Published in final edited form as:

J Clin Epidemiol. 2013 August ; 66(8): 818–825. doi:10.1016/j.jclinepi.2013.02.009.

A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research

Ravi Varadhan^{a,b,e,*}, Jodi B. Segal^{a,c}, Cynthia M. Boyd^{a,c}, Albert W. Wu^{a,c}, and Carlos O. Weiss^{1,d}

^aDepartment of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

^bDepartment of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD 21205, USA

^cDepartment of Health Policy and Management, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD 21224, USA

^dDepartment of Family Medicine, College of Human Medicine, Michigan State University, East Lansing, MI 49503, USA

^eThe Center on Aging and Health, Johns Hopkins University, Baltimore, MD 21205, USA

Abstract

Individuals vary in their response to a treatment. Understanding this heterogeneity of treatment effect is critical for evaluating how well a treatment can be expected to work for an individual or a subgroup of individuals. An overemphasis on hypothesis testing has resulted in a dichotomy of all heterogeneity of treatment effect analyses into confirmatory (hypothesis testing) and exploratory (hypothesis finding) analyses. This limited view of heterogeneity of treatment effect is inadequate for creating evidence that is useful for informing patient-centered decisions.

An expanded framework for heterogeneity of treatment effect assessment is proposed. It recognizes four distinct goals of heterogeneity of treatment effect analyses: hypothesis testing, hypothesis finding, reporting subgroup effects for meta-analysis, and individual-level prediction. Accordingly, two new types of heterogeneity of treatment effect analyses are proposed: descriptive and predictive. Descriptive heterogeneity of treatment effect analyses report treatment effects for prespecified subgroups in accordance with prospectively specified analytic strategy. They need not be powered to detect heterogeneity of treatment effect. They emphasize estimation and reporting of subgroup effects rather than hypothesis testing. Sampling properties (e.g., standard error) of descriptive analysis can be characterized, thus facilitating meta-analysis of subgroup effects. Predictive heterogeneity of treatment effect analyses estimate probabilities of beneficial and adverse responses of individuals to treatments and facilitates optimal treatment

© 2013 Elsevier Inc. All rights reserved.

*Corresponding author. Division of Geriatric Medicine and Gerontology, The Johns Hopkins Center on Health and Aging, 2024 East Monument Street, Suite 2-700, Baltimore, MD 21205-2223, USA. Tel.: +1-410-502-2619; fax: +1-410-614-9625. rvaradhan@jhmi.edu (R. Varadhan).

Conflict of interest: None of the authors have any relevant financial interests or other conflicts to disclose.

decisions for different types of individuals. Procedures are also suggested to improve reliability of heterogeneity of treatment effect assessment from observational studies.

Heterogeneity of treatment effect analysis should be identified as confirmatory, descriptive, exploratory, or predictive analysis. Evidence should be interpreted in a manner consistent with the analytic goal.

Keywords

Subgroup analysis; Comparative effectiveness research; Prediction; Meta-analysis; Clinical trials; Observational studies

1. Introduction

Individuals vary in their response to a treatment: some derive substantial benefit, some derive little benefit, and others are harmed. Heterogeneity of treatment effect can be defined as nonrandom explainable variability in the direction and magnitude of individual treatment effects, including both beneficial and adverse effects. Understanding this heterogeneity of treatment effect 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 payors. Heterogeneity of treatment effect is central to patient-centered outcomes research (PCOR) [1], which aims to help stakeholders make informed personalized health care decisions.

Heterogeneity of treatment effect analyses are commonly divided into confirmatory (hypothesis testing) and exploratory (hypothesis finding) analyses. This dichotomy is the result of an overemphasis on hypothesis testing. It offers a limited view of heterogeneity of treatment effect that is inadequate for creating useful evidence for informing patient-centered decisions. Therefore, we propose that the main goals of heterogeneity of treatment effect analyses are to (1) test hypothesis regarding differences in subgroup effects, (2) generate hypothesis related to subgroup effects, (3) estimate and report subgroup effects for future synthesis, and (4) predict whether an individual might benefit from or be harmed by a treatment. Subgroup analysis (SGA) and the analysis of interactions are the most common analytic approaches for addressing goals (1)–(3). Predicting beneficial and adverse responses of individuals to different treatments in terms of multiple end points is essential for informing individualized treatment decisions. This is the fourth goal of heterogeneity of treatment effect analysis.

Analysis of heterogeneity of treatment effect, that is, estimation of treatment effects in subgroups and for individuals, is challenging. It is difficult to obtain valid inferences regarding these effects. Although much has been written about the perils of SGA, examples abound of inappropriate conduct, reporting, interpretation, and application of subgroup analytic results [2,3]. Lack of clarity regarding the goals of SGA is likely the main reason for this state of affairs.

This conflict can be resolved by recognizing the distinct purposes served by heterogeneity of treatment effect analyses, and how the different elements of heterogeneity of treatment effect analyses relate to those purposes (Table 1). An expanded framework of heterogeneity of treatment effect can identify the distinct inferential goals of heterogeneity of treatment effect analyses and the corresponding approaches that are in accordance with the inferential goals.

2. Existing approach to SGAs

SGA is the most prevalent form of heterogeneity of treatment effect analysis. A number of authors have proposed checklists or guidelines for the conduct and interpretation of SGA [3–6]. Two types of SGA are commonly recognized: confirmatory and exploratory.

2.1. Confirmatory heterogeneity of treatment effect analysis

The goal of confirmatory heterogeneity of treatment effect analysis is to rigorously test hypotheses. Required elements include prespecification of a small number of well-defined subgroups, strong biological rationale and prior evidence to support subgroup hypotheses, adequate power to test subgroup hypotheses, prespecification of the analytic plan, control of family wise type I error, and the presence of a significant overall treatment effect [3–6]. Additionally, the outcome measure, exposure (i.e., intensity, timing, and duration of treatment), and important prognostic variables must be of high quality.

A distinguishing feature of confirmatory analysis is that the sampling properties (e.g., standard errors of effect estimates) can be assessed accurately. This is difficult in exploratory analysis, which tends to be data driven. Based on recent systematic reviews, including our own, confirmatory heterogeneity of treatment effect is rare [5].

2.2. Exploratory heterogeneity of treatment effect analysis

Heterogeneity of treatment effect analyses that do not meet the requirements of confirmatory heterogeneity of treatment effect are usually deemed exploratory. The goal of exploratory heterogeneity of treatment effect analyses is to generate hypotheses for further study. Compared with confirmatory heterogeneity of treatment effect, exploratory analyses enjoy more flexibility for identifying baseline characteristics that are most strongly related to heterogeneity of treatment effect. In reports of exploratory analyses, the number of hypotheses examined is often not stated. In some situations, it is even difficult to know how many hypotheses were tested (e.g., as in covariate selection). It is difficult to characterize the sampling properties for exploratory heterogeneity of treatment effect. Limitations in study design and poor measurement quality vis-à-vis the specific scientific question may also render heterogeneity of treatment effect analysis exploratory. Therefore, the results from exploratory analyses, irrespective of biological plausibility or statistical significance, should be regarded as preliminary until replicated. Responsible conduct and reporting of exploratory SGAs include attention to inflated type I error and being transparent about the number of hypotheses examined, de-emphasizing *P*-values for within-subgroup tests of treatment effects, clearly labeling exploratory results as such, and not overinterpreting their significance.

3. Lack of clarity on the goals of heterogeneity of treatment effect analyses

In spite of the abundance of guidance on SGAs, the literature is rife with examples of inappropriate analysis, reporting, interpretation, and use of SGAs: lack of clear identification of exploratory analyses; not being transparent about the number of SGAs performed; absence of interaction tests; emphasis on *P*-values of subgroup effects; disregard of multiple comparisons; and overinterpretation of exploratory subgroup results, to name a few. Two examples are given to highlight that lack of clarity regarding the goals of SGA can lead to inappropriate reporting and interpretation of SGA results.

The Valsartan Heart Failure Trial (ValHeFT) trial tested the effectiveness of valsartan, an angiotensin receptor blocker, for preventing mortality and morbidity (cardiac arrest and hospitalization for heart failure) [7]. A post hoc analysis showed significant interaction of treatment with baseline use of angiotensin-converting enzyme (ACE) inhibitors or beta blockers (a four-category subgroup). There was a significantly increased mortality rate in the valsartan arm in those using both medications at baseline but a decreased mortality rate in the valsartan arm in the other three groups. In particular, among 366 (7%) patients who were not receiving an ACE inhibitor, whether a beta blocker had been prescribed, there was a significantly lower risk of the combined end point in the valsartan group than that in the placebo group. Interestingly, this post hoc finding resulted in the Food and Drug Administration (FDA) approving valsartan for heart failure patients who cannot tolerate ACE inhibitors [8].

The Bypass Angioplasty Revascularization Investigation trial provides another good example of the lack of clarity regarding the goal of an SGA [9]. This study randomized patients with multivessel coronary artery disease to one of two procedures: percutaneous transluminal coronary angioplasty (PTCA) (*N* = 915) or coronary artery bypass grafting (CABG) (*N* = 914). The primary end point was all-cause mortality. The study protocol prespecified five subgrouping variables: severity of angina (three levels), number of diseased vessels (two or three), left ventricular function (normal/abnormal), the complexity of lesions (class C lesion present/absent), and a combination of number of diseased vessels and left ventricular function (four possibilities). After the study got underway, the Data Safety Monitoring Board requested that the treatment effect in diabetics and nondiabetics also be monitored as additional subgroups because of concerns aroused by a previous study of PTCA in diabetics. Significance levels of 0.01 and 0.005 were used in the prespecified and diabetes-based subgroups, respectively, to account for multiple testing. There was no significant difference between the two procedures on primary end point over an average follow-up period of 5.4 years. The only significant difference occurred in the subgroup of patients with treated diabetes: 5-year survival of 65.5% in the PTCA group vs. 80.6% in the CABG group (log-rank test, *P* = 0.003). This SGA clearly does not satisfy the requirements for confirmatory analysis, rendering the reliability of this subgroup finding as questionable. It was not preplanned by study investigators; there was relatively little prior evidence suggesting the plausibility of difference according to diabetic status; diabetes exposure was not measured well; it was one of six subgrouping variables suggesting a lack of focused hypothesis testing; and the overall treatment effect was null. Yet, this result was emphasized in the abstract of the primary trial report. The National Institutes of Health (NIH) issued a

clinical alert advising that the treated diabetic would probably fare better with CABG than PTCA as an initial treatment (http://www.nlm.nih.gov/databases/alerts/bypass_diabetes.html).

Because the SGAs in these examples were exploratory, it would arguably have been appropriate to require independently confirmation of the subgroup findings. However, both the FDA and the NIH acted definitively on these exploratory subgroup results. Indeed, policy makers may be compelled to act by the seriousness of disease without alternative therapies although evidence is incomplete or imperfect. These cases justify the concern that reporting the SGA results is akin to opening Pandora's box, in that it can have unintended consequences. Clarity regarding the purposes of SGAs can help dampen this overzealousness and encourage appropriate interpretation and use of subgroup results.

4. An expanded framework for heterogeneity of treatment effect analyses

Existing guidelines view SGA as a hypothesis-testing problem rather than as an estimation problem. This view has unfortunately resulted in the dichotomization of SGA into confirmatory (hypothesis testing) and exploratory (hypothesis generating) analyses. These two categories are inadequate to capture the richness of heterogeneity of treatment effect assessment in PCOR. An expanded analytic framework that emphasizes estimation of subgroup effects is proposed. This framework contains two additional types of heterogeneity of treatment effect analyses: descriptive and predictive. The four types of heterogeneity of treatment effect analyses differ in important ways that are listed in Table 1.

4.1. Descriptive heterogeneity of treatment effect analyses

The goal of descriptive heterogeneity of treatment effect analysis is to facilitate future synthesis (e.g., meta-analysis) by reporting treatment effect estimates (and their standard errors) in subgroups. There are two types of descriptive heterogeneity of treatment effect analysis: the estimation of primary treatment effects in generally important subgroups (descriptive SGA) and the estimation of treatment effects on multiple efficacy and safety end points (descriptive outcomes analysis). The subgroups and efficacy end points must be prespecified.

Which types of subgrouping variables should be considered for descriptive SGA? Some potentially important classes of variables are demographics (e.g., age and sex), behavior (e.g., smoking), pathophysiology (e.g., measures of disease severity), genetic markers/mutations, and comorbid conditions (e.g., diabetes status in cardiovascular disease trials). Rothwell [4] provides a comprehensive set of characteristics that may be useful determinants of heterogeneity of treatment effect. Descriptive heterogeneity of treatment effect might also be considered for subpopulations for which limited evidence is available in the literature such as the priority populations specified by the Agency for Healthcare Research Quality including women, children, minorities, elderly, individuals with disabilities, and rural populations [10].

A few journals actually recommend reporting results for important subgroups. For example, the *Journal of the National Cancer Institute* [11] instructs authors that “where appropriate,

clinical and epidemiologic studies should be analyzed to see if there is an effect of sex or any of the major ethnic groups. If there is no effect, it should be so stated in Results.” Recently, the FDA issued draft guidance [12] for industry trial sponsors of medical device clinical studies that recommended sex-specific heterogeneity of treatment effect analyses. Notably, the draft guidance says that even in the absence of a significant sex-by-treatment interaction, it is preferable to report results by sex. A workshop sponsored by the Institute of Medicine (IOM) [13] also discussed reporting of sex-specific results from studies. Some of the journal editors at the workshop cautioned that the reporting of sex-specific results can be misleading and subject to misinterpretation because of the potential for type II errors, that is publishing comparisons by sex that do not show a significant difference. They argued that if a study was not adequately powered to look for such differences, then showing no differences was meaningless. However, it was also suggested that when there is insufficient power to analyze sex differences within a study, it may be possible to combine data from various studies and conduct meta-analysis or apply advanced statistical methods, such as the use of Bayesian inference. This is a critically important insight that captures the essence of descriptive analysis.

For example, a sex-specific meta-analysis of 18 randomized controlled trials (RCTs) of statins for primary or secondary prevention of cardiovascular events showed that statins decreased the risk of cardiovascular events and all-cause mortality in both men and women [14]. The proportion of women in these trials ranged from 14% to 69%. This analysis would not have been feasible had the original trials not reported separate treatment effects for men and women. The published meta-analysis excluded 32 eligible trials because they did not report sex-specific results, leaving one to wonder if the inferences might have been different had the sex-specific results been reported in those trials.

Age, being a continuous variable, creates problems for descriptive SGA. How should age-based subgroups be defined? How should the potential inconsistency in subgroup definitions across studies be addressed? Zulman et al. [15] found that 42 of 109 trials conducted age-specific SGAs. Of these, 3 studies analyzed age as a continuous variable, 22 studies as dichotomous, 14 studies as categorical, and 6 studies did not state the definition. Of these, only three studies clearly prespecified the definition of age groups for SGA. This inconsistency would make it challenging to undertake an age-based meta-analysis of treatment effects. More generally, this issue highlights the importance of establishing consistency in how continuous variables should be defined for descriptive SGA.

Pathophysiological conditions that could be involved in causal pathways of treatment efficacy should be included. These may be measures of disease severity, indicators of comorbid conditions, or stage in the natural history of disease. The individual’s genotype is, sometimes, a determinant of intended response to treatment and susceptibility to adverse reactions for several drugs. Hence, genetic mutations, if known to be causal, should be included. An example of this is the SGA reported by Karapetis et al. [16] showing that the drug cetuximab is less effective than supportive care alone for the prevention of death from advanced colorectal cancer in individuals with a specific genetic mutation (mutation of the K-ras gene).

The analytic plan for descriptive SGA should be prespecified. The study need not be powered to test subgroup treatment effects. Consequently, inference on these multiple effects is likely to be weak. The results of descriptive heterogeneity of treatment effect analyses may be presented as a table in the main report or as a forest plot, with a vertical line representing the overall treatment effect. Alternatively, results may be made available as an appendix or as electronic supplemental material. It should be understood that the purpose of descriptive heterogeneity of treatment effect analysis is to capture the information in the study and summarize it in a form suitable for use by meta-analysts. Currently, meta-analysts interested in estimating subgroup effects are hamstrung because of nonreporting of subgroup effects in individual studies. As a result, they resort to meta-regression approaches, which are susceptible to aggregation (ecological) bias. Our proposal of reporting descriptive subgroup effects (with valid uncertainty estimates) would facilitate meta-analysis that is free from aggregation bias. We also gain evidential strength by combining multiple trials using the adequate information reported in the descriptive heterogeneity of treatment effect analyses.

Table 1 shows how descriptive heterogeneity of treatment effect analysis differs from the conventional types of confirmatory and exploratory heterogeneity of treatment effect analyses. Although descriptive heterogeneity of treatment effect is prespecified, unlike confirmatory heterogeneity of treatment effect, it does not rely on strong prior evidence. Descriptive analyses need not be powered to detect heterogeneity within a study because the goal is to report results for future synthesis. Consequently, results from a descriptive heterogeneity of treatment effect analysis from an individual study should not be overinterpreted (e.g., reporting the P-values should be avoided). Descriptive heterogeneity of treatment effect analysis differs from exploratory heterogeneity of treatment effect analyses mainly in that it is prespecified and its sampling properties (e.g., standard error) can be characterized making future synthesis feasible. In contrast, the results from exploratory heterogeneity of treatment effect are not appropriate for meta-analysis.

4.2. Predictive heterogeneity of treatment effect analyses

The main goal of predictive heterogeneity of treatment effect analyses is to develop algorithms for predicting which of the two (or more) treatments is better for a given individual [17]. For two treatments, A and B, the prediction algorithm will estimate the probabilities of experiencing a specific outcome (within a time interval) for a specific individual under treatments A and B, along with estimates of uncertainty regarding the probabilities. The two probabilities can be used to make an informed decision about which treatment should be chosen for the individual.

Prediction of individual responses can also be viewed as an extreme version of SGA, in which 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. Data from RCTs are ideally suited for predictive heterogeneity of treatment effect analysis because of the availability of reliable information on exposure, outcomes, and important prognostic variables (cf. Section VI).

Dorresteijn et al. [18] predicted the effect of rosuvastatin on cardiovascular events for individual patients using data from an RCT. They evaluated the net benefit of treatment decisions for individuals based on predicted risk difference (absolute risk reduction) because of the treatment. They used the existing risk models (Framingham and Reynolds risk scores) and a prediction model developed using the trial data to calculate baseline risk of cardiovascular outcomes for all individuals without treatment. The overall treatment effect (relative risk or hazard ratio) was applied to calculate individual treatment effects ($ITE = \text{baseline risk} \times (1 - OTE)$).

Cai et al. [19] proposed a systematic two-stage estimation procedure for individual-level treatment differences for future patient's disease management and treatment selection. Their method is applicable to parallel group randomized trials for comparing two treatments. In the first stage, parametric or semiparametric models are fitted separately to each treatment group to estimate individual-level differences. An index score system is calculated from these differences for grouping individuals. The average treatment difference is consistently estimated within each stratum of the index using a nonparametric function estimation method. Simultaneous confidence intervals are constructed to make valid inferences about stratum-specific treatment differences. There is an optimal treatment for individuals belonging to strata in which the average treatment difference is significantly different from zero.

A multivariable risk score approach was proposed by Kent et al. [20], with the goal of exploring treatment effect heterogeneity in terms of baseline risk (i.e., risk of outcome in the absence of treatment). The authors recommended reporting both absolute risk reduction and relative risk reduction as a function of the baseline risk. They recommended the use of risk scores obtained from preexisting validated risk models (e.g., Framingham risk model), which limits this approach to situations in which such validated risk models are available.

Numerous predictive learning (or machine learning) methods are available including linear/logistic/Cox regression, neural nets, support vector machines, boosted decision trees, and random forests [21]. It is easy to generate prediction rules that fit the data too well, but predict poorly in independent data (over fitting or generalization error). Thus, the central challenge in predictive learning is to minimize the generalization error. The main idea is to promote parsimony and penalize model complexity when it adds little to model fitness. Various strategies exist for doing this, including the use of information criteria, cross-validation, and split-sample validation [22]. Sampling properties of estimates are difficult to characterize (e.g., bias and variance), and consequently, predictions are not reliable unless they are subjected to independent validation. Breiman [23], and the discussants of his article, provide a thought-provoking discussion of predictive modeling.

5. Heterogeneity of treatment effect in observational studies

Randomized controlled trials are central to evidence-based medicine, in which observational studies play a secondary role. Observational studies, however, will likely have a more central role in PCOR. There are different types of observational study designs, but the essential difference between experimental and observational studies is that in the

experimental study, the assignment of treatment to the units in the study is under the control of the investigator. Another critical difference, which is underappreciated, is that in experimental studies, the design and data collection are targeted toward estimating a treatment's effects. In observational studies, on the other hand, the study design and procedures are not always targeted toward estimating the treatment effect. Consequently, the type and quality of data that are collected are not optimal for estimating treatment effect. For example, data on initiation, duration, and intensity of treatment are often unavailable.

Observational studies are susceptible to ascertainment and selection biases in exposure to treatment, measurement error in assessment of health outcomes, and lack of information on important prognostic variables. These biases and measurement errors can introduce apparent heterogeneity of treatment effect when in fact none is present, or conversely, obscure heterogeneity of treatment effect when it is actually present [24]. David Byar [25], three decades ago, cogently argued that sound inferences on comparing treatments would not generally be possible from analyzing observational data because of difficulties with bias in treatment assignment, nonstandard definitions, definitions changing in time, specification of groups to be compared, missing data, and multiple comparisons. Although there have been significant advances in the statistical methodology for making valid inference from observational data (e.g., propensity scores, instrumental variables, and principal stratification), it is difficult to overcome fundamental design and data limitations without relying on strong and, often, unjustifiable assumptions.

Despite the challenges in using observational data, it is inescapable that randomized experiments cannot answer every question to be posed by PCOR. Observational designs must be used to produce evidence to inform patient-centered decisions. Therefore, procedures must be put in place to ensure that the results from observational studies are trustworthy. Such procedures may include registering observational studies prospectively, prespecifying hypotheses, performing power calculations, prespecifying the analytic plan including how confounding and missing data will be handled. Sox et al. [26] called for registration of observational studies, analogous to NIH's clinical trials registry. Rubin [27] put forth an interesting proposal for "objective causal inference," in which emphasis is placed on understanding treatment selection, and the modeler is blinded to outcomes until the treatment assignment modeling is complete.

6. Implications of a goal-oriented framework for heterogeneity of treatment effect analyses

Existing frameworks for heterogeneity of treatment effect analysis lack clarity on the different goals of analysis. They only recognize confirmatory (hypothesis testing) and exploratory (hypothesis generating) analyses. This dichotomy is the result of an overemphasis on hypothesis testing. It offers a limited view of heterogeneity of treatment effect that is inadequate for creating useful evidence for informing patient-centered decisions. Therefore, we propose that the two main goals of heterogeneity of treatment effect analyses are to (1) estimate treatment effects in clinically relevant subgroups and (2) predict whether an individual might benefit from a treatment. Accordingly, an expanded analytic framework is proposed comprising four categories of heterogeneity of treatment

effect analyses: confirmatory, descriptive, exploratory, and predictive, each with different inferential goals (Table 1). This framework requires that the inferential goals of the heterogeneity of treatment effect analyses be made explicit. Each heterogeneity of treatment effect analysis must clearly be identified as confirmatory, descriptive, or exploratory (predictive analysis is sufficiently distinct to stand alone without causing any confusion). Consumers of heterogeneity of treatment effect analysis should interpret findings in a manner consistent with the inferential goal.

The two new elements of the proposed framework, descriptive and predictive heterogeneity of treatment effect analyses, emphasize estimation of treatment effects rather than hypothesis testing. Descriptive heterogeneity of treatment effect can help generate high-quality evidence for PCOR by facilitating meta-analysis of subgroup results. An important example of this is the sex-specific reporting of treatment effect recommended by the FDA draft guidance and the IOM report (see Section 4.1). Sex-specific analysis should be descriptive, primarily intended for future synthesis. The results from the individual study should not be overinterpreted. A finding that the treatment effect is significant in men, not in women, should be treated cautiously unless the analysis was prespecified as confirmatory. Synthesis (e.g., meta-analysis) of sex-specific effects from multiple studies can provide more reliable inference. This is the main motivation for descriptive heterogeneity of treatment effect analysis. Exploratory heterogeneity of treatment effect results should not be used for meta-analysis because the statistical properties of the estimates are unreliable.

Acknowledgments

The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the US Department of Health and Human Services. Dr. R.V. is a Brookdale Fellow of Leadership in Aging. The funding organizations played no role in any of the following: design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, and approval of the article.

Funding: This project was funded under contract no. HHS A29020050034-I-TO5 from the Agency for Healthcare Research and Quality, US Department of Health and Human Services as part of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program. Dr. C.M.B. was supported by the Paul Beeson Career Development Award Program (NIA K23AG032910; AFAR, The John A. Hartford Foundation, The Atlantic Philanthropies, The Starr Foundation, and an anonymous donor).

References

1. Washington AE, Lipstein SH. The Patient-Centered Outcomes Research Instituted—promoting better information, decisions, and health. *N Engl J Med*. 2011; 365:e31. [PubMed: 21992473]
2. Fernandez, y; Garcia, E.; Nguyen, H.; Duan, N.; Gabler, NB.; Kravitz, RL. Assessing heterogeneity of treatment effects: are authors misinterpreting their results? *Health Serv Res*. 2010; 45:283–301.
3. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991; 266:93–98. [PubMed: 2046134]
4. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005; 365:176–186. [PubMed: 15639301]
5. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007; 357:2189–2194. [PubMed: 18032770]
6. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*. 2010; 34:c117. [PubMed: 20354011]
7. Cohn JN, Tongnoni G. A randomized trial of the angiotensin receptor blocker, valsartan, for congestive heart failure. *N Engl J Med*. 2001; 345:1667–1675. [PubMed: 11759645]

8. Temple R. How FDA currently makes decisions on clinical studies. *Clin Trials*. 2005; 2:276–281. [PubMed: 16281425]
9. BARI Investigators. Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. *N Engl J Med*. 1996; 335:217–225. [PubMed: 8657237]
10. [Accessed April 1, 2013] Agency for Healthcare Research and Quality. Health care: priority populations index page. Available at <http://www.ahrq.gov/populations/>.
11. Arnold K. Journal to encourage analysis by sex/ethnicity. *J Natl Cancer Inst*. 2000; 92:1561. [PubMed: 11018090]
12. The Food and Drug Administration. Evaluation of sex differences in medical device clinical studies: draft guidance for industry and food and drug administration staff. 2011 Dec.
13. Weizemann, TR. Sex-specific reporting of scientific research: a workshop summary. Washington, DC: Institute of Medicine, National Academies Press; 2012.
14. Kostis WJ, Cheng JQ, Dobrzynski JM, Cabrera J, Kostis JB. Metaanalysis of statin effects in women versus men. *J Am Coll Cardiol*. 2012; 59:572–582. [PubMed: 22300691]
15. Zulman DM, Sussman JB, Chen X, Cigolle CT, Blaum CS, Hayward RA. Examining the evidence: a systematic review of the inclusion and analysis of older adults in randomized controlled trials. *J Gen Intern Med*. 2011; 26:783–790. [PubMed: 21286840]
16. Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med*. 2008; 359:1757–1765. [PubMed: 18946061]
17. Byar DP. Assessing apparent treatment-covariate interactions in randomized clinical trials. *Stat Med*. 1985; 4:255–263. [PubMed: 4059716]
18. Dorresteijn JAN, Visseren FLJ, Ridker PM, Wassink AMJ, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomized clinical trials. *Br Med J*. 2011; 343:d5888. [PubMed: 21968126]
19. Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*. 2011; 12:270–282. [PubMed: 20876663]
20. Kent DM, Rothwell PM, Ioannidis JPA, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*. 2010; 11:85. [PubMed: 20704705]
21. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. Springer series in statistics. New York: Springer; 2001.
22. Steyerberg, EW. Clinical prediction models. New York/Heidelberg: Springer; 2009.
23. Breiman L. Statistical modeling: the two cultures. *Stat Sci*. 2001; 16:199–231.
24. VanderWeele TJ, Mukherjee B, Chen J. Sensitivity analysis for interactions under unmeasured confounding. *Stat Med*. 2011 <http://dx.doi.org/10.1002/sim.4354>.
25. Byar DP. Why data bases should not replace randomized clinical trials. *Biometrics*. 1980; 36:337–342. [PubMed: 7407321]
26. Sox HC, Helfand M, Grimshaw J, Dickersin K, Tovey D, Knottnerus JA, Tugwell P. PLoS medicine. Comparative effectiveness research: challenges for medical journals. *Am J Manag Care*. 2010; 16:e131–e133.
27. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat*. 2008; 2:808–840.

What is new?

- An overemphasis on hypothesis testing has resulted in a dichotomy of all heterogeneity of treatment effect analyses into confirmatory and exploratory analyses. This limited view of heterogeneity of treatment effect is inadequate for creating evidence that is useful for informing patient-centered decisions.
- An expanded framework for heterogeneity of treatment effect assessment appropriate for patient-centered outcomes research is proposed. It recognizes four distinct goals of heterogeneity of treatment effect analyses: hypothesis testing, hypothesis finding, reporting subgroup effects for meta-analysis, and individual-level prediction (see Table 1). Confirmatory analyses are hypothesis testing; exploratory analyses are hypothesis finding; descriptive analyses report subgroup-specific treatment effects to facilitate meta-analysis; and predictive analyses estimate response of individuals to treatment.
- Heterogeneity of treatment effect analyses should be clearly identified as confirmatory, descriptive, exploratory, or predictive analysis. Evidence should be interpreted in a manner consistent with the inferential goal.
- Descriptive heterogeneity of treatment effect analyses must be prespecified, but they need not be powered to detect heterogeneity of treatment effect. They emphasize estimation of subgroup effects rather than hypothesis testing. They are suitable for meta-analysis because their sampling properties (e.g., standard error) can be characterized.
- Predictive heterogeneity of treatment effect analysis can facilitate optimal treatment decisions for different types of individuals. Independent validation of the prediction algorithm is essential.

Table 1

The essential characteristics of the four different types of heterogeneity of treatment effect analyses

Properties	Confirmatory heterogeneity of treatment effect analysis	Descriptive heterogeneity of treatment effect analysis	Exploratory heterogeneity of treatment effect analysis	Predictive heterogeneity of treatment effect analysis
Inferential goal	To test hypotheses related to subgroup effects	To estimate and report treatment effects for future synthesis	To generate hypotheses for further study	To predict probabilities of beneficial and adverse outcomes for individuals
Number of subgroups analyzed	A small number, typically, one or two	Moderate to large	Not made explicit but may be large	Not applicable
Scientific rationale and prior evidence for hypotheses	Strong	Immaterial	Weak or none	Not applicable but independent validation is essential
Prespecification of data analytic strategy	Fully prespecified	Fully prespecified	Not prespecified	Not prespecified
Control of family wise type I error probability	Should be done	Not needed	Difficult because it is not obvious how many related tests were performed	Not applicable
Characterization of sampling properties of the statistical estimator (e.g., standard errors)	Easy to achieve	Possible	Difficult	Difficult
Power for testing hypothesis	Ideally, study should be adequately powered	Likely to be inadequately powered, but this is immaterial	Typically, inadequate power to examine several hypotheses	Not applicable