HSR Health Services Research

DOI: 10.1111/j.1475-6773.2009.01064.x
RESEARCH ARTICLE

Assessing Heterogeneity of Treatment Effects: Are Authors Misinterpreting Their Results?

Erik Fernandez y Garcia, Hien Nguyen, Naihua Duan, Nicole B. Gabler, and Richard L. Kravitz

Objective. To determine whether investigations of heterogeneity of treatment effects (HTE) in randomized-controlled trials (RCTs) are prespecified and whether authors' interpretations of their analyses are consistent with the objective evidence.

Data Sources/Study Setting. Trials published in *Annals of Internal Medicine*, *British Medical Journal*, *Journal of the American Medical Association*, *Lancet*, and *New England Journal of Medicine* in 1994, 1999, and 2004.

Study Design. We reviewed 87 RCTs that reported formal tests for statistical interaction or heterogeneity (HTE analyses), derived from a probability sample of 541 articles.

Data Collection/Extraction. We recorded reasons for performing HTE analysis; an objective classification of evidence for HTE (termed "clinicostatistical divergence" [CSD]); and authors' interpretations of findings. Authors' interpretations, compared with CSD, were coded as understated, overstated, or adequately stated.

Principle Findings. Fifty-three RCTs (61 percent) claimed prespecified covariates for HTE analyses. Trials showed strong (6), moderate (11), weak (25), or negligible (16) evidence for CSD (29 could not be classified due to inadequate information). Authors stated that evidence for HTE was sufficient to support differential treatment in subgroups (10); warranted more research (31); was absent (21); or provided no interpretation (25). HTE was overstated in 22 trials, adequately stated in 57 trials, and understated in 8 trials.

Conclusions. Inconsistencies in performance and reporting may limit the potential of HTE analysis as a tool for identifying HTE and individualizing care in diverse populations. Recommendations for future studies on the reporting and interpretation of HTE analyses are provided.

Key Words. Heterogeneity of treatment effects, HTE analysis, subgroup analysis, individualized care, interaction analysis

Randomized-controlled clinical trials (RCTs), the cornerstone of evidence-based medicine, are designed to estimate average treatment effects. Such estimates provide a good sense of whether treatment is likely to deliver more

benefit than the control in a population but say little about the effects of treatment in individuals who depart from the average. The term *heterogeneity of treatment effects (HTE)* refers to variation in the effects of treatment across individuals, within and outside of clinical trials (Longford 1999; Kravitz, Duan, and Braslow 2004; Kraemer, Frank, and Kupfer 2006; Greenfield et al. 2007).

Subgroup analysis (SGA) is a term loosely applied to mean the practice of investigating differences in treatment outcomes within and/or between subgroups of patients. In a narrow sense, the term *SGA* is used to refer to the assessment of treatment effects within each subgroup and without a formal test of the difference of treatment effects across subgroups. Gabler et al. (2009) referred to this type of SGA as *subgroup-only analysis*. In a broader sense, the term *SGA* is sometimes used to refer to the comparison of treatment effects across subgroups, with or without a formal test for the difference. In order to avoid this ambiguity, we use the term *HTE analysis* (sometimes called moderator or modifier analysis) to refer to the assessment of variations in treatment effects across subgroups using formal tests for interaction or heterogeneity (Kravitz, Duan, and Braslow 2004; Gabler et al. 2009). The focus of this study is on trials that reported HTE analyses, irrespective of the term(s) used for these analyses in the trials themselves.

The utility of HTE analysis is hampered by the problems of multiple testing (which may result in false positives) and insufficient power (which may result in false negatives) (Stallones 1987; Yusuf et al. 1991; Cui et al. 2002; Cook, Gebski, and Keech 2004). Although recommendations for the design, reporting, and interpretation of analyses for HTE have been published (Yusuf et al. 1991; Altman et al. 2001; Lu et al. 2005; Rothwell 2005, 2007; Wang et al. 2007), actual practice does not reflect the recommendations (Assmann et al. 2000; Parker and Naylor 2000; Moreira, Stein, and Susser 2001; Pocock et al. 2002; Hernandez et al. 2006; Wang et al. 2007; Gabler et al. 2009). Deviating from appropriate design and reporting of HTE analyses

---

Address correspondence to Erik Fernandez y Garcia, M.D., M.P.H., Assistant Professor of Clinical Pediatrics, Department of Pediatrics, University of California—Davis, School of Medicine, 2516 Stockton Blvd., Ste. 341, Sacramento, CA 95817; e-mail: erik.fernandez@ucdmc.ucdavis.edu. Hien H. Nguyen, M.D., M.A.S., Associate Clinical Professor, is with the Division of Infectious Diseases, University of California—Davis, School of Medicine, Sacramento, CA. Naihua Duan, Ph.D., Professor of Biostatistics (in Psychiatry), is with the Departments of Psychiatry and Biostatistics, Columbia University, New York, NY. Erik Fernandez y Garcia, M.D., M.P.H., and Nicole B. Gabler, M.P.H., M.H.A., are with the Center for Healthcare Policy and Research, Sacramento, CA. Richard L. Kravitz, M.D., M.S.P.H., Professor and Co-Vice Chair, Research, is with the Division of General Medicine, University of California—Davis, Sacramento, CA.

increases the possibility of misinterpretation. Such misinterpretation is concerning because overstating or understating HTE can lead to inappropriately broad or narrow treatment recommendations and confuse future research priorities.

While previous studies have focused on investigating the design and reporting of subgroup-only analyses and/or HTE analyses (Assmann et al. 2000; Parker and Naylor 2000; Moreira, Stein, and Susser 2001; Pocock et al. 2002; Hernandez et al. 2006; Wang et al. 2007; Gabler et al. 2009), few have investigated interpretation (Assmann et al. 2000; Brookes et al. 2004; Hernandez et al. 2005, 2006; Parker and Naylor 2006), and none has investigated interpretation as the main focus of the study. We sought to evaluate recent practices with respect to the analysis and interpretation of HTE, ranging from one covariate and one interaction to multiple covariates and interaction terms, in RCTs published in prominent general medical journals, where formal tests for interaction or heterogeneity were performed. In so doing, we focused on two research questions. First, what is the prevalence of prespecified (versus ad hoc) HTE analyses and what were the reasons (if any) given for the inclusion of specific covariates? Second, to what extent did authors' interpretation of their HTE-related findings and the recommendations that flow from them match the objective evidence provided for or against the presence of statistically and clinically significant HTE.

METHODS

Data Sources and Searches

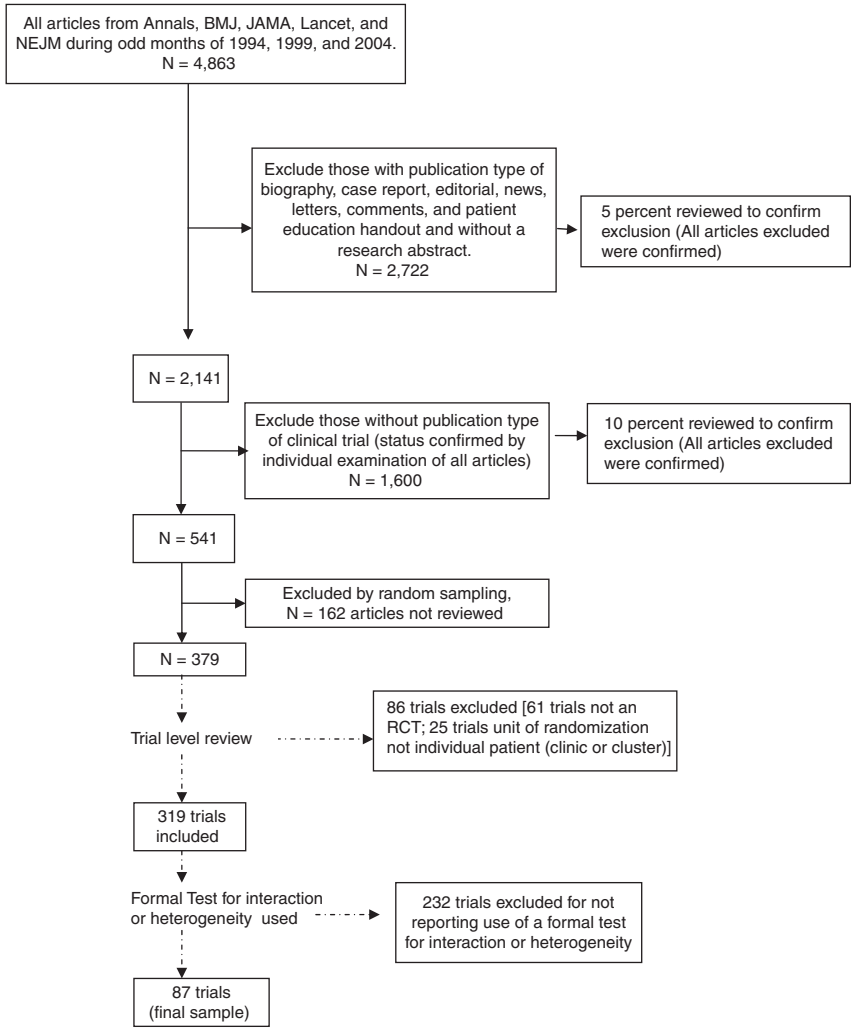
We used PubMed to identify articles reporting RCTs published in the *New England Journal of Medicine*, *Journal of the American Medical Association*, *Annals of Internal Medicine*, *Lancet*, and *British Medical Journal* during odd-numbered months in 1994, 1999, and 2004. A 10-year timeframe was chosen to reflect current practices and to assess for possible trends over time. This initial search produced 4,863 articles. Of these articles, 2,722 were excluded for not having a research abstract. A 5 percent random sample of the 2,722 excluded articles was reviewed by one investigator for inappropriate exclusion and no reports of clinical trials were found. Of the remaining 2,141 articles, a further 1,600 articles without "clinical trial" as a publication type were reviewed by one investigator and excluded for not being clinical trial reports. A second investigator additionally examined a 10 percent random sample of the 1,600 articles and no clinical trial reports were found. The remaining 541 articles on clinical

trials were randomly sorted into nine batches of 54 articles and one batch of 55. To conserve study resources, seven of these batches containing 379 articles were randomly selected. We then examined the trials within these articles, allowing for multiple trials reported within the same article (i.e., related trials with different protocols reported in the same article or reports of trials with factorial designs where each treatment arm's outcomes were compared separately with the author-designated control group's outcomes) to determine final inclusion status for these trials. All trials were reviewed for inclusion and exclusion criteria by two investigators with any disagreement resolved by a third. To be included, a trial had to have the following: (1) a human study population; (2) either a parallel group (including matched pair trials) or crossover (excluding *n*-of-1 trials) RCT design; and (3) the unit of randomization as the individual patient or as the treatment episode within a patient (for crossover trials). We excluded trials that used cluster randomization because these trials frequently focus on organizational-level treatment effects. After trial-level inclusion and exclusion criteria were applied, we identified 319 eligible RCTs reported in 303 articles among the initial random sample of 379 articles. To properly investigate HTE it is necessary to use a formal test for interaction or heterogeneity (Altman et al. 2001; Brookes et al. 2001; Cui et al. 2002; Pocock et al. 2002). Given previous documented evidence of low rates of use and in an effort to maximize the possibility of appropriate interpretation (Assmann et al. 2000; Parker and Naylor 2000; Altman et al. 2001; Brookes et al. 2001; Moreira, Stein, and Susser 2001; Cui et al. 2002; Pocock et al. 2002; Hernandez et al. 2006; Wang et al. 2007; Gabler et al. 2009), we restricted our study to those trials which used such formal tests. Therefore, of these 319 trials, 87 (27 percent) reported using a formal test for interaction or heterogeneity and this is consistent with previous reports (Assmann et al. 2000; Hernandez et al. 2006; Wang et al. 2007; Gabler et al. 2009). These 87 trials formed our final sample (Figure 1).

Data Extraction

Two investigators independently reviewed each trial with any disagreement resolved by consensus with a third investigator. (Disagreements were noted [and resolved] in 14/87 trials and involved identification of the given trial's primary outcome or covariates used for HTE analyses). We used a standard protocol, form, and database to collect the following general information: article identification number; trial identification number (if more than one trial was reported per article); and first author name. Our primary outcome was

Figure 1: Article Selection for Systematic Review of Heterogeneity of Treatment Effect in Randomized-Controlled Trial (RCTs) Published in Five General Medical Journals



whether authors' interpretations of HTE analyses, and the recommendations based on them, were supported by objective evidence. To do this, we started by recording the number, general type (e.g., morbidity, mortality, composite, etc.), and verbatim description of the primary trial outcomes. Next, we

recorded the effect measures (i.e., event rates, absolute difference in event rates, relative risk, odds ratio, hazard ratio, *p*-value, and 95 percent confidence intervals) of the primary outcomes and whether they were declared significant, not significant, or not reported. We made special note of cases where the authors only reported effect measures different than those explicitly stated as the primary outcome (e.g., primary outcome stated as mortality but effect measures only given for survival). We then examined which covariates were examined in the HTE analyses and coded whether specific reasons were given for the inclusion of all, some, or no covariates. We recorded if the inclusion of a given covariate was explicitly described by the authors as prespecified but without other more specific reasons for the inclusion; whether the reasons given for the inclusion of covariates were a priori hypotheses based on theory or prior data; or whether the inclusion was based on formal statistical model selection or variable selection, guided exploration (where a nonarbitrary guiding principle for covariate selection was provided but the guiding principle was not a statistical model selection or variable selection procedure), or exploratory and without guiding principles.

Our next step was to examine the analyses of effect measures in subgroups defined by the covariates identified. We recorded the primary outcome associated with these analyses, the covariates examined, whether the ratio measure of effect in any subgroup was at least 25 percent greater or smaller than the main effect reported in the primary analyses (clinically divergent), and whether the test for interaction or heterogeneity used in the analyses was associated with a *p*-value of .10 or less (statistically significant). While imposing specialty-specific standards for clinically important relative differences in effect for each trial would be ideal, such specificity was beyond the scope of this review; 25 percent was chosen as a clinically relevant difference across specialties and conditions. We chose $p \leq .10$ as the cutoff for statistical significance of the interaction because this value is commonly accepted in the literature (Lu et al. 2005) for detecting interactions and because sample size requirements for achieving adequate power to detect important effects under more stringent significance levels are seldom met. We also noted the significance level that authors used for hypothesis testing. A special circumstance occurred if the result of the test for interaction or heterogeneity was declared "not significant," but the actual test result was not provided and the level of significance used by the authors was not reported or reported as .05 (in which case a *p*-value between 0.51 and .10 could not be detected). These cases were deemed "unclassifiable." We used these data to classify studies as demonstrating no, weak, moderate, or strong evidence for clinicostatistical

Table 1: Scheme Used to Code the Combination of Clinical Divergence and Statistical Significance in Determining Strength of Evidence for Clinicostatistical Divergence (CSD) as an Objective Measure of the Presence or Absence of Heterogeneity of Treatment Effects (HTE)

Clinical Divergence Present?	*Statistical Significance Present?*	*Strength of Evidence for Clinicostatistical Divergence*
No	No	None
No	*	None
*	No	None
Yes	No	Weak
Yes	*	Weak
No	Yes	Weak
*	Yes	Weak
Yes	Yes (without any prespecified covariates)	Moderate
Yes	Yes (with at least one prespecified covariate)	Strong
*	*	Unable to classify

*Data were not provided in the article.

divergence (CSD). In the primary analysis, a trial was declared as showing strong evidence for CSD if the analysis was prespecified, the p-value for interaction or heterogeneity that was $\leq .10$, and the relative effect measure in any subgroup was 25 percent higher or lower than the average effect for the sample as a whole (Table 1). Such evidence was thought to be sufficient to sway most clinicians (and most guidelines committees) to consider treating patients with the identified characteristic(s) differently than the average patient. Graded departures from this standard of prespecification, statistical significance, and clinical divergence were coded as moderate, weak, or no evidence for CSD (Table 1).

As an example of this classification scheme, consider the trial reported by Drakulovic et al. (1999), a comparison of semirecumbent (intervention group) versus supine (control or usual care group) body position for prevention of clinically suspected nosocomial pneumonia (primary outcome) in ventilated intensive care unit patients. The primary analysis in the overall study sample was reported by the authors as significantly in favor of the semirecumbent position for the prevention of clinically suspected pneumonia (8 percent in the semirecumbent group versus 34 percent in the supine group, 95 percent CI for difference 10–42, $p = .003$). In regard to the HTE analysis, the authors clearly stated that the HTE analysis was preplanned and that "all risk factors (covariates used to define subgroups) tested in this analysis

had been previously described in ICU patients," and as such were classified as being based on a priori hypotheses. The authors reported statistically significant interaction between enteral feeding (the subgroup-defining covariate) and body position (intervention) in the frequency of clinically suspected pneumonia ($p<.001$). Clinical divergence between the enteral feeding group's intervention effect (41 percent: semirecumbent position 9 percent versus supine position 50 percent) and the nonenteral feeding group's intervention effect (4.6 percent: semirecumbent position 5.9 percent versus supine position 10.5 percent) in the development of clinically suspected pneumonia was also evident, with at least 25 percent difference on a relative scale from the overall intervention effect (26 percent). Therefore, the trial was classified as showing strong evidence for CSD as there was both statistical significance ($p<.10$) and clinical divergence in the presence of a prespecified HTE analysis with subgroups defined by a covariate with a priori rationale for an expected moderation of intervention effect.

Lastly, we examined the authors' interpretations of their HTE analyses and recorded verbatim descriptions of these interpretations, and any treatment recommendations that flowed from them. Two investigators coded these interpretations as showing the following: (1) evidence for HTE sufficient to support different treatment recommendations in one or more subgroups; (2) evidence for HTE insufficient to support differential treatment recommendations but sufficient to warrant further systematic research; (3) some evidence for HTE but not enough to warrant differential treatment recommendations or further systematic research; (4) evidence that HTE was absent; or (5) no interpretation of HTE-related results. Interrater agreement on this categorization was 100 percent. The most straightforward and clear statements were coded. If more than one covariate was investigated for HTE and different interpretations were given for each separate analysis, we chose to code that interpretation–covariate pair with the least discordance between interpretation and objective evidence for CSD; this would tend to bias the analysis in favor of investigators.

Data Analysis

We entered and stored data in a Microsoft Access (2000) database and produced descriptive statistics using *SAS* version 9 (SAS Institute Inc.). We compared our study outcomes by year of trial publication using Pearson Chi-square in *Stata* version 10 (StataCorp LP).

We performed two sensitivity analyses. First, we redefined statistical significance using a *p*-value of $\leq .05$ instead of $\leq .10$. Second, acknowledging the possible need for a clinically relevant and standardized measure of effect size when comparing different RCTs (Cook and Sackett 1995; Altman and Andersen 1999; Kraemer and Kupfer 2006), we attempted to calculate number needed to treat (NNT) for each trial based on reported data. We then redefined clinical divergence by whether the NNT in any subgroup was at least 25 percent greater or smaller than the NNT calculated for the entire sample. If we were unable to calculate NNT based on the information provided, these trials were deemed "unclassifiable" in terms of clinical divergence. The modified definitions of statistical significance and clinical divergence were then used to reclassify the trials' strength of evidence for CSD. We then compared the authors' interpretations with the modified definitions of CSD.

RESULTS

Prespecification of and Rationale for Covariates

Fifty-three of the 87 trials (61 percent) reported that HTE analyses were prespecified. Of the 53 trials reporting prespecified HTE analyses, 17 gave specific reasons for the inclusion of all covariates in the HTE analyses, 12 gave specific reasons for some covariates, and 24 gave no reason for the inclusion of any covariates. Therefore, overall only 29 of 87 trials (33 percent) provided prespecified covariates with at least some rationale. Of the 29 trials, which gave specific reasons for the inclusion of all or some of the covariates, 22 provided a priori hypotheses while 7 provided statistical reasons. Of the seven trials that provided statistical reasons, statistical model selection was reported in four trials while guided exploration was reported in five (multiple reasons could be given within the same trial).

Objective Evidence for HTE (Clinicostatistical Divergence)

Twenty-nine of the 87 trials (33 percent) provided insufficient data (clinical divergence and statistical significance) for us to evaluate the strength of evidence for CSD. Of the 58 remaining trials, 6 (10 percent) reported strong evidence for CSD, while 11 (19 percent) reported moderate evidence, 25 (43 percent) reported weak evidence, and 16 (28 percent) reported no evidence for CSD (Table 2).

Table 2: Comparison of Author's Interpretations of Their Heterogeneity of Treatment Effects (HTE) Analysis and the Strength of Evidence for Clinicostatistical Divergence (CSD)

	Strength of Evidence for Clinicostatistical Divergence					
Authors' Interpretation of Their HTE Analysis Results	*Unable to Classify*	*None*	*Weak*	*Moderate*	*Strong*	*Total*
Evidence for HTE was sufficient to support different treatment recommendations in one or more subgroups	0	2	4	1	3	10
Evidence for HTE was insufficient to support differential treatment recommendations but sufficient to warrant further systematic research	10	6	10	5	0	31
Evidence for HTE was present but not enough to warrant differential treatment recommendations or further systematic research	0	0	0	0	1	1
Evidence for HTE was absent	6	2	8	2	2	20
No interpretation of HTE-related results	13	6	3	3	0	25
Total	29	16	25	11	6	87

Note. Numbers represent number of trials.

Authors' Interpretations and Comparison with Objective Evidence

Twenty-five of the 87 trials (29 percent) provided neither an interpretation of HTE results nor recommendations based on the HTE analyses. Of the remaining 62 trials, 10 (16 percent) indicated that evidence for HTE was sufficient to support different treatment recommendations in one or more subgroups; 31 (50 percent) that evidence for HTE was insufficient to support differential treatment recommendations but sufficient to warrant further systematic research; 1 (2 percent) that there was some evidence for HTE but not enough to warrant differential treatment recommendations or further systematic research; and 20 (32 percent) that evidence for HTE was absent.

When comparing the authors' interpretations of their own HTE-related results to the reported evidence for CSD, we found that of the 70 trials where evidence for CSD was classified as none or weak or insufficient for coding, 6 (9 percent) overstated the HTE results by suggesting a role for different treatment recommendations in one or more subgroups. Of the 45 trials where CSD was classified as none or insufficient for coding, 16 (36 percent) claimed that evidence for HTE was insufficient to support differential treatment recom-

mendations but sufficient to warrant further systematic research. In contrast, the results of 8 (47 percent) of the 17 trials where evidence for HTE was classified as moderate or strong were understated in authors' interpretations. The eight trials' interpretations were classified as showing some evidence for HTE but not enough to warrant differential treatment recommendations or further systematic research, that evidence for HTE was absent, or not providing any interpretation of the results despite reporting strong or moderate evidence for CSD. Therefore, in 30 of 87 trial reports (34 percent) authors potentially misinterpreted their HTE-related results, with 25 percent overstating (22/87) and 9 percent understating (8/87) the strength of their HTE-related results.

Trends over Time

Of the 87 trials, which reported using formal tests for interaction or heterogeneity, 22 (25 percent), 25 (29 percent), and 40 (46 percent) were reported in 1994, 1999, and 2004, respectively. Prespecified covariates were found significantly less frequently in trials published before 2004 as opposed to trials published in 2004 (pre-2004, 21/47 trials [45 percent] versus 2004, 30/40 [80 percent]; $p = .004$). Studies which did not provide sufficient data for classification of CSD were reported significantly more frequently in trials published before 2004 than those published in 2004 (pre-2004, 22/47 [47 percent] versus 2004, 7/40 [18 percent]; $p = .004$). Authors' possible misinterpretation of their HTE-related results occurred with equal frequency when comparing pre-2004 trial reports to those from 2004 (pre-2004, 14/47 [30 percent] versus 2004, 16/40 [40 percent]; $p = .32$).

Sensitivity Analyses

Changing the definition for statistical significance to include only trials with a p-value for interaction or heterogeneity of $\leq .05$ did not affect the number of trials classified as overstating HTE-related results but did decrease the number of trials classified as understating those results (8/87 trials [9 percent] when using a p-value of $\leq .10$ versus 4/87 trials [5 percent] when using a p-value of $\leq .05$). All four of these reclassified trials were found to also have clinically divergent subgroup effects. Furthermore, there were no trials found with p-values from .051 to .10 whose interpretation called for differential treatment. Changing the definition of clinical divergence by using NNT as the standard measure of effect size did not lead to a notable difference in the proportion of trials potentially misinterpreted (30/87 trials [34 percent] when using the

original clinical divergence definition versus 28/87 trials [32 percent] when using NNT), overstated (22/87 trials [25 percent] versus 22/87 trials [25 percent]), or understated (8/87 trials [9 percent] versus 6/87 trials [7 percent]). Combining the effect of both sensitivity analyses yielded a small overall decrease in the proportion of trials with potentially misinterpreted HTE-related results (26/87 trials [30 percent] overall misinterpreted; 23/87 trials [26 percent] overstated, and 3/87 trials [4 percent] understated]).

DISCUSSION

Identifying and interpreting evidence for HTE is becoming more salient, for clinicians as the populations they serve become more diverse, and for researchers as they try to maximize the external validity of their trials (Greenfield et al. 2007; McGuire et al. 2008) and conform to regulatory and funding agency mandates to increase the diversity of trial populations (Baird 1999). The utility of HTE analysis depends on the appropriate selection of covariates used to define subgroups (through prespecification and application of an explicit rationale); appropriate statistical analysis (through use of formal tests for interaction or heterogeneity); and balanced interpretation of the evidence (Bulpitt 1988; Yusuf et al. 1991; Altman et al. 2001; Lu et al. 2005; Rothwell 2005, 2007; Lagakos 2006; Wang et al. 2007). Previous studies have enumerated the problems with statistical analysis (Assmann et al. 2000; Parker and Naylor 2000; Moreira, Stein, and Susser 2001; Pocock et al. 2002; Hernandez et al. 2006; Wang et al. 2007; Gabler et al. 2009). In this review of a sample of RCTs published in leading medical journals over a 10-year period, we identified deficiencies in both the proximal (prespecification/rationale) and distal (interpretation) portions of this important pathway. There is some suggestion, however, that these deficiencies are improving over time.

In general, covariates examined as effect moderators in HTE analysis may be prespecified, supported by a sound rationale, both, or neither. In the ideal analysis, covariates are prespecified (thus limiting the problem of multiple testing) and accompanied by a sound rationale (thereby increasing the prior probability that the covariate plays an important moderating role and reducing the chance of false-positive associations). In the data presented here, about two-fifths of trials performing HTE analysis failed to prespecify any covariates; among trials that successfully prespecified at least some covariates, most did not consistently provide an explicit rationale. While there is a role for exploratory HTE analysis in which analyses are generated in an ad hoc or post

hoc fashion (Gheorghiade et al. 1991; Yusuf et al. 1991; Kraemer et al. 2002), authors need to clarify the status of such analyses so that readers may properly interpret them. A significant interaction term ($p < .05$ or .10) means one thing if the covariate was prespecified based on theory and prior data, and quite another if discovered adventitiously. Currently, readers must accept the authors' assertions of prespecification at face value. RCT registries offer the possibility of direct verification of a priori design of HTE investigation in the near future (Krleza-Jeric et al. 2005; De Angelis et al. 2005; Sim 2008).

In the absence of a standardized framework for interpreting the results of HTE analyses, readers depend upon authors to place the findings in context. However, our data suggest that authors may frequently misinterpret their results. The consequences of *overstating* HTE-related findings are obvious: clinicians may inappropriately reject good treatments or accept bad ones for specific patient subsets, and researchers may initiate research programs based on exaggerated evidence. The consequences of *understating* HTE-related findings are more subtle but arguably just as pernicious: clinicians may accept average effects as broadly applicable when they are not, and researchers may not be motivated to pursue appropriate follow-up studies to obtain more precise estimates of HTE in specific clinical situations.

Our study had four main limitations. First, we reviewed a limited number of journals, years, and trials. However, our sample comes from a group of journals that disproportionately affect treatment recommendations and future research (Garfield 1986; Chew, Villaneuva, and Van Der Weyden 2007). Second, the appropriateness of the significance level of the criteria, which we used to construct the measure of CSD, could be debated for different disease entities or treatments. We chose levels of statistical significance and clinical divergence that we felt were most appropriate for a review of general medical topics. Our sensitivity analyses varying definitions of statistical significance and clinical divergence showed similar results. Third, classification of authors' interpretations of their HTE-related results required some investigator judgment. However, all interpretations were coded individually by two investigators with no disagreement noted. Fourth, it is possible that our analysis was biased by inclusion of trials in which small sample size would make evaluation of HTE inappropriate. However, in a companion report based on the same set of trials (Gabler et al. 2009), the authors delineated the percentages of trials in which HTE analyses, subgroup-only analyses, or no subgroup analyses were performed. Limiting the examination to those trials in which sample size for investigating HTE would be most appropriate (> 250 participants with at least 100 participants in each treatment arm), Gabler and colleagues reported that

low rates of HTE analysis and high rates of subgroup-only analysis persisted when compared with the overall sample of trials.

Given the lack of a generally accepted algorithm for the evaluation of HTE-related evidence and the nontrivial misinterpretation rate reported here, we have developed a set of recommendations based on our coding criteria for CSD that will aid readers in making their own evaluations of the evidence for HTE. The recommendations provide readers with an approach for identifying the objective evidence for HTE (i.e., presence of tests for HTE and reporting of specialty-specific, clinically relevant absolute differences in outcome measures), assessing the likelihood of false-positive or false-negative results (i.e., prespecification and number of HTE analyses performed), and evaluating authors' interpretations of subgroup differences as evidence for or against HTE given the objective data. We have also provided recommendations for authors and journal editors to improve the design and publication of analyses for identifying HTE based on our findings and on previous published recommendations (Yusuf et al. 1991; Altman et al. 2001; Carneiro 2002; Cui et al. 2002; Simes, Gebski, and Keech 2004; Lu et al. 2005; Rothwell 2005, 2007; Wang et al. 2007). These recommendations will provide readers with a more systematic framework for identifying likely HTE (Box 1).

In conclusion, there is significant opportunity for improvement of the design, reporting, and interpretation of analyses used for the identification of HTE. The need for such improvement is especially urgent as Congress enacted large increases in funding for comparative effectiveness analysis (Conway and Clancy 2009; Federal Coordinating Council for Comparative Effectiveness Research 2009; Garber and Tunis 2009; Institute of Medicine, Committee on Comparative Effectiveness Research Prioritization, Board on Healthcare Services 2009). While potentially treacherous, the enterprise of HTE analysis for identifying HTE can be fruitful. If recommendations for the design and reporting of HTE analyses are utilized and authors' interpretations are cautiously evaluated in light of the objective evidence for CSD, the likelihood of under- or overstated HTE inappropriately influencing treatment or future research is decreased. The existence of HTE within clinical trials and the implications of HTE for real patient groups should be an important motivator for improving the process by which it is accurately identified. While possible improvements in reporting and interpretation of HTE analyses over time provide cause for cautious optimism, continued attention to this area is needed if the fruits of clinical research are to be properly harvested.

Box 1: Recommendations for the Interpretation (for Users) and Design and Reporting (for Producers) of Heterogeneity of Treatment Effects (HTE) Analysis

For Users

1. Start with examination of the data.
   A. Are the HTE analyses prespecified?
   B. Are the reasons for the inclusion of given covariates presented as a priori rationales with plausible reasons to anticipate HTE? If not, are they labeled as exploratory? This information will help place any positive results in context.
   C. How many HTE analyses were performed? This information will place a positive finding for heterogeneity into context of possible false-positive results.
   D. Was a formal test for heterogeneity or interaction used? At what level of significance given your discipline and the subject matter would you feel comfortable saying that a test indicates plausible interaction or heterogeneity (how important is missed opportunity of identifying HTE)?
   E. What is the difference between subgroup and the main average treatment effect? Given the subject matter, would you feel comfortable stating that a given difference is clinically meaningful?
2. Using a reasonable scheme for coding CSD, determine the level of strength of evidence for the presence of HTE.
3. Consider the authors' interpretation of their HTE-related results in light of your own determination of CSD and base your future research or treatment recommendations on this comparison.

For Producers and Disseminators

Researchers

1. Plan all HTE analyses with a priori rationales for the inclusion of given covariates, or clearly label the analyses as exploratory.
2. Report the total number of HTE analyses performed.
3. Use formal tests for interaction or heterogeneity when inferring HTE rather than comparing *p*-values between the two groups. Clearly state the level of significance used in the tests for HTE.
4. Present data from all HTE analyses performed, including *p*-values, effect measures, and confidence intervals. Forrest plots provide a concise method for doing so.
5. Interpret the HTE analyses performed in the discussion of the paper. Discuss the role of multiplicity when discussing positive results. Stress the exploratory nature of results if the analyses were not prespecified and/or power was insufficient. Refrain from recommending differential treatment unless confident that your HTE analyses were hypothesis testing rather than hypothesis generating and that objective evidence is strong enough to support your recommendations.

Editors

1. Ensure that authors reporting any HTE analysis provide readers with enough information to place the results into context. This includes rationales for the HTE analysis; whether the analyses were primarily exploratory; the number of HTE analyses performed; whether tests for interaction or heterogeneity were used; level of significance used for the tests; and all the results from HTE analyses (Internet-only appendices and/or forest plots recommended)
2. Ensure that authors fully discuss any HTE analyses reported, even if not significant.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This research was supported in part by a grant from Pfizer Inc. (all authors); by a Primary Care Faculty Development grant (D55 HP00232) from the Health Resources and Services Administration of the U.S. Department of Health and Human Services (Dr. Fernandez y Garcia); and by a research infrastructure grant (UL1 RR024146) from the National Center for Research Resources (Drs. Nguyen and Kravitz).

The authors would like to acknowledge Diana Liao at the University of California, Los Angeles for her assistance in the early stages of statistical analysis and programming.

Disclosure: Pfizer Inc. had a right to review and comment on the manuscript 30 days before original submission. There are no other disclosures.

Disclaimer: The funding sources had no role in the collection of the data, analysis, interpretation, or reporting of the data or in the decision to submit the manuscript for publication.

REFERENCES

Altman, D. G., and P. K. Andersen. 1999. "Calculating the Number Needed to Treat for Trials Where the Outcome Is Time to an Event." *British Medical Journal* 319: 1492–5.

Altman, D. G., K. F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. C. Getzche, and T. Lang. 2001. "The Revised CONSORT Statement for Reporting Trials: Explanation and Elaboration." *Annals of Internal Medicine* 134 (8): 663–94.

Assmann, S. F., S. J. Pocock, L. E. Enos, and L. E. Kasten. 2000. "Subgroup Analysis and Other (Mis)uses of Baseline Data in Clinical Trials." *Lancet* 355: 1064–9.

Baird, K. L. 1999. "The New NIH and FDA Medical Research Policies: Targeting Gender, Promoting Justice." *Journal of Health Politics, Policy, and Law* 24 (3): 531–65.

Brookes, S. T., E. Whitely, M. Egger, G. D. Smith, P. A. Mulheran, and T. J. Peters. 2004. "Subgroup Analyses in Randomized Trials: Risks of Subgroup-Specific Analyses; Power and Sample Size for the Interaction Test." *Journal of Clinical Epidemiology* 57: 229–36.

Brookes, S. T., E. Whitely, T. J. Peters, P. A. Mulheran, M. Egger, and G. D. Smith. 2001. "Subgroup Analyses in Randomised Controlled Trials: Quantifying the Risks of False-Positives and False-Negatives." *Health Technology Assessment* 5 (33): 1–56.

Bulpitt, C. J. 1988. "Subgroup Analysis." *Lancet* 2: 31–4.

Carneiro, A. V. 2002. "Subgroup Analysis in Therapeutic Trials." *Revista Portuguesa de Cardiologia* 21 (3): 339–46.

Chew, M., E. V. Villaneuva, and M. B. Van Der Weyden. 2007. "Life and Times of the Impact Factor: Retrospective Analysis of Trends for Seven Medical Journals (1994–2005) and Their Editors' Views." *Journal of the Royal Society of Medicine* 100 (3): 142–50.

Conway, P. H., and C. Clancy. 2009. "Comparative Effectiveness Research-Implications of the Federal Council's Report." *New England Journal of Medicine* 361 (4): 328–30.

Cook, D. I., V. J. Gebski, and A. C. Keech. 2004. "Subgroup Analysis in Clinical Trials." *Medical Journal of Australia* 180 (6): 289–91.

Cook, R. J., and D. L. Sackett. 1995. "The Number Needed to Treat: A Clinically Useful Measure of Treatment Effect." *British Medical Journal* 310: 452–4.

Cui, L., H. M. J. Hung, S. J. Wang, and Y. Tsong. 2002. "Issues Related to Subgroup Analysis in Clinical Trials." *Journal of Biopharmaceutical Statistics* 12 (3): 347–58.

De Angelis, C. D., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. M. Overbeke, T. V. Schroeder, H. C. Sox, and M. B. Van Der Weyden. 2005. "Is This Trial Fully Registered? A Statement from the International Committee of Medical Journal Editors." *Journal of the American Medical Association* 293 (23): 2927–9.

Drakulovic, M. B., A. Torres, T. T. Bauer, J. M. Nicolas, S. Nogue, and M. Ferrer. 1999. "Supine Body Position as a Risk Factor for Nosocomial Pneumonia in Mechanically Ventilated Patients: A Randomised Trial." *Lancet* 354: 1851–8.

Federal Coordinating Council for Comparative Effectiveness Research. 2009. *Report to the President and Congress* [accessed on August 15, 2009]. Washington, DC: Department of Health and Human Services. Available at http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf

Gabler, N. B., N. Duan, D. Liao, J. G. Elmore, T. Ganiats, and R. L. Kravitz. 2009. "Dealing with Heterogeneity of Treatment Effects: Is the Literature Up to the Challenge?" *Trials* 10: 43, doi: 10.1186/1745-6215-10-43.

Garber, A. M., and S. R. Tunis. 2009. "Does Comparative Effectiveness Research Threaten Personalized Medicine?" *New England Journal of Medicine* 360 (19): 1925–7.

Garfield, E. 1986. "Which Medical Journals Have the Greatest Impact?" *Annals of Internal Medicine* 105 (2): 313–20.

Gheorghiade, M., L. Schultz, B. Tilley, W. Kao, and S. Goldstein. 1991. "Subgroup Analysis of Clinical Trials." *American Journal of Cardiology* 67: 330–1.

Greenfield, S., R. Kravitz, N. Duan, and S. H. Kaplan. 2007. "Heterogeneity of Treatment Effects: Implications for Guidelines, Payment, and Quality Assessment." *American Journal of Medicine* 120 (4A): S3–9.

Hernandez, A. V., E. Boersma, G. D. Murray, J. D. F. Habbema, and E. W. Steyerberg. 2006. "Subgroup Analyses in Therapeutic Cardiovascular Clinical Trials: Are Most of Them Misleading?" *American Heart Journal* 151 (2): 257–64.

Hernandez, A. V., E. W. Steyerberg, G. S. Taylor, A. Marmarou, J. D. F. Habbema, and A. I. R. Maas. 2005. "Subgroup Analysis and Covariate Adjustment in Randomized Clinical Trials of Traumatic Brain Injury: A Systemic Review." *Neurosurgery* 57: 1244–53.

Institute of Medicine, Committee on Comparative Effectiveness Research Prioritization, Board on Healthcare Services. 2009. *Initial National Priorities for Comparative Effectiveness Research* [accessed on August 15, 2009]. Washington, DC: National Academies Press. Available at http://www.nap.edu/catalog.php?record_id=12648#toc

Kraemer, H. C., E. Frank, and D. J. Kupfer. 2006. "Moderators of Treatment Outcomes. Clinical, Research, and Policy Importance." *Journal of the American Medical Association* 296 (10): 1286–9.

Kraemer, H. C., and D. J. Kupfer. 2006. "Size of Treatment Effects and Their Importance to Clinical Research and Practice." *Biological Psychiatry* 59: 990–6.

Kraemer, H. C., G. T. Wilson, C. G. Fairburn, and W. S. Agras. 2002. "Mediators and Moderators of Treatment Effects in Randomized Controlled Trials." *Archives of General Psychiatry* 59: 877–83.

Kravitz, R. L., N. Duan, and J. Braslow. 2004. "Evidenced-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages." *Millbank Quarterly* 82 (4): 661–87.

Krleza-Jeric, K., A. W. Chan, K. Dickersin, I. Sim, J. Grimshaw, and C. Gluud. 2005. "Principles for International Registration of Protocol Information and Results from Human Trials of Health Related Interventions: Ottawa Statement (part 1)." *British Medical Journal* 330 (7497): 956–8.

Lagakos, S. W. 2006. "The Challenge of Subgroup Analysis-Reporting without Distorting." *New England Journal of Medicine* 354 (16): 1667–70.

Longford, N. T. 1999. "Selection Bias and Treatment Heterogeneity in Clinical Trials." *Statistics in Medicine* 18: 1467–74.

Lu, M., P. D. Lyden, T. G. Brott, S. Hamilton, J. P. Broderick, and J. C. Grotta. 2005. "Beyond Subgroup Analysis: Improving the Clinical Interpretation of Treatment Effects in Stroke Research." *Journal of Neuroscience Methods* 143: 209–16.

McGuire, T. G., J. Z. Ayanian, D. E. Ford, R. E. M. Henke, K. M. Rost, and A. M. Zaslasky. 2008. "Testing for Statistical Discrimination by Race/Ethnicity in Panel Data for Depression Treatment in Primary Care." *Health Services Research* 43 (2): 531–51.

Moreira, E. D., Z. Stein, and E. Susser. 2001. "Reporting on Methods of Subgroup Analysis in Clinical Trials: A Survey of Four Scientific Journals." *Brazilian Journal of Medical and Biological Research* 34: 1441–6.

Parker, A. B., and C. D. Naylor. 2000. "Subgroups, Treatment Effects, and Baseline Risks: Some Lessons from Major Cardiovascular Trials." *American Heart Journal* 139 (6): 952–61.

———. 2006. "Interpretation of Subgroup Results in Clinical Trial Publications: Insights from a Survey of Medical Specialists in Ontario, Canada." *American Heart Journal* 151 (3): 580–8.

Pocock, S. J., S. F. Assmann, L. E. Enos, and L. E. Kasten. 2002. "Subgroup Analysis, Covariate Adjustment, and Baseline Comparisons in Clinical Trial Reporting: Current Practices and Problems." *Statistics in Medicine* 21: 2917–30.

Rothwell, P. M. 2005. "Treating Individuals 2. Subgroup Analysis in Randomised Controlled Trials: Importance, Indications, and Interpretations." *Lancet* 365: 176–86.

———. 2007. "Reliable Estimation and Interpretation of the Effects of Treatment in Subgroups." In *Treating Individuals: From Randomized Trials to Personalized Medicine, The Lancet*, pp. 169–82. Spain: Elsevier Limited.

Sim, I. 2008. "Trial Registration for Public Trust: Making the Case for Medical Devices." *Journal of General Internal Medicine* 23 (Suppl 1): 64–8.

Simes, J. R., V. J. Gebski, and A. C. Keech. 2004. "Subgroup Analysis: Application to Individual Patient Decisions." *Medical Journal of Australia* 180: 467–9.

Stallones, R. A. 1987. "The Use and Abuse of Subgroup Analysis in Epidemiological Research." *Preventative Medicine* 16: 183–94.

Wang, R., S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen. 2007. "Statistics in Medicine-Reporting of Subgroup Analyses in Clinical Trials." *New England Journal of Medicine* 357 (21): 2189–94.

Yusuf, S., J. Wittles, J. Probstfield, and H. A. Tyroler. 1991. "Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials." *Journal of the American Medical Association* 266 (1): 93–8.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.