interactions.

Treatment Effect Modification

If two or more exposure variables act in concert to cause disease, we will observe that the effect of exposure on outcome (treatment effect) differs according to the level of the other factor(s). A number of terms have been used to describe this phenomenon, including "joint" effects, "synergism," "antagonism," "interaction," "effect

modification," and "effect measure modification." Where effect modification exists, sound inferences will require accounting for factors that modify the effect of the exposure of primary interest. Accounting for this HTE may be required even when the variable that modifies treatment effect is not a risk factor for the outcome in the untreated group (e.g., a receptor that determines how a drug is metabolized).

Four perspectives have been advanced on the concept of interaction and the relevance of the effect modification in terms of its implication:[3]

Biological perspective: This perspective is that the interaction elucidates how factors act at the biological (mechanistic) level. The implications of this perspective are that the interaction is a representation of an underlying causal structure. Example: The finding that hypertension and smoking have a greater than additive effect on heart attack risk is a representation of some underlying biological processes that may enhance our understanding of heart attack etiology.

Statistical perspective: This perspective is that the interactions represent nonrandom variability in data unaccounted for by a model that contains only first-order terms (main effects). The implication is that the model needs to be reformulated to more accurately reflect the data. Example: A differently structured model will appropriately account for the underlying variability in the data on hypertension, smoking, and heart attack risk.

Public health perspective: This perspective is that the interactions represent a departure from additivity and highlight populations (subgroups) in which an intervention can be expected to have particularly beneficial effects. Example: The finding that hypertension and smoking have a greater than additive effect on heart attack risk suggests that limited public health resources might be most efficiently directed at patients who have hypertension and are smokers.

The individual decisionmaking perspective: This perspective is that the interactions represent a departure from additivity so that combined effects in an individual are greater than their sum. Example: Someone with hypertension can reduce heart attack risk even more by quitting smoking than someone with normal blood pressure.

Since an effect modifier changes the magnitude or direction of the association under study, different study populations may yield different results concerning the association of interest. Therefore, HTE is often suggested as a reason for differences in findings across studies. If two studies include people with different characteristics and the effect of the treatment is different in the portion of the population that differs between the studies, then HTE is a plausible explanation of the difference. Furthermore, HTE can be an explanation of differences in treatment effect between interventional and observational studies, since observational studies often include patients with different characteristics than interventional studies. Such a hypothesis might be addressed through reweighting subgroup effects according to prevalence (standardization) across studies.

Unlike potential confounders, modifying variables cannot create the appearance of an association (for exposed vs. unexposed) where none exists. But the proportion of the study population that has a greater susceptibility will influence the strength of the association. Therefore, to achieve comparability across studies, it is necessary to control for the effect of the modifying variables, generally by carrying out a separate analysis at each level of the modifier.

Additionally, the different strength of association between the exposure and outcome within strata of the effect modifier may lead to a need to be more precise in the measurement and specification of the exposure variable (such as more clearly within strata of the effect modifier).

Goals of HTE Analysis

There are two main goals of HTE analyses: (1) to estimate treatment effects in clinically relevant subgroups (subgroup analysis) and (2) to predict whether an individual might benefit from a treatment (predictive learning).[2] The first goal of HTE is highlighted in the definition of CER) proposed by the Congressional Budget Office: "An analysis of comparative effectiveness is simply a rigorous evaluation of the impact of different treatment options that are available for treating a given medical condition *for a particular set of patients.*"[1] The second goal of HTE analysis is individual-level prediction. Predicting beneficial

and adverse responses of individuals to different treatments in terms of multiple endpoints is essential for informing individualized treatment decisions. One version of this goal has been described as answering the question: "Who will benefit most from Treatment A and who will benefit most from Treatment B?"[4] Creating such a narrowly defined subgroup (the individual patient) leads to an extremely challenging problem, which has not been adequately studied and for which there are few reliable methods that provide protection against spurious findings.[5] Subgroup analysis, on the other hand, has been extensively studied.[6] Hence, we will focus on subgroup analysis.

Subgroup Analysis

Subgroup analysis is the most commonly used analytic approach for examining HTE. This method usually evaluates the treatment effect for a number of subgroups, one variable at a time, usually a baseline or pretreatment variable. A test for interaction is conducted to evaluate if a subgroup variable has a statistically significant interaction with the treatment indicator. If the interaction is significant, then the treatment effect is estimated separately at each level of the categorical variable used to define mutually exclusive subgroups (e.g., men and women).

It should be cautioned, however, that the interaction test generally has low power to detect differences in subgroup effects.[7] For example, when compared with the sample size required for detecting ATE of a particular size, a sample size roughly four times as large is required for detecting a difference in subgroup effects of the same magnitude as ATE for a 50:50 subgroup split; a sample size approximately 16 times as large is required for detecting a difference that is half of ATE (at significance level 0.05).

Even though the interaction test has low power to detect a true difference in subgroup effects, there is a danger of falsely detecting a difference in subgroup effects if we perform separate interaction tests for multiple subgrouping variables. That is, suppose we perform separate interaction tests for 100 subgroup variables. The interaction test will be statistically significant (at a significance level of 0.05), on average, for about five subgroup variables, when in truth the treatment effect is homogeneous. If we make a Bonferroni correction for multiple testing in order to maintain the correct Type-I error probability, we would be further increasing the Type-II error probability, which increases the likelihood of not identifying true heterogeneity in subgroup effects.

It should also be noted that a statistical test of interaction does not correspond to an assessment of biological interaction. The presence or absence of statistical interaction depends on various mathematical aspects of the regression model (e.g., scale of dependent variable, covariates present in the model, distributional assumptions). These considerations are largely irrelevant for biological interactions.[3]

A useful illustration of the potential for subgroup analyses (and implied HTE) to lead to erroneous inferences came from a large randomized trial of therapies for myocardial infarction. In 1988, the results of the Second International Study of Infarct Survival (ISIS-2) study, a randomized 2x2 factorial study of the effect of streptokinase and aspirin for treatment of myocardial infarction, were published.[8] This study provided evidence indicating that either streptokinase or aspirin reduced mortality following myocardial infarction, and that the combination of streptokinase and aspirin improved survival over either treatment alone. In the aspirin-treated subjects, there was a reduction in mortality (804 deaths among 8,587 people, 9.4%) relative to subjects not treated with aspirin (1,016 deaths among 8,600 people, 11.8%, $p<0.05$). Numerous subgroup analyses were conducted, most of which indicated relatively consistent effects of aspirin. However, one particular subgroup analysis, astrological birth sign, suggested heterogeneity of effect. In the subgroup of patients born under the astrological sign Gemini or Libra, there were more deaths (150 of 1,357, 11.1%) among the aspirin-treated patients than there were among the non–aspirin-treated patients (147 of 1,442, 10.2%) (p not significant).

This apparent heterogeneity in the effect of aspirin served as a caution about the causal interpretation of findings from unfocused, exploratory subgroup analyses. Rather than inferring that aspirin should not be used in the treatment of

myocardial infarction if the patient is a Gemini or a Libra, the authors pointed to the potential for overinterpreting results of subgroup analyses. When the ATE is clearly positive (both aspirin and streptokinase reduce mortality in patients with myocardial infarction) and many subgroup analyses are conducted, false positive or negative findings are to be expected. Findings from such unfocused, exploratory subgroup analyses should be interpreted with caution even if a plausible biologic mechanism exists, and with greater caution if the apparent heterogeneity of treatment is not supported by a plausible mechanism (as with the astrological sign subgroup).

The ISIS-2 study conducted additional subgroup analyses to assess the consistency of the subgroup findings from an earlier randomized trial of streptokinase (GISSI) that found no benefit of streptokinase among persons older than 65, those with a previous infarct, and those presenting more than 6 hours after the onset of pain. In contrast to GISSI, the ISIS-2 study found a mortality benefit for streptokinase among these subgroups, a finding that further underscores the need for caution when drawing inferences from subgroup results.When there are plausible a priori reasons that a treatment may not be effective (such as in patients with contraindications to the therapy) and subgroup analyses find no benefit in that subgroup, stronger inferences might be drawn.

Types of Subgroup Analysis

Three different types of subgroup analyses may be distinguished: (a) confirmatory, (b) descriptive, and (c) exploratory.[2] See Table 3.1 for a summary of the essential characteristics of these three types of subgroup analyses.

Confirmatory Subgroup Analysis

The main goal is to test and confirm hypotheses about subgroup effects. The essential elements of this type of analysis are: clear definition and prespecification of subgroups; clear definition and prespecification of endpoints related to outcomes; prespecification of a small number of hypotheses about subgroup effects, including the direction in which the effects are expected to vary in subgroups; availability of strong a priori biological and epidemiological evidence; detailed description of a statistical analysis plan for how testing will be done; and adequate power to test subgroup hypotheses. Essentially, the study intent, design, and analysis are all focused on the subgroup hypotheses to be tested. Due to these stringent requirements, the findings from a confirmatory analysis are potentially actionable.

Descriptive Subgroup Analysis

The main goal of descriptive subgroup analysis is to describe the subgroup effects for future evaluation and synthesis. The essential elements of this type of analysis are: clear definition and prespecification of subgroups, clear definition and prespecification of endpoints related to outcomes, prespecification of hypotheses relating to subgroup effects, and detailed description of a statistical analysis plan for how testing will be done. The results of these subgroup analyses may be presented as a table in the main report and as a forest plot, with a vertical line representing the overall treatment effect (ATE). See Antman et al. for a good example of such a forest plot.[9] Alternatively, the results may be made available as an appendix or as electronic supplemental material in order to facilitate future evaluation and for synthesis and meta-analysis by systematic reviewers. A detailed discussion of descriptive subgroup analysis is presented in Varadhan et al.[2]

Exploratory Subgroup Analysis

Exploratory subgroup analyses are done mainly to identify subgroup hypotheses for future evaluation. Typically, exploratory subgroups are not prespecified. Compared to confirmatory and descriptive HTE analyses, exploratory analyses enjoy more flexibility for identifying baseline characteristics that interact with treatment. Definition of subgroups, endpoints, hypotheses, and modeling parameters are usually derived in response to the data. An example of this would be the use of a stepwise model selection approach to identify treatment by covariate interactions. A major problem with these analyses is that it is extremely difficult to obtain the sampling properties of subgroup effect estimators (e.g., standard errors). Often, it is not clear how many hypotheses were tested (e.g., using stepwise model

Table 3.1. Essential characteristics of three types of subgroup analyses[2]

Properties	Confirmatory	Descriptive	Exploratory
Goal	To test hypotheses related to subgroup effects	To report treatment effects for future synthesis	To generate hypotheses for further study
Number of hypotheses examined	A small number, typically one or two	Moderate and prespecified	Not made explicit, but may be large, and not prespecified
Prior epidemiological or mechanistic evidence for hypothesis	Strong	Weak or none	Weak or none
Prespecification of data analytic strategy	Prespecified in complete detail	Prespecified	Not prespecified
Control of familywise type I error probability	Necessary	Possible, but not essential since the goal is not to test hypotheses	Not essential
Characterization of sampling error of the statistical estimator	Easy to achieve	Possible	Difficult to characterize sampling properties (e.g., confidence intervals)
Power of testing hypothesis	Study may be explicitly designed to have adequate power	Likely to be inadequately powered	Inadequate power to examine several hypotheses

selection to identify HTE). Post hoc exploratory subgroup analyses may sometimes identify promising hypotheses that could be subject to more rigorous future examination. The results of these subgroup analyses, while potentially important, should be clearly labeled as exploratory.

Potentially Important Subgroup Variables

Important subgroups are ones for which limited data are typically available, such as the AHRQ priority populations (e.g., women, men, children, minorities, elderly, rural populations, individuals with disabilities, etc.).[10]

Subgroup variables must be true covariates, that is, variables that are defined before an individual is exposed to the treatment or variables that are known to be unaffected by the treatment. Variables that change in response to treatment and post-randomization variables are not covariates. Some additional important types of subgroup variables are: (1) demographic variables (e.g., age); (2) pathophysiologic variables (e.g., timing after stroke, stable or unstable angina); (3) comorbidities (e.g., presence of renal disease when treating hypertension); (4) concomitant exposures (e.g., beta-blockers, aspirin); and (5) genetic markers (e.g., interaction between K-ras gene mutation and cetuximab for colorectal cancer). Sex and age should always be evaluated for interaction with treatment, although it is not obvious how to define the age categories. Notwithstanding, the definition of age categories should be prespecified. The other subgroup variables should be considered when there is prior epidemiological or mechanistic evidence suggesting some potential for interaction with the treatment.

Subgroup Analyses: Special Considerations for Observational Studies

General Considerations

Randomized trials generally have broad exclusion criteria that serve several purposes. These criteria reduce the heterogeneity of the study population so that there is less variability with respect to

outcome measures, thereby improving statistical power for a given sample size. Exclusion criteria also serve to protect patients who might be harmed by a treatment (such as those with a contraindication to the treatment). Since the aim of many observational studies is to describe the effect of treatment as actually used, fewer exclusions are typically applied, and those that are often applied are for the purpose of improved confounder control. As a result, observational studies often include patients for whom no randomized data of treatment effect exists. For example, a patient with a relative contraindication for a treatment might be excluded from a randomized trial, but a treating clinician may decide that the benefits outweigh the risks for this patient and apply the therapy.

The study of treatment effects can be challenging in observational studies. Observational studies are susceptible to confounding by indication, ascertainment biases in exposure to treatment, measurement error in assessment of health outcomes, and lack of information on important prognostic variables (in studies using existing data). These biases and measurement errors can introduce apparent HTE when in fact none is present, or conversely, obscure true HTE. Because heterogeneity in observational studies can be due to chance or bias, investigators must evaluate the observed HTE to determine whether a finding is indicative of true heterogeneity. To do this, chance findings should be evaluated by testing for interaction; biases should be avoided by adhering to sound study design principles and by evaluating balance on covariates within subgroups to assess the potential for confounding.

There are several potential sources of heterogeneity in observational studies, and these tend to mirror the potential explanations for a finding of an overall effect (ATE). As such, many of the approaches for reducing the potential for an incorrect inference are the same. Careful attention to study design principles is an important starting point for avoiding incorrect inferences with respect to overall findings and also benefits the identification of potential HTE. The use of the incident (new) user design reduces the potential for inclusion of immortal person-time (i.e., person-time during which a study outcome cannot occur; see chapter 4 for a detailed discussion).[11] Contemporaneous followup of exposed and unexposed subjects (parallel group design) avoids calendar time differences in exposure/covariate/outcome identification. Measures of exposure, outcome, and covariates should address misclassification and seek to limit potential for information bias.

Despite the challenges in using observational data for HTE analysis, randomized experiments cannot be performed to answer all clinically important questions regarding HTE attributable to patient characteristics. Therefore, a huge demand will be placed on observational studies to produce evidence to inform decisions. Hence, procedures must be put in place to ensure that the results from observational studies are trustworthy. A key principle here is that the observational studies should be designed and analyzed in the same manner as randomized controlled experiments. Some potential steps include registering observational CER studies prospectively, publishing the study protocol (including clear definitions of subgroups and outcomes, prespecified hypotheses, and power calculations), and developing a detailed analytic plan (including how confounding, missing data, and loss to followup will be handled). Sox has called for registration of observational studies, along the lines of the National Institutes of Health's clinical trials registry.[12] Rubin has put forth an interesting proposal for "objective causal inference," in which greater emphasis is placed on understanding treatment selection. The modeler is blinded to outcomes until the treatment assignment modeling is completed and made available to scrutiny.[13] This places the emphasis on study design and treatment assignment, and the investigator only observes outcomes at the end, as in randomized experiments. This ensures some degree of objectivity in the outcome modeling. These proposals are worth serious consideration.

Prediction of Individual Treatment Effects

This chapter has focused on analytic approaches to subgroups within a population, but variations of effect can also occur within individuals. The individual causal effects (Box 3.1), $\theta_i = g(E[Y_i(1)]) - g(E[Y_i(0)])$, are not identifiable from the data without untestable assumptions. For acute or transient outcomes, methods such as crossover

designs or N-of-1 trials may be appropriate for estimating individual effects. For nonacute outcomes, prediction models may be developed for predicting the response of individuals to different treatments. Prediction of individual responses can also be viewed as an extreme version of subgroup analysis, where individuals are cross-classified by a large number of covariates. It is quite likely that most covariate profiles viewed as cells in a high-dimensional contingency table would be either empty or sparsely populated. Consequently, individual-level predictions can be highly variable and sensitive to modeling assumptions. An example of a prediction model is by Dorresteijn et al., who predicted the effect of rosuvastatin on cardiovascular events for individual patients using data from an RCT.[14] They evaluated the net benefit of treatment decisions for individuals based on predicted risk difference (absolute risk reduction) due to the treatment. They used existing risk models (Framingham and Reynolds risk scores), as well as a prediction model developed using the trial data to calculate baseline risk of cardiovascular outcomes for all individuals without treatment. The average treatment effect (ATE) (relative risk) was applied to calculate individual treatment effects (ITE) (ITE = baseline risk * (1-ATE)). It is important to note that prediction models must be appropriately validated in order for them to be acceptable.

Value of Stratification on the Propensity Score

A study by Kurth and colleagues illustrates the use of summary score stratification as a means to assess HTE in observational studies.[15] Since many strokes are the result of thrombosis in cerebral or precerebral arteries, a highly specific thrombolytic therapy became available in the form of recombinant tissue plasminogen activator (TPA). Three randomized studies showed that TPA neither decreased nor increased mortality substantially in people who had recently experienced a stroke. However, observational studies of the same question consistently indicated that TPA therapy increased mortality, and the reasons for the discrepancy in results between observational and interventional studies were not readily apparent. With data sourced from a German stroke registry, Kurth and colleagues were able to reproduce the observational effect of an increase in mortality with TPA with careful attention to study design and regardless of adjustment for measured covariates. However, different analytic approaches (particularly matching on the propensity score) provided results more comparable to the randomized trials than was obtainable from adjusted analyses. By stratifying patients according to propensity to receive TPA and conducting analyses of TPA effect within strata, this study found that much of the observational result was being driven by a few subjects with low propensity to receive TPA who were highly influential in analyses that included them (the covariate–adjusted, propensity score–adjusted, propensity score–stratified, and the inverse probability–weighted analyses). However, the propensity score–matched analyses excluded these influential subjects, and the standardized mortality ratio results downweighted their influence so that these results were similar to the RCTs. As a summary of propensity to receive a medication or strength of indication, propensity score identifies clinically relevant subgroups. If heterogeneity is observed in the propensity score, further investigation is warranted. Stratification of results by summary variables such as propensity scores or disease risk scores, or other clinically relevant profiles may inform the analysis.

Conclusion

RCTs often exclude individuals with characteristics that may cause variation in response to treatment, limiting the generalizability of findings from these studies. Observational studies often have broad inclusion/exclusion criteria, allowing for the assessment of comparative effectiveness in large, diverse populations in "real-world" settings. With the increase in generalizability comes the potential for HTE. Investigators should understand the potential for HTE prior to conducting an observational CER study, and clearly state if and how subgroups will be defined and analyzed. If subgroup analysis is intended to be confirmatory, investigators should ensure adequate statistical power to detect proposed subgroup effects, and adjust for multiple testing as appropriate. When an interaction test is significant, subgroup effects should be reported, and a discussion of the potential clinical importance of the findings

should be included. When an interaction test is not significant, the investigator should report the ATE and discuss plausible reasons for null findings in relation to other studies. Exploratory analyses should be clearly labeled as such, and the corresponding results should not be emphasized in the abstract of the study report. Reporting of results from descriptive analysis of subgroups defined by priority populations using an informative forest plot is encouraged.

Checklist: Guidance and key considerations for the development of the HTE/ subgroup analysis section of an observational CER protocol

Guidance	Key Considerations	Check
Summarize prior knowledge of treatment effect modifiers and reference sources		☐
Prespecify subgroups to be evaluated.	- Note if priority populations with limited effectiveness data will be included in the study and evaluated as subgroups. - Subgroups should be defined by variables measured at baseline or variables known to be unaffected by exposure	☐
Specify the hypothesized direction of effect within subgroups and the significance levels that will be used to assess statistical significance.	- If confirmatory analyses, do power calculations. - Describe methods to adjust for multiple testing, if applicable.	☐
Describe how confounding will be addressed.	- Assess covariate balance between the treatment groups within each stratum of the subgrouping variable.	☐
Describe statistical approaches that will be used to test for interactions for prespecified covariates.	If the interaction test is not significant: - Report ATE. - Discuss plausible reasons for null findings in relation to other studies and plausible biological mechanism.	☐
Describe how overall (ATE) and subgroup effects will be reported if interaction test is or is not significant.	- Clearly distinguish subgroup results as confirmatory, descriptive, or exploratory analyses. - Report subgroup effects in a table and/or a forest plot with a vertical line representing the overall treatment effect (ATE).	☐

References

1. Research on the Comparative Effectiveness of Medical Treatments, A CBO Paper. U.S. Congress, Congressional Budget Office, 2007.
2. Varadhan R, Segal JB, Boyd CM, et al. Heterogeneity of treatment effect in patient-centered outcomes research. Accepted for publication in the Journal of Clinical Epidemiology.
3. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol. 1980;112:467–70.
4. Sox HC. Defining comparative effectiveness research: the importance of getting it right. Med Care. 2010 Jun 48;(6 Suppl):S7-8.
5. Cai T, Tian L, Wong PH, et al.Analysis of randomized comparative clinical trial data for personalized treatment selection. Biostatistics. 2011 Apr 12;270-82.
6. Wang R et al. Statistics in medicine – reporting of subgroup analyses in clinical trials. NEJM. 2007; 357: 2189-94.

7. Brookes ST, Whitley E, Peters TJ, et al. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technology Assessment. 2001;5:1-56.

8. ISIS-2 (Second International Study of Infarct Survival) collaborative group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction. Lancet. 1988; ii:349-60.

9. Antman EM et al. Enoxaparin versus unfractionated heparin with fibrinolysis for ST-elevation myocardial infarction. NEJM. 2006;354:1477-88.

10. Agency for Healthcare Research and Quality. Health Care: Priority Populations Index Page. Retrieved from http://www.ahrq.gov/populations/. Accessed September 21, 2012.

11. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167:492-9.

12. Sox HC, Helfand M, Grimshaw J, Dickersin K; PLoS Medicine Editors, Tovey D, Knottnerus JA, Tugwell P. Comparative effectiveness research: challenges for medical journals. Am J Manag Care. 2010. May;1;16(5):e131-3

13. Rubin DB. For objective causal inference, design trumps analysis. Ann Appl Stat. 2008; 2(3):808–40.

14. Dorresteijn JAN, Visseren FLJ, Ridker PM, et al. Estimating treatment effects for individual patients based on the results of randomized clinical trials. Br Med J. 2011;343:d5888.

15. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol. 2006;163:262–70.