Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Alemayehu. J Biomet Biostat 2011, 2:5
DOI: 10.4172/2155-6180.1000125

Research Article Open Access

Evaluation of Heterogeneity of Treatment Effects in Comparative Effectiveness Research

Demissie Alemayehu

Abbrevations: AHRQ: Agency for Healthcare Research and Quality; RCTs: Randomized Controlled Trials; CER: Comparative Effectiveness Research

Introduction

Reliable knowledge of heterogeneity of treatment effects is critical in decision making about the relative risks and benefits of competing treatment options for individual patients and subgroups of patients. Despite their widely recognized attributes as the gold-standard for generating evidence relating to comparative efficacy, randomized controlled trials (RCTs) are designed to provide data on treatment efficacy on average for a target population, and thus fall short of generating evidence pertinent at the individual or subgroup level. Extrapolation of results from RCTs to real-world situations is limited by factors inherent to the design of the studies, which include stringent entry criteria that exclude subgroups of interest, logistical constraints to study broader subgroups, and other practical considerations that impose constraints that are inconsistent with real-world practice and adherence.

To optimize patient care, it is important to ensure generalizability of results from RCTs to different individuals, subgroups or settings. Currently, the paradigm for drug development and regulatory approval is not amenable to address this issue. Pivotal studies used to establish the efficacy of a new drug are typically designed without regard to the risks and benefits at the individual of subgroup level, and follow protocols which impose constraints that are at variance with real-world conditions.

Advances in personalized medicine have not yet fully succeeded in customizing healthcare practices to meet the needs of individual patients. The systematic use of genetic or other information about an individual patient to select or optimize that patient's care is at best a work in progress. While there are positive incremental steps in this direction, especially in the field of oncology [1], a more effective strategy would require an ambitious research agenda that aims at synthesizing molecular level information with an individual's clinical history to formulate a treatment algorithm that optimizes a treatment option for a given individual. Until this happens, it is critical to address the issues of assessing heterogeneity in the framework of current drug development and utilization, while recognizing the inherent limitations.

The assessment of heterogeneity in traditional systematic reviews is challenging, and becomes even more complex when one conducts confirmatory analyses, as is done in a typical comparative effectiveness research (CER) exercise. In contrast to exploratory analyses, which are aimed at generating hypotheses or searching for signals in the data, confirmatory analyses generally require pre-specification of methods and general investigational approaches. In this paper, we outline some of the statistical and conceptual issues associated with heterogeneity in confirmatory CER, and chart measures that must be in place to effectively assess and manage heterogeneity. Section 2 elaborates the distinction between clinical and statistical heterogeneity, and in Section 3, we review statistical aspects of heterogeneity, with particular reference to their relations to familiar issues associated with traditional subgroup analysis. Section 4 suggests points to consider when assessing heterogeneity, and Section 5 provides concluding remarks.

Clinical vs Statistical Heterogeneity

In a recent AHRQ publication attempts were made to distinguish between clinical and statistical heterogeneity [2]. At a first glance, the distinction may appear superfluous, but a closer examination may help elucidate the subtle differences, which may have relevance in understanding some of the conceptual and practical issues in addressing heterogeneity.

Statistical heterogeneity relates to the assessment of the degree of variability in the observed treatment effects beyond what would be expected by play of chance. On the other hand, clinical heterogeneity refers to the "variation in study population characteristics, coexisting conditions, cointerventions, and outcomes evaluated across studies included in a systematic review or CER that may influence or modify the magnitude of the intervention measure of effect" [2].

Obviously clinical heterogeneity may lead to statistical heterogeneity, but the converse is not necessarily true. In addition to heterogeneity in underlying population characteristics and clinical conditions, the latter may result from other factors, including methodological differences, sensitivity of statistical procedures, or simply the play of chance. In any case, an effective approach to understanding and managing heterogeneity in CER would require a careful and objective definition of clinical heterogeneity, and an appreciation of the limitations of the techniques that are in routine use to assess statistical heterogeneity. In the next section, we revisit some of the challenges commonly associated with the latter, and subsequently expound the interplay between the two seemingly distinct, but closely related concepts.

Statistical Issues with Heterogeneity Assessment

From a statistical standpoint, the issues associated with heterogeneity analysis correspond mostly to those known when one deals with subgroup analysis in clinical trials [3,4]. However in the context of CER, the problems are compounded by the fact that the number of treatment options involved is typically large, the scope of the analyses wide, and the associated issues with interpretation of the results relatively more complex.

***Corresponding author:** Demissie Alemayehu, E-mail: alemad@pfizer.com

Received October 31, 2011; **Accepted** December 02, 2011; **Published** December 25, 2011

Citation: Alemayehu D (2011) Evaluation of Heterogeneity of Treatment Effects in Comparative Effectiveness Research. J Biomet Biostat 2:125. doi:10.4172/2155-6180.1000125

The first issue with heterogeneity analysis in CER is the problem of bias. As is the case with any systematic review, most CER projects generally involve a retrospective synthesis of data. This opens the door for the introduction of bias, unless necessary measures are in place to minimize its occurrence. Bias is generally a function of the timing of specification of hypotheses being tested and the analytical procedures employed relative to the examination of data. The issue is particularly important when the analyses are motivated by inspection of data [5].

A related concern is the problem of multiplicity, particularly when statistical inference is conducted in the frequentist paradigm. In general there is no indication as to the number of inferential analyses executed, and even when the number of analyses performed is known, either appropriate adjustments are not made to mitigate the inflation of type I error probabilities, or most of the available adjustment procedures lack adequate power. Although, the implication of multiplicity on healthcare utilization is well recognized [6,7], the practice is rampant in the medical literature, and efforts to curb the problem have not been fully been successful or endorsed by authors and editorial boards.

In meta-analysis heterogeneity generally refers to the inconsistency of treatment effects across studies, and there are several proposed measures for its assessment. One popular measure is the so-called Cochran's Q test, which quantifies the deviation of each study from the pooled effect across study, weighted suitably. A limitation of this procedure is that its power is a function of the number of studies included [8,9]. A test related to the Q statistics and that describes the percentage of variation across studies that is due to heterogeneity rather than chance is the I^2. It is computed as $I^2 = 100\% \times (Q-df)/Q$ and generally is less sensitive to the number of studies considered [10]. When the effect measure is based on odds ratios and appropriate data is available, the Breslow-Day test is often used; however, this test is also highly sensitive to the sample size within each study and this limits its utility [11]. Thus, even in traditional meta-analysis setting, the available procedures are fraught with problems and practical restrictions, and their extension to CER situations is less obvious. It is therefore critical that heterogeneity test results be always considered vis-à-vis a a qualitative assessment of the combinability of studies. Although L'Abbé plot is commonly employed for this purpose in traditional systematic reviews, simple forest plots or descriptive statistics should be used to visually inspect any signal of presence or absence of consistency of results across subgroups or other strata of interest.

When heterogeneity is suspected in systematic reviews, the usual approach is to attempt to perform an analysis either building the heterogeneity into the model (e.g., random effects models) or to perform the analyses in homogenous subgroups [12,13]. Neither approach, however, is desirable. The former introduces problems of interpretability of the ensuing results, while the latter is subject to loss of power, multiplicity of testing, and bias resulting from the post-hoc nature of the analyses. An alternative strategy is to explain the cause of heterogeneity by including relevant covariates either at the patient or study level in a regression analysis. When data is available only at the study level, the method commonly referred to as meta-regression is employed in routine meta-analysis, with obvious extensions to indirect comparisons. While the method appears to be intuitively appealing, it is associated with basic conceptual problems. First, the post-hoc nature of the definition of the covariates may introduce bias. Second, since the analyst has to deal with only the available information, there is no certainty that all relevant covariates could be included in the model. Lastly, such an approach may suffer from the so-called ecological fallacy, i.e., association present at patient level may not be necessarily true at the study level. In fact, it is well known that a model that includes a covariate that is an aggregate of person-level characteristics rather than a study characteristic can produce biased results [14,15].

In CER, indirect and mixed treatment comparisons play an important role, in the absence of head-to-head comparative data [16,17]. As alluded to earlier, the extension of the standard statistical tests for homogeneity to such analyses is still not well developed. In practice, the tendency is to perform separate analyses in subgroups of interest, thereby resulting in inflated type I error rate, and loss of efficiency. An alternative approach, which has little theoretical justification, is to extrapolate the heterogeneity assessment results from the original head-to-head trials to the indirect comparisons. While this may provide supportive evidence to analyses performed using other procedures, it by no means provides definitive results. Given the inadequacy of the current approaches, there is a need to explore other options, including simulations and more complex hierarchical models that incorporate the heterogeneity term in the primary analyses designed for indirect or mixed treatment comparisons.

Approaches for a General Framework

The discussion in the preceding section elucidates the challenges and opportunities of assessing heterogeneity in the context of CER, and the importance of formulating a coherent strategy to address the issue to guide important healthcare decision making. In the following, we provide a few points that may serve as guidelines for practitioners of CER, including policymakers and healthcare providers to ensure proper assessment of the risks and benefits of treatments at the individual or subgroup level.

Pre-specification

To minimize the possibility of bias associated with the definition of objectives following inspection of the data, it is vitally important to specify in a study protocol the research hypothesis, including the variables of interest, the methods to be used, and any other aspects of the analysis, as well as potential adjustments for multiplicity. The practice of attempting to explain heterogeneity through the execution of numerous post hoc analyses is tantamount to data dredging, and cannot be a basis for making healthcare decisions in a CER framework.

Analytical strategy

Since there are no universally accepted or applicable statistical procedures to test for heterogeneity in CER settings, caution should be exercised in the use of results from tests that are designed to address narrow objectives in the classical meta-analysis models. When modified tests are employed in indirect and mixed treatment comparisons, there should be adequate justification for the validity of the underlying assumptions, including adequacy of power and appropriateness of adjustments for multiple testing.

Control for Potential Bias

Investigation of heterogeneity of treatment effects in CER or other systematic reviews inherently shares the limitations of observational studies, including potential bias through confounding by observed or unobserved variables. The analytical strategy should rule out the effects of such confounders before drawing conclusions about subgroup differences. See, e.g., [18] for an example of adjusted analysis in indirect comparisons.

Clinical significance of findings

When the data suggests statistical heterogeneity, the next step

should be to carefully assess whether the magnitude of the difference is clinically important enough to warrant different recommendations for different subgroups. This generally requires a careful clinical evaluation of the findings vis-à-vis the subgroups studied. In particular, if there is a well established minimally important difference (MID) for the parameter of interest, it should be noted that the MID established for the population may not be constant across subgroups.

Validation and sensitivity analysis

To ensure the robustness of results of heterogeneity analyses, it is essential to assess the internal and external validity of the results. The former may include analyses to assess the sensitivity of the results to departures from definition of clinical variables, inclusion/exclusion of studies and model assumptions. When the analyses involve indirect or mixed treatment comparisons, it may be worthwhile to inspect whether the subgroups results are replicated in the original RCTs. In addition, one should look for external evidence supporting the findings, and the consistency of the evidence with established clinical consensus.

Reporting of results

Finally, homogeneity results should be presented with transparency and fair balance. More specifically, there should be full disclosure about pres-specification, definition of clinical variables, number of subgroup analyses performed and the rationale for the statistical procedures employed. The limitations of the analyses should be prominently acknowledged, and the biological plausibility of the findings and their consistency or inconsistency with current clinical literature should be highlighted.

Concluding Remarks

The primary goal of CER is to generate evidence to help informed decision making on what treatment is best for a specific situation. This is predicated on the fact that not all treatments may work for everyone, and that in fact certain treatments may have more benefits or risks for some patients than others. Thus, the effective assessment of heterogeneity in CER is critical for reliable decision making by diverse stakeholders, including clinicians, patients, policymakers and other healthcare providers.

In this review paper, it is stressed that a valid assessment of heterogeneity requires verifiable pre-specification of criteria, methods and general investigational approaches, as well as adequate data. Factors related to clinical heterogeneity should be defined based on underlying science, with input from experts and a review of the relevant literature. It is recognized that there are gaps in the development of analytical procedures to address heterogeneity in the context of CER. However, the usual techniques used in traditional systematic review may still prove helpful, provided their limitations are understood. This is particularly the case when conducting indirect and mixed treatment comparisons, where it is essential to verify consistency of heterogeneity findings with those observed in the original RCTs.

Arguably, any investigation of heterogeneity in CER using secondary data has the known limitations of non-randomized studies. It is therefore quite prudent to take all necessary steps that one would take in the analysis and reporting of observational studies. In particular, the results of such studies should be reported with fair balance, transparency and a discussion of the study limitations. Until more rigorous analytical tools are available to address heterogeneity in CER, the suggestions put forth in this paper will serve to underscore the underlying issues and to mitigate potential adverse impacts ensuing from lack of awareness of the pitfalls.

References

1. Schilsky RL (2010) Personalized medicine in oncology: the future is now. Nature Reviews Drug Discovery 9: 363-366.
2. West SL, Gartlehner G, Mansfield AJ, Poole C, Tant E, et al. (2010) Comparative Effectiveness Review Methods: Clinical Heterogeneity. Agency for Healthcare Research and Quality.
3. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM (2007) Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. N Engl J Med 357: 2189-2194.
4. Oxman AD, Guyatt GH (1992) A consumer's guide to subgroup analyses. Ann Intern Med 116: 78-84.
5. Smith GD, Ebrahim S (2002) Data dredging, bias, or confounding. BMJ 325: 1437-1438.
6. Ioannidis JPA (2005) Why most published research findings are false. PLoS Medicine 2: e124.
7. Lord SJ, Gebski VJ, Keech AC (2004) Multiple analyses in clinical trials: sound science or data dredging? The Medical Journal of Australia 181: 452-454.
8. Gavaghan DJ, Moore AR, McQay HJ (2000) An evaluation of homogeneity tests in meta-analysis in pain using simulations of patient data. Pain 85: 415-424.
9. Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. British Medical Journal 327.
10. Higgins JPT, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21: 1539-1558.
11. Breslow NE, Day NE (1994) Statistical Methods in Cancer Research.Volume II--The Design and Analysis of Cohort Studies. IARC Sci Publ 1-406.
12. Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. London: Academic Press.
13. Petitti DB (2001) Approaches to heterogeneity in meta-analysis. Statistics in Medicine 20: 3625-3633.
14. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J (2004) Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol 57: 683-697.
15. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman KA (2002) Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Statistics in Medicine 21: 371-387.
16. Bucher HC, Guyatt GH, Griffith LE, Walter SD (1997) The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 50: 683–691.
17. Song F, Altman DG, Glenny AM, Deeks JJ (2003) Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published metaanalyses. BMJ 326: 472.
18. Song F, Harvey I, Lilford R (2008) Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. J Clin Epidemiol 61: 455–463.