Review of Guidance Documents for Selected Methods in Patient Centered Outcomes Research:
Standards in Addressing Heterogeneity of Treatment Effectiveness in Observational and Experimental Patient Centered Outcomes Research

A Report to the
PCORI Methodology Committee Research Methods
Working Group

Ravi Varadhan, PhD

Elizabeth A. Stuart, PhD

Thomas A. Louis, PhD

Jodi B. Segal, MD, MPH

Carlos O. Weiss, MD, MHS

March 29, 2012

Introduction:

Individuals vary in their response to a treatment: some derive substantial overall benefit; some derive little benefit, while others are harmed. Understanding this heterogeneity of treatment effect (HTE) is critical for evaluating how well a treatment can be expected to work for an individual or a group of similar individuals. Assessment of HTE is essential in patient-centered outcomes research (PCOR), which aims to help stakeholders make informed personalized health care decisions. HTE may be defined as the non-random, explainable variability in the direction and magnitude of treatment effect. Understanding HTE is critical for decisions that are based on knowing how well a treatment is likely to work for an individual, or group of similar individuals; and it is relevant to most stakeholders, including patients, clinicians, and policy makers.

There are two main goals of HTE analyses: (1) to estimate treatment effects in clinically relevant subgroups (subgroup analysis ([SGA]), and (2) to predict whether an individual might benefit from a treatment (predictive learning). For the first goal, subgroup analysis (SGA) is the most common analytic approach for examining HTE. The second goal of HTE analysis is to predict the probabilities of beneficial and adverse responses of individuals to a treatment. We only propose minimum standards for SGA since little guidance is available for the prediction of individual responses to treatments.

i. Background on Testing for HTE

While the motivation for examining HTE is all too obvious, reliable identification of HTE is far from trivial. Use of SGA is susceptible to well-known problems that can result in increased likelihood of falsely detecting HTE (Type-I error) or failing to detect true HTE (Type-II error), when compared to examining whether the treatment works, on average, in the overall sample. Prior planning, careful analysis, and responsible reporting are critical when examining HTE in order that the consumers of are not misled, and can benefit from this information. Good methodological practices are therefore clearly valuable.

There are two types of SGA: estimating treatment effect separately within levels of baseline covariates (e.g., men and women), and modeling the interaction between the treatment and covariates. It is important to distinguish between these two types of SGA. Simple notation follows, to illustrate this distinction clearly. Consider a single subgrouping variable, X, that is binary (X = 0, 1). Let the treatment be represented by a binary indicator variable, A (A=0 for treatment, A=1 for no treatment). Finally, let Y be the outcome, which can be continuous, binary, or time-to-event (subject to censoring). For a given individual, the observed data is $\{X_i, A_i, Y_i\}$. The treatment effect, the effect of A on Y, is a comparison of some function of the average (expectation) of Y when A=0 to when A=1. In other words, the treatment $\theta = f(g(E[Y|A=0], g(E[Y|A=1))$. For example, let us suppose that Y is binary. Let g(.) be the logistic function, and let f(.) be the difference function. Now, the treatment effect θ is the log-odds ratio of Y=1 when A=1 versus when A=0.

A common practice in SGA is to estimate the treatment effect separately within the strata of covariate X, (i.e., the treatment effect is estimated separately in X=1 and X=2). Let these two estimated effects be θ_1 and θ_2. Then HTE is inferred when θ_1 is statistically significant and θ_2 is not, or vice-versa. While this approach is correct for estimating stratified treatment effects, it is incorrect for inferring HTE. This stratified approach is cumbersome to implement when X has

more than two levels. Brookes[1] showed that when using this flawed approach, even when there is no actual HTE, there is up to 66 percent probability of inferring the presence of HTE when the overall treatment effect is significant; and that there is up to 25 percent probability of inferring HTE when the overall treatment effect is non-significant. To correctly test for HTE, we need to test whether the difference between the two stratified treatment effects is zero using a Wald test.

A better way to assess HTE is to model the interaction between A and X. In the interaction modeling approach, fit the regression model:

$$g(E[Y \mid X, A]) = b_0 + b_A A + b_X X + b_{A.X} A * X \qquad (1)$$

In model (1), the term $b_{A.X}$ denotes the interaction between treatment A and the binary covariate X. Since model (1) is a saturated model, the stratified effects θ_1 and θ_2 and the coefficients of the interaction model (1) are equivalent: $b_A = \theta_1$ and $b_A + b_{A.X} = \theta_2$. Therefore, the interaction term $b_{A.X} = \theta_2 - \theta_1$. Thus, the interaction term actually is the same as the difference between the two stratified treatment effects. We can test for significant interaction using a Wald test, where we simply compute the P-value for the t-statistic of the interaction term $b_{A.X}$. Or alternatively, we can do a likelihood ratio test that compares model (1) to a simpler model without the interaction term. These two tests should yield very similar results in large subgroups. The interaction test is simpler to implement than a test that evaluates the significance of the difference $\theta_2 - \theta_1$. The interaction test (especially, the likelihood ratio test) is also easier and the best approach for detecting HTE due to covariate X, when X has more than two levels. The interaction test always maintains the correct 5 percent probability regardless of whether the overall treatment effect was significant or not. Obviously, one can also use the interaction testing using model (1) when X is continuous, whereas the stratified approach is not applicable without using cut-points to categorize X.

ii. A Framework for HTE Analysis in PCOR

There are numerous commentaries that discuss SGA. Even though much has been written about the perils of SGA and sound principles for conducting them, examples abound of inappropriate conduct, reporting, interpretation, and application of subgroup analytic results.[2,3] One important reason is the myopic view of SGA as a hypothesis testing problem rather than as an estimation problem. This focus on hypothesis testing has unfortunately resulted in the dichotomization of SGA into confirmatory (hypothesis-testing) and exploratory (hypothesis-generating) analyses. We have proposed an expanded analytic framework that allows both hypothesis testing and HTE estimation. In particular, we have introduced a new type of HTE analysis called descriptive HTE analysis. In descriptive HTE analysis, the focus is on estimation of treatment effects in prespecified subgroups. Each study reports these effect estimates and their standard errors in order to facilitate future meta-analysis.

Which types of subgrouping variables should studies use for descriptive subgroup analysis? Some potentially important classes of variables are: demographics (e.g., age and sex), behavior (e.g., smoking), pathophysiology (e.g., measures of disease severity), genetic markers, and comorbid conditions (e.g., diabetes status in cardiovascular disease trials). Studies might use descriptive HTE for sub-populations for which limited evidence is available in the literature, such as the priority populations specified by the Agency for Healthcare Research Quality, including women, children, minorities, elderly, individuals with disabilities, and rural populations.[4] We suggest that descriptive HTE should always include sex and age. A few journals actually recommend reporting results for important subgroups. For example, the Journal of the National

Cancer Institute[5] instructs authors that "where appropriate, clinical and epidemiologic studies should be analyzed to see if there is an effect of sex or any of the major ethnic groups. If there is no effect, it should be so stated in Results." Recently, the Food and Drug Administration (FDA) issued draft guidance[6] for industry trial sponsors of medical device clinical studies that recommended sex-specific HTE analyses. Notably, the draft guidance says that even in the absence of a significant sex-by-treatment interaction, it is preferable to report results by sex. A workshop sponsored by the Institute of Medicine[7] also discussed reporting of sex-specific results from studies. Some of the journal editors at the workshop cautioned that the reporting of sex-specific results can be misleading and subject to misinterpretation due to the potential for type II errors, (i.e., publishing comparisons by sex that do not show a significant difference). They argued that if a study was not adequately powered to look for such differences, then showing no differences was meaningless. However, it was also suggested that when there is insufficient power to analyze sex differences within a study, it may be possible to combine data from various studies and conduct meta-analysis or apply advanced statistical methods, such as the use of Bayesian inference. This is a critically important insight that captures the essence of descriptive analysis.

Table 1 depicts the essential characteristics of the three types of HTE analyses, confirmatory, exploratory, and descriptive. Table 1 shows how descriptive HTE analysis differs from the conventional types of confirmatory and exploratory HTE analyses. While descriptive HTE is pre-specified, unlike confirmatory HTE it does not rely on strong prior evidence. Descriptive analyses need not be powered to test for heterogeneity within a study, because the focus is on estimation of subgroup-specific estimates for future synthesis. Consequently, researchers should not over interpret the results from a descriptive HTE analysis from an individual study. In particular, researchers should avoid reporting the P-values. Descriptive HTE analysis differs from exploratory HTE analysis mainly in that it is pre-specified and its sampling properties (e.g. standard error) can be characterized, making future synthesis feasible. In contrast, the results from exploratory HTE are not appropriate for meta-analysis because their sampling properties cannot be characterized.

We propose minimum standards according to this framework for inferential goals of HTE analysis in PCOR.

Methods:

To perform a literature review, we worked with an information specialist with experience conducting systematic reviews (Claire Twose, Johns Hopkins Welch Medical Library). The search sought several types of guidance documents: official guidance (e.g., from the Institute of Medicine), expert groups (e.g., Cochrane Collaboration), and the primary methodology literature. Because of the paucity of official and expert group guidance available for the topic of HTE, the search also sought expert, peer-reviewed published recommendations accompanied by scientific rationale.

i. Search Strategy for Identifying Relevant Reports

Overall, the search strategy included database searches and other searches.

Database searches included the National Library of Medicine Books, National Library of Medicine Catalog, and Current Index to Statistics databases. We considered searching the MEDLINE database. However, in a previous research project, "Methods to Study the Heterogeneity of Treatment Effects in Comparative Effectiveness Research", funded by the Agency for Healthcare Research and Quality in 2010-2011, we learned that a MEDLINE search strategy was neither sensitive nor specific for a methodological review. Similarly, for this project, we confirmed that a search for "heterogeneity of treatment effect (HTE)" in MEDLINE resulted in "Quoted phrase not found." We also confirmed that the National Library of Medicine has not introduced any Medical Subject Heading descriptor or entry term that relate to HTE. We performed a crosscheck with MEDLINE after the main search strategy, which we described below. The database searches are summarized in Table 2.

Other searches included a review of documents from a 2010 literature review of primary methodological studies addressing HTE. In that review of primary methods articles, we had conducted a structured literature review followed by a concept mapping exercise. We had reviewed 54 articles culled from the investigators' article libraries and had identified 81 different key words related to HTE that mapped onto seven topics: data designs (e.g., trials), sampling designs (e.g., selection bias), detection of effect modification (e.g., treatment effect moderators or a treatment-covariate interaction), major types of treatment effect modification (e.g., qualitative heterogeneity and harm), known dimensions of HTE (e.g., competing risks), broad approaches to HTE (e.g., decision theory formulations), and analytic methods (e.g., structural micro-simulation). We used citation search strategy to find sources in a variety of disciplines. We had used ISI Web of Science[8] to retrieve all articles (n=1,702 unique articles) that cited seed methodological articles. We found that 8.6 percent of the articles were in statistical journals. We had retrieved all these for review of the full text. For remaining articles in non-statistical journals, two investigators had reviewed the citation information. After resolving disagreements through a tie-breaking process by the other investigators, we had included a total of 315 articles for full text review. To find guidance documents for this current project, we performed a manual review of studies identified during that previous project.

Other searches also included citation review. We maintained an open list of documents to undergo citation review. In other words, for any additional document identified as a potential guideline we performed an additional citation review.

Finally, other searches included website searches. We searched the websites of 25 major regulatory agencies and organizations: FDA; National Academy of Sciences; Institute of Medicine; Agency for Healthcare Research and Quality; National Research Council; National

Institutes of Health; Canadian Agency for Drugs and Technologies in Health; German Institute for Quality and Efficiency in Health Care; European Medicines Agency; U.K. Medical Research Council; U.K. Medical Research Council; National Clinical Guideline Centre; National Institute for Health and Clinical Excellence; Institut National de la Santé et de la Recherche Médicale; Pharmaceuticals and Medical Devices Agency, Japan; Danish Centre for Health Technology Assessment; ICH Harmonization for Better Health; Australian Regulatory Agency; Therapeutic Goods Administration; National Health and Medical Research Council; International Association for Health European Medicines Agency; European Network for Health Technology Assessment; World Health Organization; World Bank; Cochrane Library; GRADE Working Group; American Statistical Association; and International Biometric Society. We searched the websites between February 17-27, 2012, with variations on the terms heterogeneity and guidelines or standards using a Google site: search and, when available, each site's own search interface.

As a final cross-check concerning the utility of a MEDLINE search, we conducted a query through the PubMed interface: (((heterogen* AND effect*) OR (effect* AND modif*) OR (interaction*) OR ((subgroup* OR subpopulation* OR subset*) AND analy*) OR ((subgroup* OR subpopulation*) AND effect*))) AND ("guidelines as topic"[mesh] OR guideline[all fields] OR "best practice"[all fields] OR "best practices"[all fields] OR "user guide"[all fields] OR "user guides"[all fields] OR "user's guide"[all fields]). This search had 4,405 results on 3/20/12. We reviewed the first 800 results. There was one hit and we already identified it through the main search strategy.

ii. Screening and Assessing for Eligibility of Retrieved Reports

Two investigators (CW, EL) independently screened search results through review of bibliographic information, including abstracts. We kept the document if either investigator classified it as possibly eligible. We then used the full text review of screened documents to assess eligibility. We used endnote software to manage citations.

iii. Inclusion and Exclusion Criteria

In order to be included in this project, the document had to be a guidance document. The working definition we used for a guidance document was: "it must provide guidance, not only discussion or a presentation of methods". We considered the following key words as likely indicators of a guidance document: advisory, best practice(s), criteria, guidance(s), guideline(s), standard(s) and statement(s). Because relatively few documents met the inclusion criterion of being a guidance document concerning HTE, there were no exclusion criteria per se. We included guidelines from organizations and groups outside of the U.S. We obtained draft guidelines when possible.

iv. Abstraction

Rather than perform abstraction, we extracted (i.e., copied verbatim) all recommendations from guidance documents that relate to HTE. This created a single document compiling recommendations (Appendix A).

v. Methods for Selection of Standards

Each author considered a minimum standard to mean that there were good methodological justifications (i.e., sound scientific and statistical principles) and there was at least one example of application of the standard in clinical research practice. Each main investigator (RV, CW)

drafted a standards table. At this point we considered several organizational schemes for the standards (e.g., according to stage of investigation: design, analysis, reporting, interpreting; according to study design: trial, observational study, meta-analysis; or, inferential goal: confirmatory, exploratory or descriptive). The main investigators discussed and compared the first draft standards tables to ensure consistency of intent and format. Then each main author independently drafted remaining standards.

vi. Synthesis, i.e. Approach to Selection of Minimum Standards

The main investigators discussed the independently-created minimum standards then merged or deleted standards through consensus-building to create a final list of draft standards. The goal was to create standards that would make unambiguous the fundamental challenges to studying HTE. The main investigators showed the final list of minimum standards to co-investigators for comment. An expert in medical writing (EV) reviewed the wording of the standards.

Results:

i. Search Results with Flow Chart

The flow of the document search and results at each step are presented in Figure 1.

ii. Main Findings: Minimum Standards for the study of HTE in PCOR

The main investigators independently created 38 standards (RV 21; CW 17). Through discussion and consensus-building we chose inferential goals of HTE analysis as the main organizational scheme. Then we merged or deleted the independently-written standards to reduce the final number of standards to 14.

Box 1 lists the proposed minimum standards.

Box 1. Proposed Minimum Standards for HTE Analysis in PCOR.

State the Goals of HTE Analyses
For Confirmatory HTE Analyses, Prespecify a Few Subgroup Hypotheses
For Confirmatory HTE Analyses, Involve Stakeholders in the Selection of Subgroups and Outcomes
For Confirmatory HTE Analyses, Report *a priori* Statistical Power
For Confirmatory and Descriptive HTE Analyses, Report Sufficient Information on Treatment Effect Estimates
For Descriptive HTE Analyses, Prespecify Subgroups and Outcomes
For Descriptive HTE Analyses, Involve Stakeholders in the Selection of Subgroups and Outcomes
For Exploratory HTE Analysis, Document the Number of Subgroups and Outcomes Analyzed
For Exploratory HTE Analyses, Discuss Findings in the Context of Study Design and Prior Evidence
For Any HTE Analysis in Observational Data, Explicitly Assess Study Quality for Making Causal Inference
For Any HTE Analysis, Describe the Analytical Methods in Detail
For Any HTE Analysis, Perform an Interaction Test
For Any HTE Analysis, Report All Prespecified Analyses
For Any HTE Analysis, Use Appropriate Methods for Post-treatment Subgroups

We present each proposed standards as a stand-alone table (Tables 3-16). Each table discusses the rationale for choosing the standard, provides reference sources for the standard and reference sources for publications that have adhered to the standard. We followed the PCORI template, based on the CONSORT explanation and elaboration model. Appendix Table 1 provides a description of guidance documents included as justification for the proposed minimum standards.

iii. State of the Art Methods Guidance Not Included in the Main Findings

There were a few important issues in the assessment of HTE that guidance articles addressed with different emphases and varying levels of detail. However, there has been no consensus on how to resolve these issues. Therefore, we did not make any direct recommendations that directly addressed those issues.

The first issue is that of formally accounting for multiple testing in HTE analyses. This is a hotly-debated methodological issue. Our view is that a focus on the inferential goals of HTE analysis is most effective in addressing the multiple-testing problem. For example, the proposed

framework makes it clear that the P-values and attendant concerns regarding multiple testing are irrelevant in a descriptive HTE analysis, whose main purpose is to estimate and report effects for future meta-analysis. Furthermore, the proposed minimum standards are aimed at producing reliable inferences on HTE by ensuring prespecification and transparency, rather than prescribing statistical machinations for multiple testing. Our position is consistent with the view of Don Berry, "Neither Type I error adjustments, nor ignoring the problem of multiplicities, is consistent with the scientific method. Researchers and statisticians should take a reasoned approach that recognizes the possibility that observed differences may be the result of random variability as well as the possibility that they are real."[9]

A second issue is the recommendation that reporting in trials should focus on the overall treatment effect, and should avoid the examination of HTE when the overall treatment effect is null.[1,10-13] This is an example of a recommendation that may be justifiable in one context, but it is incongruous in another context. The main motivation behind this is to prevent sponsors from "salvaging" a trial that has shown no overall treatment effect. This recommendation is justifiable in a regulatory context for approval of drugs. However, it is incongruent with the goals of PCOR, which is very much interested in identifying subgroups exhibiting different responses to different treatments. Furthermore, this recommendation may be inappropriate in the context of comparative effectiveness research where two active treatments are compared because a null overall treatment effect can mask qualitative differences in treatment effect. In other words, treatment A is better than treatment B in one subgroup, but treatment B is better in a mutually exclusive subgroup. Such qualitative differences are more likely when two active treatment arms are being contrasted compared to when one arm is an inactive treatment. Finally, and perhaps most importantly, we were unable to find a theoretical or statistical justification for why HTE should not be examined when the overall treatment effect is non-significant. Therefore, we did not adopt this standard. For PCOR, the focus should be on reliably estimating subgroup effects, either by replicating exploratory HTE results or by conducting meta-analysis of HTE results based on descriptive HTE analyses.

Appendix Table 2 provides a description of guidance documents not included as justification for the proposed minimum standards. We did not create an Appendix Table 3 because none of the guidance documents were focused on PCOR.

Discussion of Major Challenges and Gaps:

The main challenge that we encountered in developing minimum standards for HTE analysis was that there are no formal guidance documents devoted to the analysis of HTE. There are numerous commentaries. These articles have put forth numerous principles for proper design, analysis, reporting and interpretation of subgroup analyses. Despite the plethora of articles, it is surprising that there are no formal guidance documents. Agencies such as the FDA and the European Medicines Agency do not have any formal guidance documents devoted to subgroup analyses (we understand that both agencies are actively working on preparing formal guidance).

We also identified some major gaps (i.e., situations that are likely to occur frequently in patient-centered research but for which it is difficult to propose justifiable standards due to a lack of adequate methodological understanding and empirical experience). The literature on HTE analysis is essentially all about subgroup analysis in placebo-controlled experimental studies. Few papers, if any, have been written from the perspective of patient-centered outcomes research. Notably, none of the articles address active comparator trials and pragmatic (simple, large) trials.

Another major gap is the lack of guidance for HTE analysis in observational designs, which we believe will be of paramount importance in PCOR. Experimental studies target design and data collection towards estimating a treatment's effects. This is often not the case in observational studies. Consequently, the type and quality of data these studies collect are not optimal for estimating treatment effects. For example, data on initiation, duration, and intensity of exposure are often unavailable. Observational studies are susceptible to major problems: ascertainment and selection biases in exposure to treatment, measurement error in assessment of health outcomes, and lack of information on important prognostic variables. These biases and measurement errors can introduce apparent HTE when in fact none is present, or conversely, obscure HTE when it is actually present.[14] David Byar wrote in 1980 that sound inferences on comparing treatments would not generally be possible from analyzing observational data because of difficulties with bias in treatment assignment, nonstandard definitions, definitions changing in time, specification of groups to be compared, missing data, and multiple comparisons.[15] While there have been significant advances since 1980 in the statistical methodology for making valid inference from observational data (e.g., propensity scores, instrumental variables, principal stratification), they cannot overcome fundamental design and data limitations without relying upon strong and, often, unjustifiable assumptions.

In spite of all these problems, the governing principles of prespecification and transparency are as important in observational studies as they are in trials. The scientific community must put procedures in place to ensure that the results from observational studies are trustworthy. These procedures include: registering observational studies prospectively; pre-specifying hypotheses; doing power calculations; and performing detailed pre-specification of analytic plans including how confounding, missing data, and how loss-to-follow up will be handled. Sox called for registration of observational studies, along the lines of National Institutes of Health's clinical trials registry.[16] Rubin put forth an interesting proposal for "objective causal inference," where a greater emphasis is placed on understanding treatment selection, and the modeler is blinded to outcomes until the treatment assignment modeling is complete.[17] These ideas are worthy of serious consideration.

Other areas of HTE analysis where little guidance is available are: Bayesian methods, prediction modeling for individual responses to treatments, and appropriate outcome scale for HTE analysis (e.g., risk difference, risk ratio, log of odds-ratio). We note that this is by no means an exhaustive list.

Conclusions:

Two major governing principles emerged from our study for developing the minimum standards for HTE analysis in PCOR. These are prespecification and transparency of reporting. At the design stage, prespecification of the subgroup hypotheses and the analytic plan for HTE analysis contributes directly to scientific rigor. At the reporting stage, transparency regarding all analyses enables an unbiased assessment of the significance of the results and fosters trustworthiness of the findings. We believe that these two considerations are fundamental to reliable assessment of HTE. Therefore, we based our minimum standards on these two principles.

References:

1. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technology Assessment.* 2001;5(33):1-56.

2. Fernandez YGE, Nguyen H, Duan N, Gabler NB, Kravitz RL. Assessing Heterogeneity of Treatment Effects: Are Authors Misinterpreting Their Results? *Health Services Research.* Nov 19 2009.

3. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. *New England Journal of Medicine.* Nov 22 2007;357(21):2189-2194.

4. Agency for Healthcare Research and Quality. Priority Populations. 2012; http://www.ahrq.gov/populations/. Accessed March 29, 2012.

5. Arnold K. Journal to encourage analysis by sex/ethnicity. *Journal of the National Cancer Institute.* Oct 4 2000;92(19):1561.

6. Food and Drug Administration. *Draft Guidance for Industry and Food and Drug Administration Staff: Evaluation of Sex Differences in Medical Device Clinical Studies.* Rockville, MD: U.S. Department of Health and Human Services; December 19, 2011 2011.

7. Institute of Medicine. *Sex-Specific Reporting of Scientific Research - Workshop Summary.* Washington, D.C.: National Academies Press;2012.

8. Thompson Reuters. Web of Science. 2010; http://thomsonreuters.com/content/PDF/scientific/Web_of_Science_factsheet.pdf. Available at: http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science. Accessed June 01, 2010.

9. Berry DA. Subgroup analyses. *Biometrics.* Dec 1990;46(4):1227-1230.

10. Food and Drug Administration. Guideline for the Format and Content of the Clinical and Statistical Sections of an Application. Rockville, MD. 1988.

11. Food and Drug Administration. E9 Statistical Principles for Clinical Trials Guidance for Industry. In: (CBER) URockville, MD. 1998.

12. Fayers PM, King MT. How to guarantee finding a statistically significant difference: the use and abuse of subgroup analyses. *Quality of life research.* Jun 2009;18(5):527-530.

13. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the Amerian Medical Association.* Jul 3 1991;266(1):93-98.

14. Vanderweele TJ, Mukherjee B, Chen J. Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine.* Oct 4 2011.

15. Byar DP. Why data bases should not replace randomized clinical trials. *Biometrics.* Jun 1980;36(2):337-342.

16. Sox HC, Helfand M, Grimshaw J, et al. Comparative effectiveness research: challenges for medical journals. *PLoS Med.* 2010;7(4):e1000269.

17. Rubin DB. For objective causal inference, design trumps analysis. *Annals of Applied Statistics.* 2008;2(3):808-840.

18. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet.* Jan 8-14 2005;365(9454):176-186.

19. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *American Heart Journal.* Feb 2006;151(2):257-264.

20. Bulpitt CJ. Subgroup analysis. *Lancet.* Jul 2 1988;2(8601):31-34.

21. Varadhan R, Seeger J. Estimation and Reporting of Heterogeneity of Treatment Effects in Observational CER. In: Velentgas P, Dreyer NA, eds. *User Guide for Developing a Protocol for Observational Comparative Effectiveness Research (OCER)*. Rockville, MD: Agency for Healthcare Research and Quality; [in press].
22. Zulman DM, Sussman JB, Chen X, Cigolle CT, Blaum CS, Hayward RA. Examining the evidence: a systematic review of the inclusion and analysis of older adults in randomized controlled trials. *Journal of General Internal Medicine.* Jul 2011;26(7):783-790.
23. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Annals of Internal Medicine.* Jan 1 1992;116(1):78-84.
24. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet.* Mar 25 2000;355(9209):1064-1069.
25. Lagakos SW. The challenge of subgroup analyses--reporting without distorting. *New England Journal of Medicine.* Apr 20 2006;354(16):1667-1669.
26. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *Bmj.* 2010;340:c117.
27. Petticrew M, Tugwell P, Kristjansson E, Oliver S, Ueffing E, Welch V. Damned if you do, damned if you don't: subgroup analysis and equity. *Journal of epidemiology and community health.* Jan 2012;66(1):95-98.
28. Weiss CO, Segal JB, Boyd CM, Wu A, Varadhan R. *A Framework to Identify and Address Heterogeneity of Treatment Effect in Comparative Effectiveness Research.* Rockville, MD2010.
29. Goodman SN. A comment on replication, p-values and evidence. *Statistics in Medicine.* May 1992;11(7):875-879.
30. Janse AJ, Gemke RJ, Uiterwaal CS, van der Tweel I, Kimpen JL, Sinnema G. Quality of life: patients and doctors don't always agree: a meta-analysis. *Journal of Clinical Epidemiology.* Jul 2004;57(7):653-661.
31. Moreira ED, Susser E. Guidelines on how to assess the validity of results presented in subgroup analysis of clinical trials. *Revista do Hospital das Clinicas.* Mar-Apr 2002;57(2):83-88.
32. Grouin JM, Coste M, Lewis J. Subgroup analyses in randomized clinical trials: statistical and regulatory issues. *Journal of biopharmaceutical statistics.* 2005;15(5):869-882.
33. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine.* Oct 15 2002;21(19):2917-2930.
34. Wactawski-Wende J, Kotchen JM, Anderson GL, et al. Calcium plus vitamin D supplementation and the risk of colorectal cancer. *The New England Journal of Medicine.* Feb 16 2006;354(7):684-696.
35. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine.* Apr 2001;134(8):663-694.
36. Gabler NB, Duan N, Liao D, Elmore JG, Ganiats TG, Kravitz RL. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? *Trials.* 2009;10:43.
37. Food and Drug Administration. Guidance for Industry Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products. Rockville, MD. 1998.
38. Food and Drug Administration. Draft Guidance for Industry, Clinical Investigators, and Food and Drug Administration Staff Design Considerations for Pivotal Clinical Investigations for Medical Devices. Rockville, MD. 2011.
39. Koopman L, van der Heijden GJ, Hoes AW, Grobbee DE, Rovers MM. Empirical comparison of subgroup effects in conventional and individual patient data meta-

analyses. *International Journal of Technology Assessment in Health Care.* Summer 2008;24(3):358-361.
40. Dhruva SS, Redberg RF. Evaluating sex differences in medical device clinical trials: time for action. *Journal of the American Medical Association.* Mar 21 2012;307(11):1145-1146.
41. Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions. Higgins JPT and Green S, eds. 2011. Updated March 2011. Accessed at www.cochrane-handbook.org on March 20, 2012.
42. Sun X, Heels-Ansdell D, Walter SD, et al. Is a subgroup claim believable? A user's guide to subgroup analyses in the surgical literature. *The Journal of bone and joint surgery. American volume.* Feb 2 2011;93(3):e8.
43. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials.* 2010;11:85.
44. Zwarenstein M, Treweek S, Gagnier JJ, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *Bmj.* 2008;337:a2390.
45. Goodman SN. Multiple comparisons, explained. *American journal of epidemiology.* May 1 1998;147(9):807-812; discussion 815.
46. Hochberg Y, Tamhane AC. *Multiple comparison procedures*. New York, NY: John Wiley & Sons; 1987.
47. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Journal of the American Medical Association.* May 26 2010;303(20):2058-2064.
48. Ioannidis JP. Why most published research findings are false. *Plos Medicine.* Aug 2005;2(8):e124.
49. Goodman S, Greenland S. Assessing the unreliability of the medical literature: A response to "Why most published research findings are false". *Johns Hopkins University, Dept. of Biostatistics Working Papers.* 2007. http://www.bepress.com/cgi/viewcontent.cgi?article=1135&context=jhubiostat.
50. Zacharski LR, Chow BK, Howes PS, et al. Reduction of iron stores and cardiovascular outcomes in patients with peripheral arterial disease: a randomized controlled trial. *JAMA : the journal of the American Medical Association.* Feb 14 2007;297(6):603-610.
51. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. The SOLVD Investigators. *N Engl J Med.* Aug 1 1991;325(5):293-302.
52. Glynn RJ, Koenig W, Nordestgaard BG, Shepherd J, Ridker PM. Rosuvastatin for primary prevention in older persons with elevated C-reactive protein and low to average low-density lipoprotein cholesterol levels: exploratory analysis of a randomized trial. *Annals of Internal Medicine.* Apr 20 2010;152(8):488-496, W174.
53. Fischhoff B. Hindsight not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. 1975. *Qual Saf Health Care.* Aug 2003;12(4):304-311; discussion 311-302.
54. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Annalls of Epidemiology.* Oct 2003;13(9 Suppl):S18-77.
55. Knol MJ, Egger M, Scott P, Geerlings MI, Vandenbroucke JP. When one depends on the other: reporting of interaction in case-control and cohort studies. *Epidemiology.* Mar 2009;20(2):161-166.
56. Stallones RA. The use and abuse of subgroup analysis in epidemiological research. *Preventive Medicine.* Mar 1987;16(2):183-194.
57. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of Clinical Epidemiology.* Dec 2011;64(12):1294-1302.

58. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of Clinical Epidemiology.* Apr 2008;61(4):344-349.
59. VanderWeele TJ, Knol MJ. Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. *Annals of Internal Medicine.* May 17 2011;154(10):680-683.
60. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. *American Journal of Epidemiol.* May 1 2006;163(9):783-789.
61. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Statistics in Medicine.* Aug 15 1999;18(15):1903-1942.
62. Paget MA, Chuang-Stein C, Fletcher C, Reid C. Subgroup analyses of clinical effectiveness to support health technology assessments. *Pharmaceutical statistics.* Nov-Dec 2011;10(6):532-538.
63. Chan AW, Hrobjartsson A, Jorgensen KJ, Gotzsche PC, Altman DG. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *Bmj.* 2008;337:a2299.
64. Boonacker CW, Hoes AW, van Liere-Visser K, Schilder AG, Rovers MM. A comparison of subgroup analyses in grant applications and publications. *American Journal of Epidemiology.* Jul 15 2011;174(2):219-225.
65. Should protocols for observational research be registered? *Lancet.* Jan 30 2010;375(9712):348.
66. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics.* Mar 2002;58(1):21-29.
67. Frangakis C. The calibration of treatment effects from clinical trials to target populations. *Clin Trials.* Apr 2009;6(2):136-140.
68. Chiba Y, VanderWeele TJ. A simple method for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology.* Apr 1 2011;173(7):745-751.
69. Follmann DA. On the Effect of Treatment among Would-Be Treatment Compliers: An Analysis of the Multiple Risk Factor Intervention Trial. *Journal of the American Statistical Association.* 2000;95(452):1101-1109.
70. Zhang JL, Rubin DB. Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death". *Journal of Educational and Behavioral Statistics* Winter 2003;28(4):353–368.
71. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association.* 1996;91(434):444-455.
72. Stuart EA, Perry DF, Le HN, Ialongo NS. Estimating intervention effects of prevention programs: accounting for noncompliance. *Prevention Science.* Dec 2008;9(4):288-298.

Tables and Figure

Table 1. The essential characteristics of the three different types of HTE analyses.

Properties	***Confirmatory HTE Analysis***	***Descriptive HTE Analysis***	***Exploratory HTE Analysis***
Inferential Goal	To test hypotheses related to subgroup effects	To report treatment effects for future synthesis	To generate hypotheses for further study
Number of subgroups analyzed	A small number, typically, one or two	Moderate to large	Not made explicit, but may be large
Scientific rationale and prior evidence for hypotheses	Strong	Immaterial	Weak or none
Pre-specification of data analytic strategy	Fully pre-specified	Fully pre-specified	Not pre-specified
Control of family-wise type I error probability	Should be done	Not needed	Difficult, since it is not obvious how many related tests were performed
Characterization of sampling properties of the statistical estimator (e.g., standard errors, type-I error rate)	Easy to achieve	Possible	Difficult
Power for testing hypothesis	Ideally, study designed to have sufficient power	Likely to be inadequately powered, but this is immaterial	Typically, inadequate power to examine several hypotheses

Table 2. Summary of database searches.

Database	Date Searched	Search strategy
NLM Bookshelf	2/16/12	((guid* OR panel* OR recommend* OR standard* OR statement* OR advi* OR rule* OR principle* OR report OR reports)) AND (((((((subgroup* OR subpopulation*) AND effect*)) OR ((subgroup* OR subpopulation* OR subset*) AND (analys*
NLM Catalog	2/16/12	(((heterogen* AND effect*) OR (effect AND modif*) OR (interaction*) OR ((subgroup* OR subpopulation* OR subset*) AND analy*) OR ((subgroup* OR subpopulation*) AND effect*))) AND ("Registries"[Mesh] OR "Registry"[all fields] OR "registries"[all fields] OR "guidelines as topic"[mesh] OR guideline[all fields] OR guidelines[all fields] OR "guidance"[all fields] OR "guidances"[all fields] OR "best practice"[all fields] OR "best practices"[all fields] OR "user guide"[all fields] OR "user guides"[all fields] OR "user's guide"[all fields])
Current Index to Statistics	3/20/12	[The following 2 Keyword/Title searches were imported into Endnote to identify the union of results: 1-standards OR best practice% OR guideline% or guidance%; 2-subgroup% OR heterogen% + treat% OR effect modif% or interaction. Searches were limited to the years 1980-2012.]

Figure 1. Flow Diagram of Search Results.

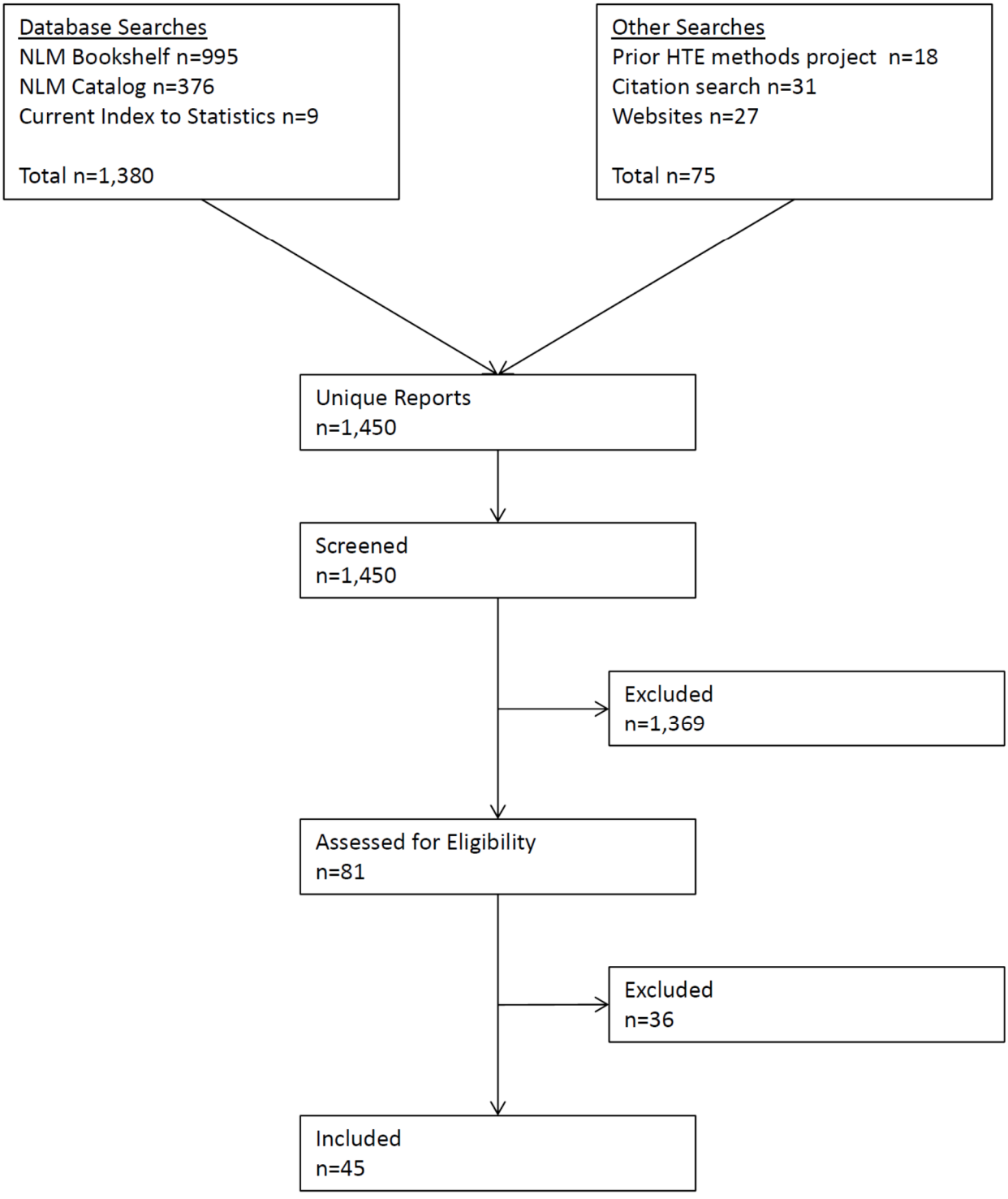

Table 3.

Name of Standard	State the Goals of HTE Analyses
Description of Standard	State the inferential goal of each HTE analysis; identify each analysis as confirmatory, descriptive, or exploratory. See Table 1 that compares the different types of HTE analyses.
Current Practice and Examples	Many reports of subgroup analysis do not clearly identify the inferential goals of the analyses. Examples of confirmatory analyses are quite rare, but it is quite common for authors to not identify HTE analyses as exploratory and to over-emphasize exploratory HTE results.[2,3,18,19]
Published Guidance	Many guidance documents and commentaries support this recommendation.[1,2,13,18,20]
Contribution to Patient Centeredness	This standard by itself contributes indirectly to patient centeredness.
Contribution to Scientific Rigor	A statement of the inferential goals of an analysis is fundamental to ensuring that the study design and analysis are appropriately focused on achieving that goal.
Contribution to Transparency	Identifying the goals of analysis is an important aspect of transparency.
Empirical Evidence and Theoretical Basis	Lack of clarity regarding HTE goals can contribute to misinterpretation of results. Not identifying the inferential goal of an HTE analysis has resulted in authors placing inappropriate emphasis on subgroup findings, and misinterpreting and misapplying of the results by the consumers of the evidence.[2] It should be noted that descriptive HTE is a relatively new category,[21] but it is very important for generating evidence on HTE for future synthesis (e.g., meta-analysis).
Degree of Implementation Issues	This is a straightforward recommendation to implement. All three types, confirmatory, exploratory and descriptive analyses, can be present in the same study. Therefore, each analysis type should be clearly identified. Table 1, which compares the different types of HTE analyses, should be helpful.
Other Considerations	None

Table 4.

Name of Standard	For Confirmatory HTE Analyses, Prespecify a Few Subgroup Hypotheses
Description of Standard	The study protocol should unambiguously prespecify planned confirmatory HTE analyses. Prespecification should include a public record with a clear statement of the hypotheses the study will evaluate, including the definitions of subgroup variables and outcomes, and the direction of the expected treatment effects. Prior evidence should be available for review and the study protocol should present this evidence clearly.
Current Practice and Examples	There are many examples of reports that do not clearly indicate whether they contained prespecified analyses.[2,3,22] Even when the analyses were reported as being prespecified, adequate information (e.g., scientific rationale, prior evidence, definitions of subgroups including categorization of continuous variables, and description of analytic method) is not provided to assess the completeness of prespecification and to ascertain whether there were any deviations from the protocol.
Published Guidance	There are numerous guidance documents that support this standard.[2,3,10,13,18,20,21,23-28]
Contribution to Patient-Centered Care	This standard by itself contributes indirectly to patient-centeredness.
Contribution to Scientific Rigor	Prespecification contributes to scientific rigor by requiring investigators to think carefully about HTE during the study design, conduct, and analysis stages of investigation. This should ensure that the main variables including subgroup definitions are measured well, and when possible, used for stratified randomization. Prespecification of a small number of hypotheses strongly motivates a critical evaluation of prior knowledge.
Contribution to Transparency	Prespecification allows readers to judge whether the reported analyses were proposed in the study protocol or were post-hoc or data-responsive. This is essential for judging the strength of prior evidence to support the hypotheses, as well for judging the reliability of the results.
Empirical Evidence and Theoretical Basis	Prespecification enhances the credibility of subgroup inferences. It serves as an antidote to the practice of selective reporting. Studies have shown that the unreliability of post-hoc subgroup findings is not conveyed by P-values.[29] Many post-hoc subgroup findings have subsequently been disproven (see Table 3 of Rothwell[13]; and Table 1 of Yusuf[18]).
Degree of Implementation Issues	There are currently no standards, nor grading systems, for prespecification. Prespecification can run the gamut from complete absence (post-hoc analysis) to full specification (scientific rationale, prior evidence, definitions of subgroups including categorization of continuous variables, direction of the expected treatment effects, and description of analytic method).
Other Considerations	This minimum standard closely complements another, to "For Any HTE Analysis, Describe the Analytical Methods in Detail"

Table 5.

Name of Standard	**For Confirmatory HTE Analyses, Involve Stakeholders in the Selection of Subgroups and Outcomes**
Description of Standard	Stakeholders should be involved in the selection of subgroups and outcomes for confirmatory HTE analyses.
Current Practice and Examples	We are unaware of data regarding how often this is done, nor of published examples illustrating stakeholder involvement in HTE analyses.
Published Guidance	Stakeholder involvement is recommended for PCOR in general. We are unaware of guidance documents that recommend stakeholder involvement in the choice of HTE analyses.
Contribution to Patient-Centered Care	This standard is the *sine qua non* of patient-centered research.
Contribution to Scientific Rigor	This standard contributes indirectly to scientific rigor.
Contribution to Transparency	This standard contributes indirectly to transparency.
Empirical Evidence and Theoretical Basis	This standard is needed to conduct patient-centered research. It is well known that health care providers are not able to accurately predict patient preferences.[30] This evidence provides strong face validity for the need to involve stakeholders to conduct patient-centered research. However, we are unaware of studies evaluating the degree of disagreement between investigators and stakeholders with respect to study design decisions.
Degree of Implementation Issues	The large number of possible distinct stakeholders presents a challenge to prioritization of a small number of confirmatory HTE analyses.
Other Considerations	None

Table 6.

Name of Standard	For Confirmatory HTE Analyses, Report *a priori* Statistical Power
Description of Standard	Studies should calculate and report the power to detect treatment effects in each subgroup and to detect the interaction between the treatment and the subgrouping variable (i.e., the power to test whether the effects are statistically different between particular subgroups).
Current Practice and Examples	Studies appear to be rarely powered to detect HTE, although they might be more common in the future due to the availability of large healthcare databases. For example, in a review of 63 cardiovascular trials, Hernandez et al.[19] did not find any examples of a trial powered to detect a subgroup effect. However, Moreira et al.[31] found 4 of 17 articles where power for subgroup effect was addressed. It is important to note that this standard does not call for studies to have sufficient power (e.g., 80 percent), but only requires that they report the *a priori* power to study subgroup effects and to detect interaction for given subgroup sample sizes.
Published Guidance	Several guidance documents and commentaries call for adequate sample size to detect HTE.[1,12,13,18,31-33]
Contribution to Patient-Centered Care	This standard contributes indirectly to patient-centeredness.
Contribution to Scientific Rigor	Power calculation promotes increased rigor by ensuring that the investigators examine prior evidence in order to collect information on plausible effect sizes and their variation across subgroups.
Contribution to Transparency	This standard contributes to transparency by helping the readers assess the reliability of study findings, especially when the study did not identify any heterogeneity in the treatment effect.
Empirical Evidence and Theoretical Basis	An estimate of the study power can help in evaluating the reliability of a null finding (i.e., no significant difference in treatment effects between subgroups). A study with low power to detect HTE is much less likely to detect HTE. Consequently, when such a study finds no HTE, it provides little confidence that the treatment effects are not heterogeneous.
Degree of Implementation Issues	A major challenge to implementing this requirement is to obtain reliable information on both the treatment effects in subgroups and the expected sample sizes of subgroups. However, this ought to be possible for confirmatory HTE analyses.
Other Considerations	None

Table 7.

Name of Standard	**For Confirmatory and Descriptive HTE Analyses, Report Sufficient Information on Treatment Effect Estimates**
Description of Standard	Within each subgroup level, studies should present treatment effect estimates, standard errors, and 95 percent confidence intervals. Studies should also report the P-value for the interaction test for each subgrouping variable. For descriptive analyses, studies should also consider presenting a forest plot as a visual summary of the results, although such forest plots should not be used to infer HTE.[34]
Current Practice and Examples	Current practice is quite varied. There are many examples of studies that present P-values for subgroup-specific effects,[1,35] but they less frequently present confidence intervals. Studies seldom present the standard error for the treatment effect estimate. Forest plots are becoming more common in leading journals for presenting HTE analytic results.
Published Guidance	Guidance documents recommend reporting confidence intervals, but not P-values, for subgroup effects.[36] They also recommend reporting interaction tests and their P-values.[1,35] Lagakos[25] recommends a forest plot for a visual summary of subgroup-specific treatment effects.
Contribution to Patient-Centered Care	This standard contributes indirectly to patient-centeredness.
Contribution to Scientific Rigor	Presentation of appropriate statistical summary of results contributes directly to scientific rigor.
Contribution to Transparency	This standard contributes directly to transparency.
Empirical Evidence and Theoretical Basis	Studies should present standard errors of treatment effect estimates because they are essential for meta-analysis. Even though one can extract standard errors from 95 percent confidence intervals, it's better to actually report them. P-values for interaction tests are useful to infer the presence of the.
Degree of Implementation Issues	It is straightforward to implement this standard.
Other Considerations	None

Table 8.

Name of Standard	For Descriptive HTE Analyses, Prespecify Subgroups and Outcomes
Description of Standard	The study protocol should prespecify all subgroups for descriptive HTE analyses. The hypotheses need not be prespecified, rather the subgroups to be studied, because the goal is to facilitate future meta-analyses.
Current Practice and Examples	As a prominent example, the FDA promotes routine examination of prespecified special populations such as the elderly, children, racial, and sex groups.[37,38] When there is insufficient power to analyze prespecified subgroup differences within a study, it may be possible to combine data from various studies and conduct meta-analysis or apply advanced statistical methods, such as Bayesian methods. However, meta-analysis or Bayesian inference is hampered by frequent lack of sufficient and consistent reporting of subgroup treatment effects and their standard errors.[39]
Published Guidance	There is an increasing call for sex-specific reporting of treatment effects, which is an example of the value of descriptive HTE analyses.[6,7,40] Some medical research journals, such as the Journal of the National Cancer Institute, recommend sex and race-specific reporting.[5] The priority populations identified by Agency for Healthcare Research and Quality may also be considered for HTE analysis.[21]
Contribution to Patient-Centered Care	This standard contributes indirectly to patient centeredness.
Contribution to Scientific Rigor	Prespecification should ensure that the main variables including subgroup definitions are measured well. This standard facilitates future meta-analysis, which is an important method for combining evidence.
Contribution to Transparency	Prespecification is essential for judging transparency. It allows readers to judge whether the reported analyses were proposed in the study protocol or were post-hoc or data-responsive. This is essential for judging the reliability of the results.
Empirical Evidence and Theoretical Basis	Prespecification enhances the credibility of subgroup inferences. It serves as an antidote to the practice of selective reporting. Inconsistent or selective reporting of subgroups or outcomes creates bias during meta-analysis. [41] When there is insufficient power to analyze prespecified subgroup differences within a study, it may be possible to combine data from various studies and conduct meta-analysis or apply advanced statistical methods, such as Bayesian methods. This is the theoretical basis for descriptive HTE analysis.
Degree of Implementation Issues	While the FDA approval process already includes this standard, a broader audience for this standard will invite discussion of which subgrouping variables should be prioritized for descriptive HTE analysis. Consistency of reporting across studies is critical for descriptive HTE analyses.
Other Considerations	None

Table 9.

Name of Standard	**For Descriptive HTE Analyses, Involve Stakeholders in the Selection of Subgroups and Outcomes**
Description of Standard	Stakeholders should be involved in the selection of subgroups and outcomes for descriptive HTE.
Current Practice and Examples	We are unaware of data regarding how often this is done, nor of published examples illustrating stakeholder involvement in HTE analyses.
Published Guidance	Stakeholder involvement is generally recommended for PCOR. We are unaware of any guidance recommending stakeholder involvement when planning HTE analyses.
Contribution to Patient-Centered Care	Stakeholder involvement is considered a *sine qua non* of patient-centered research.
Contribution to Scientific Rigor	This standard contributes indirectly to scientific rigor.
Contribution to Transparency	This standard contributes indirectly to transparency.
Empirical Evidence and Theoretical Basis	This standard is essential to patient-centered research. It is well known that health care providers are not able to accurately predict patient preferences.[30] While there could be disagreement between investigators and stakeholders regarding study designs, few, if any, studies have evaluated this.
Degree of Implementation Issues	The large number of possible distinct stakeholders presents a challenge to prioritization of a relatively small number of descriptive HTE analyses.
Other Considerations	None

Table 10.

Name of Standard	For Exploratory HTE Analysis, Document the Number of Subgroups and Outcomes Analyzed
Description of Standard	Reports of exploratory HTE analyses should clearly document the number of subgroups and outcomes analyzed.
Current Practice and Examples	There are many examples of trials that do not clearly indicate how many subgroup analyses were performed.[3] Reviews find many examples of trials that report results for greater than 10 subgroups,[22,36,42] yet it was not clear if the trials reported all subgroup analyses. The total number of subgroup analyses may reasonably be assumed to equal the product of the number of subgrouping variables and the number of outcomes studied.[42]
Published Guidance	Several guidance documents support this standard.[2,3,12,21,28,31,36,43,44]
Contribution to Patient-Centered Care	This standard contributes indirectly to patient centeredness.
Contribution to Scientific Rigor	An understanding of the total number of analyses is necessary to assess the role of chance in exploratory findings and to make appropriate inferences regarding whether exploratory findings provide evidence that supports future confirmatory analysis or other actions.
Contribution to Transparency	Clear documentation is synonymous with transparency.
Empirical Evidence and Theoretical Basis	It can be shown that the type I error probability increases as the number of analyses under the null hypothesis of no treatment effect increases.[25,45] While there are numerous techniques to adjust for testing multiple HTE hypotheses,[46] simply reporting the number of hypotheses that were examined is often adequate to give the readers a sense for the likelihood of false positives.[3]
Degree of Implementation Issues	This standard depends on accurate reporting by researchers. Such accuracy is difficult to confirm without prespecification, but prespecification may not be appropriate for some exploratory, data-responsive analyses.
Other Considerations	None

Table 11.

Name of Standard	For Exploratory HTE Analyses, Discuss Findings in the Context of Study Design and Prior Evidence
Description of Standard	Exploratory HTE analyses should be presented in the context of whether they are consistent with prior evidence and how well the study design addresses the HTE question. These considerations are more important than P-values for inferences.
Current Practice and Examples	There are many examples of trials where exploratory HTE results have been over-emphasized.[2,19,47] This has resulted in a high rate of non-replication of research discoveries (i.e., false discovery rate) in the scientific literature.[48,49] There are also examples of appropriately cautious presentation of exploratory HTE findings.[50-52]
Published Guidance	Several guidance documents support this standard.[2,13,19,21,26,28,32]
Contribution to Patient-Centered Care	This standard contributes indirectly to patient centeredness.
Contribution to Scientific Rigor	This standard contributes to scientific rigor by requiring investigators to think carefully about how well the study design addresses the HTE question. For example high-quality measurement of exposure to treatment is of critical importance for studying HTE. Placing findings in the context of prior evidence is also essential for scientific rigor, especially for exploratory analyses that were not prespecified.
Contribution to Transparency	This standard will encourage readers to judge whether the reported analyses provide evidence to support future confirmatory analyses or other actions.
Empirical Evidence and theoretical Basis	Studies can't measure statistical properties, such as bias and error, in exploratory analyses. Investigators' post-hoc prior probabilities are subconsciously biased.[53] In fact, many post-hoc subgroup findings have subsequently been disproven (see Table 3 of Rothwell[13]; and Table 1 of Yusuf[18]). The unreliability of post-hoc subgroup findings is not conveyed by P-values.[29] Therefore interpretation should focus on study quality and consistency with prior evidence.
Degree of Implementation Issues	Post-hoc judgments tend to be psychologically biased, which can also affect the (*a posteriori*) discussion of study design and prior evidence.
Other Considerations	None

Table 12.

Name of Standard	For Any HTE Analysis in Observational Data, Explicitly Assess Study Quality for Making Causal Inference
Description of Standard	HTE analyses based on observational data should assess study quality for making causal inference.
Current Practice and Examples	The goal of HTE analyses is to make inferences regarding a treatment's effect. This requires that studies isolate the treatment effect from confounding factors. Observational studies often do not focus on a specific treatment, its confounders, or treatment-specific outcomes. There are many examples of observational studies that meet this standard.[54-56]
Published Guidance	Two guidance documents address this standard for HTE analyses.[21,28] There are other guidance documents to assess the comprehensiveness of reporting of observational studies.[57,58] These, however, do not directly address the study of HTE.
Contribution to Patient-Centered Care	This standard contributes indirectly to patient centeredness.
Contribution to Scientific Rigor	This standard contributes to scientific rigor by requiring that investigators in observational studies provide an explicit assessment of study quality when making a causal inference regarding treatment effect.
Contribution to Transparency	This standard contributes indirectly to transparency.
Empirical Evidence and Theoretical Basis	Observational studies usually collect information on many variables and allocate resources broadly rather than in a focused manner. Some observational datasets collect information that was not collected for research purposes (e.g., billing data). Some observational datasets contain a high number of participants, making statistically significant findings likely, regardless of bias. There are numerous examples where strong observational evidence has been overturned in experimental studies. Ascertainment and selection biases in exposure to treatment, measurement error in assessment of health outcomes, and lack of information on important prognostic variables can result in biased estimates of treatment effect and its heterogeneity (e.g., introduce apparent HTE when in fact none is present, or conversely, obscure HTE when it is actually present).[59]
Degree of Implementation Issues	None
Other Considerations	None

Table 13.

Name of Standard	For Any HTE Analysis, Describe the Analytical Methods in Detail
Description of Standard	Studies should describe analytic methods in sufficient detail to facilitate replication by external investigators. Studies should also clearly identify prespecified, post-hoc, and data-responsive elements of the analytic plan. Any deviations from the prespecified plan should be acknowledged and justified.
Current Practice and Examples	There are many positive and negative examples of clear reporting of prespecified analyses.[2,3,22] Even when reports stated that analyses were prespecified, they often did not provide analytic details.
Published Guidance	While virtually all guidance documents recommend detailed statistical reporting, or appear to assume it is part of good scientific practice, there are several guidance documents that support prespecification of the analytic plan for HTE.[11,13,21,28,35,37] An outline of standards for reproducible research has been proposed.[60]
Contribution to Patient-Centered Care	This standard contributes indirectly to patient centeredness. Stakeholder input may be considered for this step (for PCOR in general). For example, stakeholder input could possibly be informative for selecting cutoff values for continuous variables used to define subgroups.
Contribution to Scientific Rigor	Detailed description of the analytic plan makes it possible to attempt replication and to understand statistical bias and error. Reporting this will convey information regarding the detail and strength of prior evidence.
Contribution to Transparency	Detailed description is virtually synonymous with transparency of statistical methods. It allows readers to judge whether the reported analyses were proposed in a protocol, were post-hoc, or data-responsive. This is essential for judging the inferential strength of analyses. This transparency, rather than strict adherence to a prespecified protocol, is the major goal of this minimum standard.
Empirical Evidence and Theoretical Basis	Detailed description of the analytic plan makes it possible to attempt replication and to understand statistical bias and variance. A detailed case study in promoting reproducible research is available.[60]
Degree of Implementation Issues	The FDA approval process already includes this standard. However, it is sometimes challenging for investigators to publish a full description of analytic plans due to journal word number limits. Investigators may also feel that it is inappropriate to prespecify all analytic details.
Other Considerations	None

Table 14.

Name of Standard	For Any HTE Analysis, Perform an Interaction Test
Description of Standard	To detect differences in treatment effect between subgroups, use an interaction test (i.e., test whether the interaction between the treatment indicator and the subgroup variable is statistically significant).
Current Practice and Examples	When the treatment effect is significant in one of the subgroups, but not in the other, it is very common for investigators to infer the presence of an interaction between the treatment and a binary subgroup variable.[2,19,26] This is flawed. The correct approach is to use an appropriate test of interaction. There are many examples of proper use of an interaction test.[2,3,19,22]
Published Guidance	There are numerous guidance documents and commentaries that support this recommendation.[1-3,12,13,18,19,25,26,33,35,36,43,61,62]
Contribution to Patient-Centered Care	This standard contributes to indirectly patient-centeredness.
Contribution to Scientific Rigor	This standard contributes to scientific rigor. An interaction test is the correct way to assess HTE. It has the correct 5 percent probability of rejecting the null hypothesis of no HTE, when in fact there is no HTE.
Contribution to Transparency	This standard contributes indirectly to transparency.
Empirical Evidence and Theoretical Basis	An interaction test is a direct way to gauge whether the treatment effect varies across levels of a subgroup factor. A flawed, but widely used, approach infers the presence of an interaction between the treatment and a binary subgroup variable when the treatment effect is significant in one of the subgroups, but not in the other. Brookes[1] showed that when using this flawed approach, even when there is no actual HTE, there is up to 66 percent probability of inferring the presence of HTE when the overall treatment effect is significant; and that there is up to 25 percent probability of inferring HTE when the overall treatment effect is non-significant. The interaction test, however, maintains the correct 5 percent probability regardless of whether the overall treatment effect was significant or not.
Degree of Implementation Issues	This recommendation can be implemented easily using standard software.
Other Considerations	It is important to understand that interaction tests have low power. For a 50:50 subgroup split, it requires nearly 4 times the sample size to detect an interaction effect of the same magnitude as the overall treatment effect; and it requires nearly 16 times the sample size to detect an interaction effect that is half the size of the overall effect. Therefore, a non-significant interaction test does not necessarily mean that there is no THE; the null finding could be due to a lack of power.

Table 15.

Name of Standard	For Any HTE Analysis, Report All Prespecified Analyses
Description of Standard	Studies must report the results of all the HTE analyses that were prespecified in the study protocol or grant application, regardless of their statistical significance.
Current Practice and Examples	There are both positive and negative examples of discrepancies between the description of HTE analysis in the study protocols and grant applications.[2,63,64]
Published Guidance	There are several guidance documents that support this recommendation.[2,18,44]
Contribution to Patient-Centered Care	This standard contributes indirectly to patient-centeredness.
Contribution to Scientific Rigor	This standard contributes to scientific rigor by eliminating the bias that can be introduced if the investigators only reported favorable, interesting, or statistically significant results.
Contribution to Transparency	This standard contributes directly to transparency in reporting HTE results. The practice of selective reporting can be more deleterious to scientific progress than flawed design or analysis, since such flaws are open to the scrutiny and can be detected, whereas selective reporting is difficult to detect.
Empirical Evidence and Theoretical Basis	Prespecification enhances the credibility of subgroup inferences. It serves as an antidote to the practice of selectively reporting interesting findings from post-hoc subgroup analyses. Some authors insist on reporting all HTE analyses, both prespecified and post-hoc.[18,43] However, this might create too much information. It seems more sensible to report all the prespecified analyses and a count of the total number of analyses that the study conducted.
Degree of Implementation Issues	It is somewhat difficult to ensure the fulfillment of this recommendation. One way to ensure this would be to establish a public registry of protocols for studies funded by PCORI.[16,65]
Other Considerations	None

Table 16.

Name of Standard	For Any HTE Analysis, Use Appropriate Methods for Post-treatment Subgroups
Description of Standard	Studies must use appropriate statistical methods when conducting subgroup analysis based on post-treatment variables (i.e., variables that are likely to be affected by the treatment).
Current Practice and Examples	Examples of such subgroup analyses are the "responders" analysis and "compliers" analysis.[13,20] There are examples of appropriate use of methods for post-treatment subgroups.[66-70]
Published Guidance	Many guidance documents and commentaries recommend that studies not use post-randomization variables as subgroups.[13,20] Recent advances in causal inference have proposed methods for providing valid treatment effect estimates that attempt to capture the "biological" effect of the treatment (principal stratification[66]; instrumental variables[71]). For studies to use these methods appropriately, however, requires expert statistical knowledge. There is little guidance available for the broader use of these techniques.
Contribution to Patient-Centered Care	This standard contributes indirectly to patient-centeredness.
Contribution to Scientific Rigor	This standard contributes to scientific rigor by addressing the bias in treatment effects that can result from analyzing subgroups whose definition is impacted by the treatment.
Contribution to Transparency	This standard contributes indirectly to transparency.
Empirical Evidence and Theoretical Basis	Analysis of subgroups, defined on the basis of variables that are affected by the treatment, can result in biased estimation of the treatment effects in those subgroups.[66,72]
Degree of Implementation Issues	Studies should avoid post-treatment subgroups, unless expert statistical knowledge is available to handle the analytic challenges properly.
Other Considerations	None

Appendix Table 1. Description of Guidance Statements Included in the Recommended Minimum Guidelines

Guideline	Organization or Authors	Year	Program	Country or Region	Guideline subjected to independent external review?	Research Design	Description
The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration	Douglas G. Altman, DSc; Kenneth F. Schulz, PhD; David Moher, MSc; et al. for the CONSORT Group	2001	CONSORT Group	United Kingdom	Yes	Trials	Explanatory and elaboration document is intended to enhance the use, understanding, and dissemination of the Consolidated Standards of Reporting Trials statement. Addresses subgroup reporting, interaction testing and multiplicity.
Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives	ST Brookes; E Whitley; TJ Peters; PA Mulheran; M Egger; G Davey Smith	2001	**International Network of Agencies for Health Technology Assessment**	International	Yes	Clinical Trials (Simulation studies)	Considers subgroup-specific and formal interaction tests; sample size, magnitude of treatment effect, number of treatment groups; continuous variables, binary outcomes and survival times.
Subgroup Analysis	Christopher J. Bulpitt	1988	Royal Postgraduate Medical School	United Kingdom	Yes	Clinical Trials (Review)	Addresses pre-specification, potential bias, magnitude of overall treatment effect

Cochrane Handbook for Systematic Reviews of Interventions	Cochrane Collaboration	2011	Cochrane Group	U.K.	Yes	Systematic reviews and meta-analysis	A handbook for conducting meta-analyses.
How to guarantee finding a statistically significant difference: the use and abuse of subgroup analyses	Peter M. Fayers; Madeline T. King	2009	University of Aberdeen Medical School and Norwegian University of Science and Technology; University of Sydney	Europe, Australia	Yes	Clinical Trials (Review)	Reviews guidelines for when subgroup analysis is appropriate and how to conduct and report these analyses.
Guidance for Industry Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products	U.S. Department of Health and Human Services Food and Drug Administration	1998	Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER)	United States	Yes (drafts published for public comment)	Clinical Trials (Guidance)	Outlines the quantity and quality of evidence required to support effectiveness claims.
Guidance for Industry E9 Statistical Principles for Clinical Trials	U.S. Department of Health and Human Services Food and Drug Administration	1998	CDER, CBER	United States	Yes (drafts published for public comment)	Clinical Trials (Guidance)	Presents detailed instructions to drug application sponsors, including for special population subgroup analyses.
Guideline for the Format and Content of the Clinical and Statistical Sections of an	U.S. Department of Health and Human Services Food and Drug	1988	Center for Drug Evaluation and Research (CDER)	United States	Yes (drafts published for public comme	Clinical Trials (Guidance)	Presents detailed instructions to drug application sponsors, including for special population subgroup analyses.

Application	Administration				nt)		
Draft Guidance for Industry, Clinical Investigators, and FDA Staff Design Considerations for Pivotal Clinical Investigations for Medical Devices	U.S. Department of Health and Human Services Food and Drug Administration	2011	CDRH, CBER	United States	Underway	Clinical Trials (Guidance)	Encourages appropriate use of stratified subgroup design to ensure representation of target populations
Assessing Heterogeneity of Treatment Effects: Are Authors Misinterpreting Their Results?	Erik Fernandez y Garcia; Hien Nguyen; Naihua Duan; Nicole B. Gabler; Richard L. Kravitz	2010	UC Davis; Columbia University; Center for Healthcare Policy and Research, Sacramento, CA	United States	Yes	Clinical Trials (Review)	Reviews the literature and suggests guidelines to improve consistency in interpretation and reporting of HTE analysis.
Dealing with heterogeneity of treatment effects: is the literature up to the challenge?	Nicole B Gabler, Naihua Duan, Diana Liao, Joann G Elmore, Theodore G Ganiats and Richard L Kravitz	2009	UC Davis; UC San Diego; Washington University	United States	No	Clinical Trials (Review)	A literature sample to track the use of HTE analyses over time, examine the appropriateness of the statistical methods used, and explore the predictors of such analyses.
Subgroup Analyses in Randomized Clinical Trials: Statistical and Regulatory Issues	Jean-Marie Grouin; Maylis Coste; John Lewis	2006	University of Rouen; I.R.I. Servier; Silverton Lodge, UK	European Union	Yes	Clinical Trials	Outlines reasons for subgroup analysis, considerations for trial design, analysis and reporting.

GRADE guidelines: 7. Rating the quality of evidence-inconsistency	Gordon H. Guyatt; Andrew D. Oxman; Regina Kunz; James Woodcock; et al.	2011	GRADE Working Group	International	Yes	Systematic Reviews (Guideline)	Addresses inconsistency of treatment effects across subgroups during guideline development.
Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading?	Adrian V. Hernandez; Eric Boersma; Gordon D. Murray; J. Dik F. Habbema; Ewout W. Steyerberg,	2006	Erasmus University Medical Center, Rotterdam; University of Edinburgh Medical School	Netherlands; United Kingdom	Yes	Clinical Trials (Review)	Reviews practices related to reporting HTE in cardiovascular journals.
Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal	David M Kent, Peter M Rothwell, John PA Ioannidis, Doug G Altman, Rodney A Hayward	2010	Tufts Medical Center, John Radcliffe Hospital, Oxford, UK, University of Ioannina School of Medicine, University of Oxford, University of Michigan	U.S., U.K., Greece	No	Clinical trials	Based upon recent evidence on optimal statistical approaches to assessing HTE, we propose a framework that prioritizes the analysis and reporting of multivariate risk-based HTE.
The Challenge of Subgroup Analyses — Reporting without Distorting	Stephen W. Lagakos	2006	Harvard School of Public Health	U.S.	No	Clinical trials	Illustrates role of chance in false positive results with multiple tests.

Reporting on methods of subgroup analysis in clinical trials: a survey of four scientific journals	E.D. Moreira Jr., Z. Stein and E. Susser	2001	Fundação Oswaldo Cruz, Salvador, Brasil Columbia University, New York, NY,	Brazil, U.S.	No	Clinical trials	A survey and list of important methodological items.
A Consumer's Guide to Subgroup Analyses	Andrew D. Oxman and Gordon H. Guyatt	1992	McMaster University, Ontario, CA	Canada	No	Trials and meta-analyses	Presents guidelines to readers of trials and meta-analyses
Subgroup analyses of clinical effectiveness to support health technology assessments	Marie-Ange Paget, Christy Chuang-Stein, Christine Fletcher and Carol Reid	2011	Eli Lilly, France; Pfizer, USA; Amgen Ltd, UK; Roche, UK	France, U.S., U.K.	No	Health technology assessments	Describes good statistical principles for subgroup analyses of clinical effectiveness to support HTAs and include case examples where HTA recommendationswere given to subpopulations only
Damned if you do, damned if you don't: subgroup analysis and equity	Mark Petticrew, Peter Tugwell, Elizabeth Kristjansson, Sandy Oliver, Erin Ueffing, Vivian Welch5	2011	London School of Hygiene and Tropical Medicine, London, UK; University of Ottawa, Ottawa, Ontario, Canada;	International	No	Diverse	This paper considers some of the methodological problems with subgroup analysis, and its applicability to considerations of equity, using both clinical and public health examples.
Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial	Stuart J. Pocock, Susan E. Assmann, Laura E. Enos and Linda E. Kasten	2002	London School of Hygiene & Tropical Medicine; London; U.K.;	International	No	Clinical trials	This paper examines how these issues are currently tackled in the medical journals, based on a recent survey of 50 trial reports in four major journals. The statistical ramifications are explored, major problems are highlighted and recommendations for

reporting: current practice and problems			New England Research Institutes; U.S.				future practice are proposed.
Subgroup analysis in randomised controlled trials: importance, indications, and interpretation	Peter M Rothwell	2005	Radcliffe Infirmary, UK	UK	No	Clinical trials	Formal rules for the planning, analysis, and reporting of subgroup analyses are proposed.
Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses	Xin Sun, Matthias Briel, Stephen D Walter,Gordon H Guyatt	2010	McMaster University, Hamilton, ON, Canada; Sichuan University, Chengdu, China; University Hospital Basel, Basel, Switzerland	International	No	Clinical trials	This article identifies new criteria and proposes a checklist for judging the credibility of subgroup analyses.
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration	Jan P. Vandenbroucke; Erik von Elm,; Douglas G. Altman; Peter C. Gøtzsche; Cynthia D. Mulrow; Stuart J. Pocock; Charles	2007	Numerous	International	Yes	Observational studies	The STROBE Statement consists of a checklist of 22 items, which relate to the title, abstract, introduction, methods, results, and discussion sections of articles.

	Poole; James J. Schlesselman; and Matthias Egger, for the STROBE initiative						
Estimation and Reporting of Heterogeneity of Treatment Effects in Observational CER	Ravi Varadhan and John Seeger	2012	Johns Hopkins University and Harvard University	U.S.	Yes	Observational studies	A chapter that focuses on defining and describing HTE and how to evaluate and report such heterogeneous effects using subgroup analysis.
Statistics in medicine — reporting of subgroup analyses in clinical trials.	Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM.	2007	New England Journal of Medicine; Harvard University	U.S.	No	Clinical trials	Recommendations for reporting subgroup analyses.
A Framework to Identify and Address Heterogeneity of Treatment Effect in Comparative Effectiveness Research	Carlos Weiss, Jodi Segal, Cynthia Boyd, Albert Wu and Ravi Varadhan	2011	Johns Hopkins University	U.S.	Yes	Observational studies and clinical trials	At the request of the Agency for Healthcare Research and Quality, developed a framework for identifying HTE in comparative effectiveness studies which includes a review of analytic methods that aim toward appropriate application of the results of studies which have HTE.
Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials	Salim Yusuf; Janet Wittes; Jeffrey Probstfield; Herman A. Tyroler	1991	National Heart, Lung, and Blood Institute; New England Research Institute; University of North	U.S.	No	Trials	Recommends examining the architecture of the entire set of subgroups within a trial, analyzing similar subgroups across independent trials, and interpreting the evidence in the context of known biologic mechanisms and patient prognosis

			Carolina, Chapel Hill				
Improving the reporting of pragmatic trials: an extension of the CONSORT statement	Merrick Zwarenstein, Shaun Treweek, Joel J Gagnier, Douglas G Altman, Sean Tunis, Brian Haynes, Andrew D Oxman, David Moher and for the CONSORT and Pragmatic Trials in Healthcare (Practihc) groups	2008	Sunnybrook Hospital, Canada; University of Toronto; Karolinska Institute, Stockholm, Sweden; University of Dundee, UK; Norwegian Knowledge Centre for the Health Services, Norway University of Oxford; Center for Medical Technology Policy; Medicine, Johns Hopkins School of Medicine; Stanford University School of Medicine; McMaster University	International	Yes	Pragmatic clinical trials	Recommends extending eight CONSORT checklist items for reporting of pragmatic trials: the background, participants, interventions, outcomes, sample size, blinding, participant flow, and generalisability of the findings.

			Faculty of Health Sciences, Canada; Ottawa Health Research Institute, Canada				

Appendix Table 2. Description of Guidance Statements Not Included in the Recommended Minimum Guidelines

Guideline	Organization or Authors	Year	Program	Country or Region	Guideline subjected to independent external review?	Research Design	Description
Subgroup Analyses	Donald A. Berry	1990	University of Minnesota	United States	No	Clinical Trials (Letter to the Editor)	Discusses two basic approaches to controlling error in subgroup analysis. Addresses multiplicities, prior knowledge, persistence of the null hypotheses, and prespecification.
The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies	Karl Claxton	1999	Harvard School of Public Health	United States	Yes	Cost-effectiveness analysis (Original Report)	A framework for decision making and establishing the value of additional information is presented which is consistent with the decision rules in cost-effectiveness analysis.
A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy	J Dinnes, J Deeks, J Kirby and P Roderick	2005	University of Southampton, UK; Centre for Statistics in Medicine, Oxford, UK	U.K.	Yes	Diagnostic test studies	Systematic review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy studies.
Expert workshop on subgroup analysis	European Medicines Agency	2011	Human Medicines Development	European Union	Not clear	Clinical Trials (Workshop	This document contains summaries of presentations and concluding remarks on trial

			and Evaluation			Report)	design and analysis.
Points to Consider on Multiplicity issues in Clinical Trials	European Medicines Agency	2002	Committee for Proprietary Medicinal Products	European Union	Not clear	Clinical Trials (Guidance)	Considers when adjustment for multiplicity is necessary, how to interpret significance of multiple secondary variables, when conclusions drawn from subgroup analysis are reliable, when it is appropriate to restrict license to a subgroup, how to interpret analysis of responders, how to handle composite endpoints with respect to regulatory claims
Reviewer Guidance: Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review	U.S. Department of Health and Human Services Food and Drug Administration	2005	Center for Drug Evaluation and Research	U.S.	Yes	Safety studies	Handbook for examining safety studies.
Guidance for Industry and FDA Staff Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials	U.S. Department of Health and Human Services Food and Drug Administration	2010	Center for Devices and Radiological Health	U.S.	Yes (drafts published for public comment)	Clinical Trials (Guidance)	Encourages appropriate use of Bayesian statistics to address multiplicity adjustments
Guidance for Industry: Drug Interaction Studies — Study Design, Data Analysis, Implications for Dosing, and Labeling Recommendations	U.S. Department of Health and Human Services Food and Drug Administrati	2012	Center for Drug Evaluation and Research	U.S.	(draft published for public comment)	Drug interaction studies	Discusses statistical consideration.

	on						
Clinical Heterogeneity in Systematic Reviews and Health Technology Assessments: Synthesis of Guidance Documents and the Literature	Gerald Gartlehner; Charles Poole; Suzanne L. West; Alyssa J. Mansfield; Elizabeth Tant; Linda J. Lux; Kathleen N. Lohr	2012	Danube University; University of North Carolina at Chapel Hill; RTI International	Internatio nal	Yes	Systematic Reviews and Health Technolog y Assessme nts (Review)	Synthesis of best practices
Statistical Principles for Clinical Trials (ICH E9) An Introductory Note on an International Guideline	John A. Lewis	1999	Medical Control Agency	Internatio nal	Yes	Clinical Trials (Guideline)	Discusses prespecification of statistical plans.
Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication	Internationa l Committee of Medical Journal Editors	2010	International Committee of Medical Journal Editors	Internatio nal	No	Biomedical Publication (Guideline)	Discusses clear reporting of which subgroup analyses were prespecified.
Statistical Analysis Plans: Principles and Practice	James R. Johnson; David Fitts	2007	Society for Clinical Trials	Internatio nal	No	Clinical Trials	An overview of statistical analysis plans that considers all subgroup analyses exploratory.
Research methods for subgrouping low back pain	Peter Kent, Jennifer L Keating and Charlotte Leboeuf-Yde	2010	Monash University, Melbourne, Australia	Internatio nal	No	Diverse	The proposed method framework proposes six phases for studies of subgroups: studies of assessment methods, hypothesis-setting studies, hypothesis-testing studies, narrow validation studies, broad validation studies, and impact analysis studies.
Meta-regression Approaches: What, Why, When, and	Morton SC, Adams JL, Suttorp MJ,	2004	Southern California—RAND	U.S.	Yes	Meta-analysis	Simulation results produced specific guidelines for meta-regression

How?	Shekelle PG		Evidence-Based Practice Center				practitioners that may be summarized in the key message that the causes of heterogeneity should be explored via the inclusion of covariates at both the person level and study level.
Subgroups and Heterogeneity in Cost-Effectiveness Analyses	Mark Sculpher	2008	University of York, UK	International	No	Cost-effectiveness analysis	Discusses the use of subgroup analysis in cost-effectiveness analysis, which raises a number of methodological questions : a need to define the possible sources of heterogeneity that exist, which extends beyond relative treatment effect; jpw heterogeneity in model parameters should be estimated and how uncertainty should be appropriately quantified.
Comparative Effectiveness Review Methods: Clinical Heterogeneity	Suzanne L. West, Gerald Gartlehner, Alyssa J. Mansfield,. Charles Poole, Elizabeth Tant, Nancy Lenfestey, Linda J. Lux, Jacqueline Amoozegar, Sally C. Morton,	2010	University of North Carolina – Chapel Hill	U.S.	Yes	. Systematic reviews and comparative effectiveness reviews	Recommends clear evidence-based guidance on addressing clinical heterogeneity in SRs and CERs is not available currently but would be valuable to AHRQ's EPCs and to others conducting SRs internationally.

	Ph.D. Timothy C. Carey, Meera Viswanathan, Kathleen N. Lohr						