Published in final edited form as:

J Clin Epidemiol. 2018 August ; 100: 22–31. doi:10.1016/j.jclinepi.2018.04.005.

Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research

Catherine R. Lesko^a, Nicholas C. Henderson^b, and Ravi Varadhan^{b,c}

^aDepartment of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St., Baltimore, MD 21205

^bDivision of Biostatistics and Bioinformatics, Sidney Kimmel Cancer Care Center, Johns Hopkins School of Medicine, 550 N. Broadway, suite 1111-E, Baltimore, MD 21205

^cDepartment of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St., Baltimore, MD 21205

Abstract

When baseline risk of an outcome varies within a population, the effect of a treatment on that outcome will vary on at least one scale (e.g., additive, multiplicative). This treatment effect heterogeneity is of interest in patient-centered outcomes research. Based on a literature review and solicited expert opinion, we assert: 1) Treatment effect heterogeneity on the additive scale is most interpretable to healthcare providers and patients using effect estimates to guide treatment decision making; heterogeneity reported on the multiplicative scale may be misleading as to the magnitude or direction of a substantively important interaction. 2) The additive scale may give clues about sufficient-cause interaction, although such interaction is typically not relevant to patients' treatment choices. 3) Statistical modeling need not be conducted on the same scale as results are communicated. 4) Statistical testing is one tool for investigations, provided important subgroups are identified a priori, but test results should be interpreted cautiously given non-equivalence of statistical and clinical significance. 5) Qualitative interactions should be evaluated in a pre-specified manner for important subgroups. Principled analytic plans that take into account the purpose of investigation of treatment effect heterogeneity are likely to yield more useful results for guiding treatment decisions.

Keywords

Interaction; Patient-Centered Outcomes Research; Treatment Effect Heterogeneity; Subgroup Analysis

Many epidemiologic investigations (including randomized trials) aim to describe the effect of a (set of) treatment(s) on some outcome to inform clinical practice or policy decisions. Populations to whom clinical guidelines or policies would be applied are composed of individuals with heterogeneous characteristics and comorbidities that confer varying baseline risk of the outcome (i.e., risk in the absence of any treatment). When baseline risk

varies, the treatment effect is likely to vary on at least one scale (e.g., additive, multiplicative). In many instances, and particularly for patient-centered outcomes research, this treatment effect heterogeneity is of substantive interest. Even treatments acknowledged to be widely beneficial (e.g., airbags for injury during a car crash) may be harmful in a subset of the population (e.g., children in the front seat) (2, 3). In other instances, subgroups may not experience any or as strong a benefit from treatments, either because they are not at risk of the outcome (in an extreme example, consider that HIV-uninfected patients will experience no survival benefit from antiretroviral therapy) or because their risk of the outcome is low (as with patients with stable coronary heart disease but low 10-year mortality risk based on a range of clinical and angiographic variables (4)). A main goal of patient-centered outcomes research is to help individuals make informed decisions on whether a particular treatment will work for them. Because it is not possible to directly observe or estimate individual level causal effects (5), one way this is operationalized is by identifying individual-level factors that contribute to treatment effect heterogeneity. Herein, we summarize the literature and expert opinion on several considerations when analyzing treatment effect heterogeneity in the context of patient-centered outcomes research.

Methods

We searched the National Library of Medicine Books, National Library of Medicine Catalog, Current Index to Statistics database, ISI web of science, and websites of 25 major regulatory agencies and organizations for papers and guidelines on study design, analysis and interpretation of treatment effect heterogeneity. Because there is not standard terminology for this topic, a structured search strategy was not sensitive nor specific and we found many resources through “snowball” searching, that is, reviewing citations in, and citations of, key methodological and policy papers.

During the literature review, we identified five key questions relevant to investigations of treatment effect heterogeneity, which we posed to a group of statistical and methodological experts during a focus group conference call: 1) What is the most relevant effect scale for patient-centered outcomes research?; 2) Does the distinction between statistical and mechanistic interaction matter, and if so, which type is most relevant?; 3) Should we try to find a transformation that minimizes heterogeneity, at least when conducting data analysis?; 4) Should we test for qualitative treatment effect heterogeneity?; and 5) What overall strategic approach is recommended for investigating treatment effect heterogeneity? Experts were identified as having an established reputation in work related to: Bayesian methodology; subgroup analysis; or clinical research methodology. We invited 15 potential experts to participate in the focus group; 14 accepted. Focus group participants are listed in the Acknowledgments of this paper. Prior to the focus group, experts were provided with a summary of the literature review described above. Experts discussed responses to each of these questions during the focus group and then individually submitted their final recommendations via email. By first allowing for group discussion on the questions, experts’ final recommendations were quite similar.

We summarize findings from the literature and experts' recommendations below by topical area. Finally, we briefly review analytic advancements for detection and reporting of clinically relevant treatment effect heterogeneity.

Definition of terms

Formal definition of terms appears in Appendix A. Briefly, however, we define a treatment effect as a comparison of some function of potential outcomes. A potential outcome is the outcome that would be observed if, possibly contrary to fact, an individual was exposed to a particular level of treatment; each individual has as many potential outcomes as there are levels of treatment. We rely on potential outcomes to make explicit the *causal* nature of the research questions of interest. Several assumptions are required to identify the contrast of potential outcomes using observable data. A sufficient set of assumptions are often met by design in a randomized trial (8, 9) and can sometimes be plausible in observational studies (10, 11).

The field of patient-centered outcomes research is often loosely interpreted as aiming to predict which treatment will work in which patients. However, we cannot estimate individual treatment effects because we can never observe both potential outcomes in the same person (5). The best we can do is to report expected treatment effect for patients *similar to* the individual faced with a treatment decision. Treatment effect heterogeneity is present when the causal contrast of interest varies across subgroups. Within whatever subgroups we define, however, there are likely to be patients for whom treatment is more or less effective, due to some additional factor. Indeed, it is likely to be 'turtles all the way down,' so to speak, with respect to residual treatment effect heterogeneity within investigator-defined subgroups. At some level, it will be impossible or impractical to further describe residual treatment heterogeneity, either due to insufficient data within a subgroup or due to challenges identifying or measuring the factors responsible for the heterogeneity.

The recommended approach to assessing treatment effect heterogeneity is to model the statistical interaction between the treatment and patient characteristics that define subgroups using a product term. The statistical interaction is *quantitative* if the subgroup-specific effects are different (the coefficient on the product term is non-zero) but both effects are in the same direction (i.e., both suggest harm or benefit) and *qualitative* (sometimes termed *crossover interaction*) if one subgroup-effects is harmful and one is beneficial (12, 13). Statistical interaction is *semi-qualitative* if treatment effect is clinically significantly harmful or beneficial in one subgroup while in the other subgroup the effect is null. VanderWeele and Knol call this type of interaction "pure" interaction (14). Prior work has not distinguished semi-qualitative interaction, perhaps because, given enough precision, the likelihood that an estimated effect would be exactly equal to the null value is small or because distinctions between clinical and statistical significance were ignored. Whether a subgroup effect is determined to be null will be context specific. We address issues of whether statistical tests are appropriate for detecting treatment effect heterogeneity in a subsequent section.

The link function and distributional form of the statistical model typically determines the scale on which treatment effect heterogeneity is measured and tested. In the presence of varying baseline risk and a non-null average treatment effect in at least one strata of

participants, it is a mathematical inevitability that there will be treatment effect heterogeneity on at least one scale, although the clinical relevance of the heterogeneity might be negligible. A quantitative interaction may be nullified by a transformation of the effect scale (although standard transformations may not suffice). A qualitative or semi-qualitative interaction cannot be removed by a scale transformation; it will be present on any scale (whether or not this statistically distinguishable is a different issue). We argue that identifying qualitative and semi-qualitative interactions is of fundamental importance in the context of patient-centered outcomes research.

Choice of analytic scale

Our panel of experts agreed that effect estimates should be reported on a scale easily interpretable by physicians and patients who will use them to make treatment decisions. Physicians and patients (and humans) most readily understand benefits or harms of a particular treatment when results are presented on the absolute (risk difference) scale, preferably alongside a personalized baseline risk estimate (15). Additive treatment effect heterogeneity is also most informative for guiding public health policy that aims to maximize the benefit or minimize the harm of an exposure by targeting subgroups (16). The relative scale (risk ratios or odds ratios) can tend to overstate treatment benefits or harms (17).

There is tension between the scale most appropriate for interpreting results, and the scale most likely to yield a parsimonious model. Recently, two meta-analyses of meta-analyses found tests of homogeneity of the risk difference were rejected more often than were tests of homogeneity of the risk ratio (18, 19). Although this finding is in part due to the difference in power between the two tests (and the fact that they are testing different null hypotheses) (20–22), differential geometry demonstrates that the homogeneity space is indeed smallest for the risk difference and largest for the odds ratio (23).

A consequence of this finding is that recommendations regarding choice of scale for analysis and reporting may conflict (24, 25). For example, the Cochrane Handbook for Systematic Reviews of Interventions states “we desire a summary statistic that gives values that are similar for all the studies in the meta-analysis and subdivisions of the population to which the interventions will be applied” and “the summary statistic should be easily understood and applied by those using the review (25).” The former recommendation would favor relative effect measures, while the latter would favor absolute measures. The European Medicines Agency recommended that “exploration of interactions...in subgroups proceeds first on the scale on which the endpoint is commonly analyzed, with supplementary analyses presented on the complementary scale for those...subgroups that become important for the risk-benefit decision (24).” The “scale on which the endpoint is commonly analyzed” is likely to be the multiplicative scale, either because of historical influence (26) or statistical convenience. As a result, important treatment effect heterogeneity that is present in the complementary scale (additive) might be overlooked because no treatment effect heterogeneity was detected on the primary analytic scale (multiplicative). An instance of this was observed in Kovalchik et al. (2012) (27).

In general, the experts asserted that the analytic model should not dictate how results are reported. They suggested using the most parsimonious, optimally predictive analytical model to predict patient-specific (subgroup-specific) outcomes under each treatment, and then reporting contrasts of outcomes on the additive scale. Given the scale dependence of treatment effect heterogeneity, some have proposed that both multiplicative and additive interactions could be reported (28).

We note briefly that risks are functions of time (e.g., 1-year, 5-year and 10-year risks will differ) and treatment effect heterogeneity for comparisons may be a function of the time-horizon of the analysis as much as of the analytic scale used to model risk. Finally, varying rates of competing events (events that preclude the occurrence of the event of interest) across subgroups may result in treatment effect heterogeneity that may or may not be of interest because the risk of the competing event bounds the maximum risk achievable in a particular subgroup (29).

Non-binary Outcomes

Much of the work on scale dependence of treatment effect heterogeneity has focused on binary outcomes, although the issue is present for other types of outcomes too. For continuous outcomes (e.g., blood pressure), where normality assumption is tenable, there is classical literature on analysis of interactions (13, 30). Interpretation of interaction effects is straightforward for untransformed continuous outcomes. However, interpretation can be challenging when transformation of the outcome variable (e.g., logarithmic or Box-Cox-type transformations) becomes necessary, either for variance stabilization or to satisfy other modeling requirements. Occasionally, the most interpretable scale of a continuous variable will be one that is transformed, e.g., $\log_{10}$ HIV RNA copies/mL is easily understood by clinicians.

For time-to-event outcomes, Cox multiplicative hazard model is popular. However, hazard ratios are not easily interpretable to stakeholders (particularly in the presence of competing risks) (31, 32). One alternative set of models is the accelerated failure time model which regresses the logarithm of the survival time over the covariates (33). Coefficients from the accelerated failure time model denote the differences in the logarithm of time-to-event, which may be more interpretable to physicians and patients than hazard ratios. Another alternative is to model the hazards additively. Additive hazard models yield differences in hazards directly (34). However, communicating the results in terms of hazard of failure is still challenging. A more easily interpretable choice of scale is the restricted mean survival time (RMST) (35, 36). For a chosen time point t , the RMST can be interpreted as the expected survival time of patients who were followed-up until time t , and treatment effect estimates may then be directly reported as differences in RMST. RMST is calculated as the area under the survival curve (perhaps estimated by a Kaplan-Meier curve, for example) up to time point t . In addition to allowing treatment effects to be interpreted as a difference in time spent free of the expected outcome, RMST has the advantage of not relying on the assumption of proportional hazards.

Effect measure modification versus mechanistic interaction

Cox and Berrington (13, 37) loosely differentiate three types of covariates involved in interactions with the primary exposure, each with different interpretation and implications. This differentiation is largely context- and research question-specific. The covariate could be: 1) another treatment or a co-exposure; 2) an intrinsic variable, which may be either manipulable or unmanipulable by the investigator (e.g., patient characteristics such as age or sex), but outside the scope of the investigation (e.g., other comorbidities or patient's environment); or 3) a nonspecific factor (e.g., study site) that are not uniquely characterized. In certain circumstances, given enough data and variability, characteristics of nonspecific factors (e.g., staff-to-patient ratio) may be able to be studied explicitly, but more often nonspecific factors should be treated as a source of unexplained variation and the analysis should reflect this either through estimation of random effects or the use of generalized estimating equations (13, 37, 38). Treatment effect heterogeneity due to another exposure/treatment or to an intrinsic covariate corresponds to different modeling strategies and interpretations of results (39).

Causal interaction is of interest when there are two or more treatments (i.e., the covariate is a treatment) and we would like to know how the outcome changes under different combinations of the treatments (39). There is a large epidemiological literature that delineates statistical interaction (alternatively termed effect measure modification) and causal/biological/mechanistic interaction (39–42) (alternatively termed “causal interdependence” (43, 44), “synergy” (43–45), “definite interdependence” (46, 47), “sufficient cause interaction” (45–47)) and defines conditions under which one is indicative of the presence of the other. A full treatment of this literature is beyond the scope of this paper. Consensus, however, is that statistical interactions on the additive scale are most relevant for investigating causal interactions (43, 46–48). If causal interaction is of interest, the most explicit analytic approach would be to estimate joint effects of the two treatments (28, 39, 49). There are three effects of interest for a binary primary and secondary exposure, where the referent is both exposures absent: 1) the effect of the primary treatment in the absence of the second treatment; 2) the effect of the second treatment in the absence of the primary treatment; and 3) the effect of both treatments administered together. Additional causal assumptions are required to interpret estimates of joint effects causally (39).

When the covariate involved in the interaction is not manipulable (e.g., an intrinsic patient characteristic) or manipulation of the covariate is unlikely or not the focus of study (e.g., smoking status or body weight), it is safe to restrict ourselves to questions about treatment effects within strata of the covariate (28, 39, 49). There are two stratum specific effects for a binary exposure and covariate: 1) the effect of the primary treatment in the absence of the covariate; and 2) the effect of the primary treatment in the presence of the covariate (where the referent is patients who did not receive treatment, but had the covariate). While additional causal assumptions are not needed to identify the effect of the covariate (since it is not estimated), confounding of stratum-specific effect estimates can limit generalizability of results (50). For example, if the effect of a drug varies by body weight, sex-specific effect estimates are likely to vary, but we may not see similar effect sizes in a population in which there is a different distribution of body weight among men and women.

Ultimately, experts reminded us that the theoretical distinction between statistical and causal interaction probably does not matter much to the patient. What is likely most relevant for patient-centered outcomes research is personalized predictions, derived from analyses that account for heterogeneity, but evaluated at the particular covariate profile of an individual patient.

Testing for treatment effect heterogeneity

Confirmatory testing and exploratory analyses for treatment effect heterogeneity should be clearly delineated (51, 52). Subgroups for confirmatory testing should be specified a priori to avoid spurious conclusions (24, 51), particularly because the role of bias and variability is often under-estimated when subgroup effects are interpreted a posteriori (53). Anticipated qualitative interactions, especially, should be specified a priori because their existence is less plausible and more likely to be spurious, if found, than quantitative interactions (54). Likewise, subgroups for which there is an a priori hypothesis about heterogeneity based on a causal mechanism should be prioritized for investigation (24). Using a validated prognostic score (i.e. for risk of the outcome in the absence of treatment) instead of individual covariates may increase power to detect meaningful heterogeneity, and improve interpretation of results (55). Exploratory analyses of subgroup effects are encouraged (51), but regulatory decisions or treatment guidelines are unlikely to be based on exploratory analyses in the absence of replication (24).

Statistical tests are helpful tools, but should not be relied on exclusively. Rothwell may have summarized it best when he said, “the best test of the validity of subgroup analyses is not significance, but replication (52).” There are many reasons to be cautious when testing for treatment effect heterogeneity in a single study. Qualitative treatment effect heterogeneity may be present and important even if there is not sufficient power to reject the null hypothesis that stratum specific effects are equivalent (12, 56–59). In contrast, in large data sets (e.g., healthcare databases) treatment effect heterogeneity may be statistically significant but not clinically significant. Stratum-specific effect estimates, and not just the results of statistical tests, should be reported (24), although it should be recognized that such estimates can be highly unstable due to small sample sizes (we discuss Bayesian approaches in the next section for obtaining stable stratum-specific estimates). Interpretation of the presence or absence of treatment effect heterogeneity should be undertaken with caution, and in the context of prior evidence (51).

Recommendations for analytic strategy

Ultimately, the analytic strategy for evaluating treatment effect heterogeneity will depend on the research goals. Is the goal to identify subgroups in which the treatment may have strongest effects or to identify subgroups likely to receive little or no benefit from treatment? That is, is the goal to define indications or counter indications for treatment? Perhaps the goal is to compare multiple treatment options and come up with treatment recommendations, based on patient characteristics?

To summarize above sections with regards to analytic recommendations: 1) assessment of treatment effect heterogeneity is scale specific and choice of scale should be purposeful,

keeping in mind physicians and patients who will need to interpret results; 2) scale for modeling and scale for reporting need not be the same; statistical modeling should be done on whichever scale best fits the data, whereas treatment effects should be reported on a scale that is of most relevance to stakeholders, primarily patients and physicians; 3) the distinction between statistical and causal interaction is not relevant in the context of treatment choice for the patient but may drive analytic approach; 4) qualitative interactions, although less likely, are highly important and should be evaluated for important, a priori-specified subgroups (while the Gail-Simon test (11) is simplest and most widely known, there are a number of other approaches to assessing qualitative interactions (56–59)); and 5) statistically significant treatment effect heterogeneity is only meaningful when the magnitude of interaction is comparable to the magnitude of the overall treatment effect, i.e., when it is clinically significant as well.

A Bayesian approach addresses many common concerns with subgroup analysis while also providing more informative characterizations of treatment effect heterogeneity (6). Theoretically, the Bayesian framework is well suited to examining treatment effect heterogeneity because it assumes that the treatment effects are random variables having an underlying distribution, rather than being fixed and deterministic. A vital feature of most Bayesian approaches to subgroup analysis is the inclusion of all subgroup-level treatment effects in a single, unified model, which allows inferences in each subgroup to be informed by all the patients in the study rather than only the patients in that particular subgroup. This stabilizes highly variable subgroup effect estimates and increases precision by “borrowing information” from other subgroups (60, 61). Furthermore, Bayesian hierarchical models can limit the probability of finding extreme results in some subgroups by pulling subgroup effects toward the average treatment effect, and thus mitigate against high false positive rate (51). Bayesian skeptical priors or penalized maximum likelihood can shrink interaction terms if the effective sample size does not support the pre-specified number of interaction parameters. Another important advantage of Bayesian analyses is that they can readily facilitate the conversion of results from one treatment effect scale to another, for example, from log-odds ratio to risk difference.

Regardless of the estimation strategy employed, models should include flexibly-modeled main effects and clinically pre-specified interactions. Our experts recommended prizing parsimony and predictive accuracy over model fit, and saw model selection as a process distinct from reporting interpretable results. Fitted models could be used to generate predicted outcomes and those outcomes should then be contrasted to present both relative and absolute treatment effects as a function of background risk and covariates. When there is a validated prognostic risk score, for example for prediction of coronary heart disease (62, 63), this might include plotting the treatment effect as a function of the prognostic risk score (51). This approach is particularly useful when there are multiple covariates that determine pre-treatment risk of the outcome since multiple cross stratification would result in an unwieldy number of strata. We caution, however, that if the same data is used to estimate predictive models to generate a risk score, and to estimate the treatment effects within strata of that risk score, this can potentially lead to bias from overfitting even in moderately large sample sizes. This overfitting problem can be overcome by choosing model form using split samples and cross-validation.

Discussion

There are other considerations in evaluating treatment effect heterogeneity that we have not explored. For example, the question remains, upon which covariates should we stratify to look for treatment effect heterogeneity? There are practical considerations, for example, if treatment effect heterogeneity is reported across strata defined by demographics or comorbidities that are easily measureable or commonly captured, treatment algorithms based on those covariates are less costly to implement, both in terms of financial costs and in terms of possible additional harm to the patient (e.g., from infection following a biopsy (64).) However, recent developments in the area of generalizability and transportability of treatment effects should invite caution (65–68). The assumption that subgroup effects are transportable to future patient populations may not hold (50) and examples abound of false-positive reports of treatment effect heterogeneity that led to mistreatment or under-treatment of countless patients before subsequent investigations failed to confirm earlier reports (52). Generalizability and transportability are not guaranteed and should be considered explicitly (66, 67). While reliable estimates are key for improving clinical care, estimates relevant to patients are likewise worth pursuing (69).

While some patients may react differently to treatment A compared to treatment A' , it is also possible that patients may react differently to A compared to A' at different times, given different time-varying characteristics (e.g., biomarkers, comorbidities, disease severity, etc. (24)). For simplicity, herein we implicitly restricted ourselves to principles for guiding decisions regarding whether or which treatment to initiate based on static patient characteristics (we did not index covariates by time). Decisions regarding when to initiate or switch treatments should be guided by many of the same principles, but may be complicated by considering covariates affected by prior treatment (or absence of treatment) (70, 71).

Ultimately, while studying treatment effect heterogeneity is important for understanding which subgroups of patients may be more or less likely to benefit from an intervention, for providing clues as to the mechanism of action of a particular treatment, and for providing insights into possible sets of interventions that would result in synergy if implemented jointly, applying these insights to patient care may be best served by taking an additional step. If the goal of the study is to integrate knowledge of treatment effect heterogeneity into guidelines for treatment – either a set of indications for or contraindications against treatment, or a set of guidelines to choose between several available treatment options – a different set of methods is required. Investigations of treatment effect heterogeneity seek to identify subgroups across which there are the largest differences in average treatment effect. Identifying a treatment algorithm has as its goal to maximize the average treatment effect for the population (and thereby also for individuals within that population). A complete treatment of methods for defining treatment algorithms are beyond the scope of this paper, but entail generating decision lists or protocols that maximize the average treatment effect by targeting particular subgroups with specific treatments (64, 72, 73).

Herein we have summarized some best practices and considerations when treatment effect heterogeneity is of interest for patient-centered outcomes research based upon a review of literature and guidance from an expert panel. Given that treatment effect heterogeneity is

always present on at least one scale, ad hoc investigations are generally bound to find it. Principled analytic plans that take into account the purpose of the investigation into treatment effect heterogeneity are likely to yield more useful results for guiding treatment decisions, refining treatment algorithms, understanding treatment mechanisms, and ultimately, improving population and individual health.

Acknowledgements:

Funding: RV and NCH were supported by a Patient-Centered Outcomes Research Institute (PCORI) award (ME-1303–5896) and the National Institutes of Health (NIH) through grant number P30CA006973. CRL was supported in part by National Institutes of Health grants U01 HL121812 and U01 AA020793.

We would like to thank the following members of the PCORI expert advisory panel for their insightful discussions and recommendations: David Banks (Duke University); Scott Berry (Berry Consultants); Brad Carlin (University of Minnesota); Ralph B. D'Agostino (Boston University); Steve Goodman (Stanford School of Medicine); Paul Gustafson (University of British Columbia); Frank Harrell (Vanderbilt University); J. Jack Lee (University of Texas MD Anderson Cancer Center); Roderick Little (University of Michigan); David Matchar (Duke University); Sharon-Lise Normand (Harvard Medical School); David Ohlssen (Novartis); Gene Pennello (U.S. Food and Drug Administration); Gary Rosner (Johns Hopkins University); and Tyler VanderWeele (Harvard University). We would also like to thank Tom Louis (Johns Hopkins University) for his valuable insights and guidance throughout the project.

Appendix A. Notation and definition of terms

Many epidemiologic investigations aim to describe the effect of a treatment A on outcome Y . Assume, for the moment, a trial in which individuals are block-randomized to a binary treatment $A = 0, 1$ conditional on their value of binary baseline covariate $X = 0, 1$. Assume complete follow-up for outcome Y , which can be continuous, binary, or time-to-event (subject to censoring). For a given individual i , the observed data is (X_i, A_i, Y_i) . $Y_i(a)$ denotes the potential outcome for individual i ; that is, $Y_i(a)$ is the outcome that would be observed if i was exposed to treatment a . We follow convention and use capital letters to denote random variables, and lower case letters to denote possible realizations of those random variables.

The treatment effect is a comparison of some function $g[\cdot]$ of the expected value of $Y(a)$ if A is set to 1 versus 0, e.g., $g[E(Y|1)] - g[E(Y|0)]$. Assuming no unmeasured confounding and that the observed outcomes for individuals with $A = a$ are equivalent to their potential outcomes if they had been given $A = a$ (i.e., there are not alternative versions of the treatment that could have been given that would have influenced the outcome) (8, 9), $E[Y(a)] = E(Y|a)$. Both of these assumptions are met by design in a randomized trial and can sometimes be plausible in observational studies (10, 11). The treatment effect can then be represented $\theta = g[E(Y|A=1)] - g[E(Y|A=0)]$.

The field of patient-centered outcomes research is often loosely interpreted as aiming to predict which treatment will work in which patients. However, we cannot predict individual treatment effects $Y_i(1) - Y_i(0)$ because we can never observe both potential outcomes in the same person (5). The best we can do is to report expected treatment effect for patients *similar to* the individual faced with a treatment decision; that is, we can report stratum-specific treatment effects for strata across which treatment effect heterogeneity is present. Treatment effect heterogeneity is present when θ varies across subgroups defined by X :

$$g[E(Y|A = 1, X = 1)] - g[E(Y|A = 0, X = 1)] \neq g[E(Y|A = 1, X = 0)] - g[E(Y|A = 0, X = 0)]$$

The recommended approach to assessing treatment effect heterogeneity is to model the statistical interaction between A and X . This involves fitting the regression model:

$$g[E(Y|A, X)] = \beta_0 + \beta_A A + \beta_X X + \beta_{A,X} AX. \quad (1)$$

In model (1), the coefficient $\beta_{A,X}$ represents the statistical interaction between the treatment A and covariate X . The treatment effects in the two subgroups $X = 0$ and $X = 1$ are, respectively, $\theta_{X=0} = \beta_A$ and $\theta_{X=1} = \beta_A + \beta_{A,X}$. The statistical interaction between A and X is *quantitative* if $\theta_{X=0} = \theta_{X=1}$, but both effects are in the same direction (i.e., both suggest harm or benefit), and is *qualitative* if $\theta_{X=0} \neq \theta_{X=1}$ but effects have opposite signs (12, 13). Statistical interaction is *semi-qualitative* if $\theta_{X=0} = \theta_{X=1}$ if one treatment effect suggests clinically meaningful harm or benefit while the other effect is null (13). Testing $\theta_{X=0} = \theta_{X=1}$ corresponds to rejecting the null hypothesis $H_0: \beta_{A,X} = 0$. We stress that this test does not provide information on the null hypotheses $H_0: \theta_{X=0} = 0$ or $H_0: \theta_{X=1} = 0$.

$g(\cdot)$ determines the scale on which treatment effect heterogeneity is measured and tested. For example, if $g(\cdot)$ is the identity link, assumed to follow a binomial distribution, $\beta_{A,X}$ is the difference of differences and a test of $H_0: \beta_{A,X} = 0$ is a test of departure from perfect additivity of effects. In the presence of varying baseline risk of Y and a non-null average treatment effect in at least one strata of participants, there will be treatment effect heterogeneity on at least one scale. This is a mathematical inevitability (51). To illustrate this point, let us assume that $E(Y|A = 1, X = 0) = 0.04$, $E(Y|A = 0, X = 0) = 0.10$, $E(Y|A = 1, X = 1) = 0.08$ and $E(Y|A = 0, X = 1) = 0.20$. The relative risk for both strata of X is 2 indicating absence of multiplicative treatment effect heterogeneity, but among the individuals with $X = 0$ the risk difference is 0.06 while among individuals with $X = 1$ the risk difference is 0.12. That is, there is additive treatment effect heterogeneity. A quantitative statistical interaction can often be nullified by a transformation of the effect scale, i.e., by using a different link function, $g(\cdot)$. A qualitative statistical interaction cannot be removed by a scale transformation; it will be present on any scale.

References

1. SOLVD Investigators, Yusuf S, Pitt B, et al. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med* 1991;325(5):293–302. [PubMed: 2057034]
2. Olson CM, Cummings P, Rivara FP. Association of first- and second-generation air bags with front occupant death in car crashes: a matched cohort study. *Am J Epidemiol* 2006;164(2):161–9. [PubMed: 16775043]
3. Newgard CD, Lewis RJ. Effects of child age and body size on serious injury from passenger air-bag presence in motor vehicle crashes. *Pediatrics* 2005;115(6):1579–85. [PubMed: 15930219]
4. Yusuf S, Zucker D, Peduzzi P, et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet* 1994;344(8922):563–70. [PubMed: 7914958]
5. Holland PW. Statistics and Causal Inference. *J Am Stat Assoc* 1986;81(396):945–60.

6. Henderson NC, Louis TA, Wang C, et al. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Serv Outcomes Res Method* 2016;16(4):213–33.
7. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA : the journal of the American Medical Association* 2007;298(10):1209–12. [PubMed: 17848656]
8. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health* 2006;60(7):578–86. [PubMed: 16790829]
9. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009;20(6):880–3. [PubMed: 19829187]
10. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19(6):766–79. [PubMed: 18854702]
11. Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;155(2):176–84. [PubMed: 11790682]
12. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985;41(2):361–72. [PubMed: 4027319]
13. Cox DR. Interaction. *International Statistical Review* 1984;52(1):1–24.
14. VanderWeele TJ, Knol MJ. A tutorial on interaction. *Epidemiol Method* 2014;3(1):33–72.
15. Fagerlin A, Zikmund-Fisher BJ, Ubel PA. Helping patients decide: ten steps to better risk communication. *J Natl Cancer Inst* 2011;103(19):1436–43. [PubMed: 21931068]
16. Blot WJ, Day NE. Synergism and interaction: are they equivalent? *Am J Epidemiol* 1979;110(1):99–100. [PubMed: 463868]
17. Poole C. Coffee and myocardial infarction. *Epidemiology* 2007;18(4):518–9.
18. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002;21(11):1575–600. [PubMed: 12111921]
19. Engels EA, Schmid CH, Terrin N, et al. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med* 2000;19(13):1707–28. [PubMed: 10861773]
20. Poole C, Shrier I, VanderWeele TJ. Is the Risk Difference Really a More Heterogeneous Measure? *Epidemiology* 2015;26(5):714–8. [PubMed: 26196684]
21. VanderWeele TJ. Sample Size and Power Calculations for Additive Interactions. *Epidemiol Method* 2012;1(1):159–88. [PubMed: 25473594]
22. White IR, Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions? *BMC Med Res Methodol* 2005;5:15. [PubMed: 15862124]
23. Ding P, VanderWeele TJ. The Differential Geometry of Homogeneity Spaces Across Effect Scales. *arXiv preprint arXiv:151008534* 2015.
24. Guideline on the investigation of subgroups in 4 confirmatory clinical trials. 2014, Committee for Medicinal Products for Human Use (CHMP). European Medicines Agency.
25. Cochrane Handbook for Systematic Reviews of Interventions. Higgins JPT, Green S, eds. The Cochrane Collaboration, 2011.
26. Poole C On the origin of risk relativism. *Epidemiology* 2010;21(1):3–9. [PubMed: 20010205]
27. Kovalchik SA, Varadhan R, Fetterman B, et al. A general binomial regression model to estimate standardized risk differences from binary response data. *Stat Med* 2013;32(5):808–21. [PubMed: 22865328]
28. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol* 2012;41(2):514–20. [PubMed: 22253321]
29. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol* 2009;170(2):244–56. [PubMed: 19494242]
30. Scheffé H. The analysis of variance. New York,: Wiley; 1959.
31. Allignol A, Schumacher M, Wanner C, et al. Understanding competing risks: a simulation point of view. *BMC Med Res Methodol* 2011;11:86. [PubMed: 21639902]
32. Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010;21(1):13–5. [PubMed: 20010207]

33. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med* 1992;11(14–15):1871–9. [PubMed: 1480879]
34. Rod NH, Lange T, Andersen I, et al. Additive interaction in survival analysis: use of the additive hazards model. *Epidemiology* 2012;23(5):733–7. [PubMed: 22732385]
35. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 2013;13:152. [PubMed: 24314264]
36. Zhao L, Claggett B, Tian L, et al. On the restricted mean survival time curve in survival analysis. *Biometrics* 2016;72(1):215–21. [PubMed: 26302239]
37. Berrington de Gonzles A, Cox DR. Interpretation of interaction: A review. *The Annals of Applied Statistics* 2007;1(2):371–85.
38. Hanley JA, Negassa A, Edwardes MD, et al. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003;157(4):364–75. [PubMed: 12578807]
39. VanderWeele TJ. On the distinction between interaction and effect modification. *Epidemiology (Cambridge, Mass)* 2009;20(6):863–71.
40. Varadhan R, Seeger JD. Estimation and Reporting of Heterogeneity of Treatment Effects In: Velentgas P, Dreyer NA, Nourjah P, et al., eds. *Developing a Protocol for Observational Comparative Effectiveness Research: A User’s Guide*. Rockville, MD: Agency for Healthcare Research and Quality, 2013.
41. Ahlbom A, Alfredsson L. Interaction: A word with two meanings creates confusion. *European Journal of Epidemiology* 2005;20(7):563–4. [PubMed: 16119427]
42. Vanderweele TJ. Invited commentary: assessing mechanistic interaction between coinfecting pathogens for diarrheal disease. *Am J Epidemiol* 2012;176(5):396–9. [PubMed: 22842718]
43. Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. *Scand J Work Environ Health* 1988;14(2):125–9. [PubMed: 3387960]
44. Miettinen OS. Causal and preventive interdependence. Elementary principles. *Scand J Work Environ Health* 1982;8(3):159–68. [PubMed: 7156935]
45. Rothman KJ. Causes. *Am J Epidemiol* 1976;104(6):587–92. [PubMed: 998606]
46. VanderWeele TJ. Sufficient cause interactions and statistical interactions. *Epidemiology* 2009;20(1):6–13. [PubMed: 19234396]
47. VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-cause framework. *Epidemiology* 2007;18(3):329–39. [PubMed: 17435441]
48. Greenland S Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology* 2009;20(1):14–7. [PubMed: 19234397]
49. VanderWeele TJ, Knol MJ. Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. *Ann Intern Med* 2011;154(10):680–3. [PubMed: 21576536]
50. Varadhan R, Wang SJ. Standardization for subgroup analysis in randomized controlled trials. *J Biopharm Stat* 2014;24(1):154–67. [PubMed: 24392983]
51. Kent DM, Rothwell PM, Ioannidis JP, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010;11:85. [PubMed: 20704705]
52. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365(9454):176–86. [PubMed: 15639301]
53. Lash TL. Heuristic thinking and inference from observational epidemiology. *Epidemiology* 2007;18(1):67–72. [PubMed: 17149141]
54. Peto R Statistical aspects of cancer trials In: Halnan KE, ed. *Treatment of cancer*. London, UK: Chapman and Hall, 1982:867–71.
55. Hayward RA, Kent DM, Vijan S, et al. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol* 2006;6:18. [PubMed: 16613605]
56. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. *Stat Med* 1993;12(13):1239–48. [PubMed: 8210823]

57. Li J, Chan IS. Detecting qualitative interactions in clinical trials: an extension of range test. *J Biopharm Stat* 2006;16(6):831–41. [PubMed: 17146982]
58. Pan G, Wolfe DA. Test for qualitative interaction of clinical significance. *Stat Med* 1997;16(14):1645–52. [PubMed: 9257418]
59. Bayman EO, Chaloner K, Cowles MK. Detecting qualitative interaction: a Bayesian approach. *Stat Med* 2010;29(4):455–63. [PubMed: 19950107]
60. Jones HE, Ohlssen DI, Neuenschwander B, et al. Bayesian models for subgroup analysis in clinical trials. *Clinical trials* 2011;8(2):129–43. [PubMed: 21282293]
61. Alosch M, Huque MF, Koch GG. Statistical Perspectives on Subgroup Analysis: Testing for Heterogeneity and Evaluating Error Rate for the Complementary Subgroup. *J Biopharm Stat* 2015;25(6):1161–78. [PubMed: 25331097]
62. D'Agostino RB Sr., Grundy S, Sullivan LM, et al. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA : the journal of the American Medical Association* 2001;286(2):180–7. [PubMed: 11448281]
63. Wilson PW, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837–47. [PubMed: 9603539]
64. Zhang Y, Laber EB, Tsiatis A, et al. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* 2015;71(4):895–904. [PubMed: 26193819]
65. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *J Causal Inference* 2013;1(1):107–34.
66. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America* 2016;113(27):7345–52. [PubMed: 27382148]
67. Lesko CR, Buchanan AL, Westreich D, et al. Generalizing study results: a potential outcomes perspective. *Epidemiology* 2017.
68. Bareinboim E, Pearl J. Transportability of Causal Effects: Completeness Results. Presented at
69. Flores L Therapeutic inferences for individual patients. *J Eval Clin Pract* 2015;21(3):440–7. [PubMed: 25511915]
70. Cain LE, Robins JM, Lanoy E, et al. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *Int J Biostat* 2010;6(2):
71. Young JG, Cain LE, Robins JM, et al. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Stat Biosci* 2011;3(1):119–43. [PubMed: 24039638]
72. Luedtke AR, van der Laan MJ. Statistical Inference for the Mean Outcome under a Possibly Non-Unique Optimal Treatment Strategy. *Annals of Statistics* 2016;44(2):713–42. [PubMed: 30662101]
73. Cai TX, Tian L, Wong PH, et al. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 2011;12(2):270–82. [PubMed: 20876663]

Box 1. A Case Study: The SOLVD Treatment Trial

Background:

The studies of left ventricular dysfunction (SOLVD) were a series of trials designed to study the effect of initiating the ACE-inhibitor enalapril on risk of hospitalization or death among persons with congestive heart failure. One of these studies, the SOLVD treatment trial, enrolled 2,569 individuals who had ejection fractions lower than 0.35 and who were suffering from overt congestive heart failure and randomized them 1:1 to enalapril or placebo (1).

Objective:

To explore the magnitude and nature (quantitative/semi-qualitative/qualitative) of variation of treatment effect in pre-specified subgroups of patients

Methods:

We split participants into 12 mutually exclusive multivariate subgroups according to the following covariates measured at baseline: sex (male/female), age (>65/ ≤65 years) and ejection fraction (grouped by tertiles) (1). We performed an unstructured interaction test to test whether or not there is at least one patient subgroup defined by all combinations of baseline patient covariates with a differential treatment effect (6); using a test that utilizes multivariate subgroups is more likely to identify important treatment effect heterogeneity than tests that consider interaction one covariate at a time (7). We performed a Gail-Simon test to investigate qualitative interactions (12). Finally, to focus on estimation of subgroup-specific effects, rather than on testing for the presence or absence of heterogeneity of treatment effects, we fit Bayesian hierarchical models to estimate the hazard ratio using Cox proportional hazards regression, and 2-year survival probabilities using Kaplan-Meier estimates (i.e., both multiplicative and additive scales) (6).

Results:

The unstructured test for interaction on the log-hazard ratio scale yielded a p-value of 0.018. The Gail-Simon test yielded a p-value of 0.85. These two tests together imply the presence of *quantitative* heterogeneity of the log-hazard ratio for hospitalization or death due to initiation of enalapril. Figure 1 shows estimates of the hazard ratio (multiplicative scale) and Figure 2 shows estimates of the difference in 2-year survival probabilities (additive scale) comparing enalapril versus placebo for each of the 12 multivariate subgroups from a fully stratified frequentist model and from a Bayesian hierarchical model. The Bayesian shrinkage estimates are less variable than the fully stratified estimates because Bayesian estimates tend to be pulled or “shrunk” towards the overall hazard ratio. The amount of shrinkage is inversely related to the number of patients within strata. Bayesian estimates suggest that modest variation in treatment effects is present. In particular, male subgroups tend to derive greater benefit than the female subgroups, and groups with high baseline ejection fractions tend to derive less benefit than those with either medium or low ejection fractions, although estimated differences are quite modest. Female subgroups have small sample size, thus subgroup-specific treatment effect estimates are shrunk towards the overall estimate very strongly. In the

absence of precise subgroup information, the Bayesian shrinkage model will typically produce “conservative” (i.e., close to the overall) estimates and associated uncertainty intervals will usually not provide any clear evidence of a differential treatment effect. In this example, the estimate of the difference in 2-year survival among females was 0.01 (95% credible interval: -0.13, 0.15) suggesting there could be semi-qualitative interaction (i.e., the treatment effect is null among women). However, the Gail-Simon test did not suggest presence of any qualitative interactions.

Conclusions:

Although there is evidence of a quantitative effect heterogeneity, and some indication that there may be semi-qualitative treatment effect heterogeneity, small sample sizes prevented us from making definitive conclusions regarding qualitative/semi-qualitative variation in treatment effect.

What is new:

1. Assessment of treatment effect heterogeneity is important for patient-centered outcomes research. Qualitative treatment effect heterogeneity should always be evaluated in a pre-specified manner for important subgroups (e.g., men versus women).
2. Treatment effect heterogeneity should be evaluated on different scales (e.g., multiplicative and additive) because it might be present on one scale, but not on another scale.
3. The scale for the analytic model need not be the same the scale in which results are communicated to stakeholders. While modeling can be done on whichever scale that best fits the data, stakeholders generally prefer to see the results communicated in terms of absolute magnitude of benefit or harm, i.e. risk differences or difference in time-to-event.
4. Statistically significant interactions are meaningful only when the magnitude of interaction is similar to the magnitude of the overall treatment effect. This is especially important in the context of large databases.
5. Bayesian hierarchical modeling is one available analytic strategy with many attractive properties for patient-centered outcomes research.

Figure 1:

Stratum-specific hazard ratios for the effect of the ACE-inhibitor enalapril on time to hospitalization or death in the SOLVD treatment trial, estimated from a fully stratified sample (black dots) and from a Bayesian hierarchical model (red dots) with 95% confidence intervals and 95% credible intervals, respectively, for 12 subgroups of patients defined by tertile of ejection fraction, age (>65 or ≤65) and sex (male/female) all measured at baseline. The solid vertical line represents the overall estimate of the hazard ratio, and the dashed vertical line is placed at one (i.e., no treatment effect).

Figure 2.

Stratum-specific differences in 2-year survival (from hospitalization or death) probabilities due to initiation of the ACE-inhibitor enalapril versus placebo in the SOLVD treatment trial, estimated from a fully stratified sample (black dots) and from a Bayesian hierarchical model (red dots) with 95% confidence intervals and 95% credible intervals, respectively, for 12 subgroups of patients defined by tertile of ejection fraction, age (>65 or ≤ 65) and sex (male/female) all measured at baseline. The solid vertical line represents the overall estimate of 2-year survival, and the dashed vertical line is placed at zero (i.e., no treatment effect).