# The Good, the Bad, and the Different: A Primer on Aspects of Heterogeneity of Treatment Effects

Daniel C. Malone, RPh, PhD; Lisa E. Hines, PharmD; and Jennifer S. Graff, PharmD


## SUMMARY

**The concept of heterogeneity is concerned with understanding differences within and across patients and studies. Heterogeneity of treatment effects is nonrandom variability in response to treatment and includes both benefits and harms. Because not all patients respond the same way, treatment decisions applied in a "one size fits all" fashion based on the average response observed in clinical trials may lead to suboptimal outcomes for some patients. Variation in outcomes among patients may be caused by observable and nonobservable factors. Changes in patients' health status over time can contribute to variability among patients. Assuming that the results from clinical trials are homogeneous across patients may fail to take into account clinically significant variability where some patients may receive benefit and others harm. Subgroup analyses and prediction models are 2 tools to explain variability observed within a study. Evidence synthesis with meta-analysis can provide useful information on the overall effectiveness and response among groups of patients undersampled in individual studies. Yet caution is warranted if the meta-analysis is missing studies or the individual studies comprising the meta-analysis are inherently different.**

**For those making clinical, coverage, and reimbursement decisions at a population level, such as clinicians and pharmacy and therapeutics committee members, understanding the variation among patients, among subpopulations or populations of patients, among clinical studies, or within a meta-analysis is important to ensuring optimal patient outcomes. This article presents a variety of tools and resources to aid decision makers as they evaluate the literature to determine when clinically relevant differences exist.**


*J Manag Care Pharm.* 2014;20(6):555-63



In health care, patients commonly report that "this medication doesn't work for me," or physicians switch therapies because of nonresponse. Basic pharmacology and clinical studies typically evaluate a medication's efficacy in a controlled setting, among patients with a narrowly circumscribed set of inclusion and exclusion criteria, and tend to report the "average" or mean treatment response. Yet, this average does not ensure that all patients will have the identical response to therapy. In fact, there is a range of both therapeutic benefits and harms, and no 2 patients will respond exactly the same way. As an example, pharmacology studies have documented up to 100-fold differences in drug metabolism.[1] Measures of what works best on average, such as mean treatment effects, may be misleading. Relying solely on the mean estimate assumes the underlying response to treatment (both benefits and harms) is consistent among all patients. To put this idea in perspective, it is similar to the ice fisherman assuming that a frozen lake with an *average* depth of 1 foot of ice will have the same depth of ice across the entire lake. However, many fishermen have broken through ice, with life-threatening consequences, because of this false assumption.

The current focus on large population-based studies of what works in health care is, in part, driven by the funding of comparative effectiveness research (CER). The premise of CER is to assist health care providers, payers, and consumers in making treatment decisions by comparing therapeutic alternatives. However, one challenge with population studies is that they focus on reporting central tendency (e.g., population mean or median). Yet, this type of focus represents only part of the story. The other part, and perhaps the most important, is dispersion of the data around the mean or median.

Recognizing this dispersion allows one to shift from asking "how alike" to "how different" are these subjects or studies. *Heterogeneity* addresses the "how different" and is defined by the Cochrane Collaboration as "variation in, or diversity of, participants, interventions, and measurement of outcomes across a set of studies, or the variation in internal validity of those studies."[2] Referring specifically to *statistical heterogeneity*, the term is used to describe the degree of variation in the effect estimates from a set of studies, and thus, the term indicates the presence of variability among studies beyond the amount expected solely caused by chance.[2]

Recently, Varadhan et al. (2013) defined heterogeneity of treatment effects as "nonrandom explainable variability in the direction and magnitude of individual treatment effects, including both beneficial and adverse effects."[3] Heterogeneity of treatment effects is a principal component of patient-centered outcomes research. It is therefore increasingly important for decision makers, such as pharmacy and therapeutics (P&T) committees, to understand the variation and diversity in treatment response as decisions shift from optimizing care for populations of patients to individual patients. In general, the optimal treatment choice for a population strikes a balance between the associated benefits and risks. However, not all patients respond the same way; consequently, treatment decisions applied in a "one size fits all" fashion, based on the average response, may lead to suboptimal clinical, humanistic, and economic outcomes for patients, providers, and health care

decision makers. This central tendency approach may result in substantial benefits for some patients, little benefit for many, and harm for a few patients.[4]

There are several sources of heterogeneity of treatment effects. In some cases, heterogeneity or differences in response to treatment clearly exist because of variation in baseline clinical characteristics, concomitant medications, or care settings such as primary versus specialist care or centers with a high volume versus low volume for particular procedures. In other cases, demographic differences, such as age, gender, or time since diagnosis, may exist between the study groups and a health plan's population, affecting the generalizability of the study results to a plan's membership. In addition, trials may differ in their reported results, making synthesis across studies more challenging. When comparing treatments, it is essential to consider the significance or relevance of the sources of differing treatment effects. Are the differences clinically meaningful, or are the differences due to variation in the trial design, population, endpoints, or comparators? Although doing so is somewhat difficult, researchers should examine why differences exist when interpreting the results of multiple studies.

Recently, the importance of considering heterogeneity in the context of clinical decision making has intensified, yet little guidance exists on how to assess or evaluate the statistical and clinical relevance of these differences.[5-7] This commentary presents a variety of tools and resources for clinicians and decision makers, such as P&T committee members, to evaluate the literature and help the reader understand when clinically relevant differences exist between individuals, populations, or clinical trials.

## Differences Among Individual Patients

Each patient is unique and responds differently than the "average" patient from a study population for a variety of reasons.[4,8,9] For example, differences in treatment response may result from factors such as age, sex, ethnicity, phenotype, and genotype. Disease-related risk factors, including severity, comorbidities, and physiologic status, may also significantly impact therapeutic response as well as propensity for harm. Other factors such as lifestyle (e.g., level of activity, alcohol consumption, dietary intake), treatment setting (e.g., acute care vs. outpatient clinics), and provider characteristics (e.g., provider quality, specialist vs. primary care) may further impact clinical outcomes. While many factors influencing treatment response are observable and identifiable via administrative data or other electronic data systems, some factors such as lifestyle, adherence, or disease severity are either nonobservable or not systematically captured, and other factors remain unknown given our current scientific understanding.

Selecting the optimal treatments for preventing stroke in patients with atrial fibrillation provides an example of differences among patients. We highlight these differences by describing 2 hypothetical patients. Patient 1 is a male aged 52 years with diabetes and hypertension, while Patient 2 is a relatively healthy female aged 52 years without comorbidities. Both patients are newly diagnosed with atrial fibrillation. Based on study results, warfarin is more effective than aspirin at reducing the risk of stroke.[10] Yet, patients on warfarin may be at increased risk of extracranial bleeding. For some patients, this risk of bleeding may be worth the benefit of stroke prevention. For other patients, the risks may outweigh the benefits. To identify where the differences lie, risks and benefits must be determined in each of these cases. CHADS2 is a summary measure, helpful in predicting the likelihood of stroke based on presence of relevant risk factors—**C**ongestive heart failure, **H**ypertension, **A**ge ≥ 75 years, **D**iabetes, and **S**troke (previous history of stroke or ischemic event).[11] The higher the CHADS2 score, the greater the risk for stroke. Patient 1 has a CHADS2 score of 2 and therefore may accept a higher risk of bleeding associated with anticoagulation therapy compared with Patient 2, who has a CHADS2 score of 0. Other factors may influence therapeutic choices, including ability to monitor anticoagulation status effectively, patient adherence, and follow-up visits. As a consequence of having a CHADS2 score of 2, Patient 1 is a candidate for anticoagulation therapy and would benefit the most from dabigatran compared with other oral anticoagulants, according to guidelines from You et al. (2012).[10] On the other hand, Patient 2 has a CHADS2 score of 0, suggesting that the bleeding risks of anticoagulation therapy outweigh the benefits.[10]

Some of the relevant risk factors for clinical outcomes associated with atrial fibrillation are readily identifiable from administrative data, such as age or use of antidiabetic medications to help identify patients with diabetes mellitus. Other risk factors, such as time from myocardial infarction (MI) or number of recent exacerbations of congestive heart failure, require more effort and sophisticated analyses for health plans to investigate.

Differences among patients may occur as a result of differing risk and optimal treatments at a single point in time, and these may change for patients over time. Changes in health status, such as pregnancy or development of new comorbidities, may impact the optimal treatment choice. Using the atrial fibrillation example, when Patient 1 experiences a MI, treatment modifications are required to address the significant increase in risk of stroke post-MI. The MI may result in physiologic changes, such as reduced renal function, or new medications may result in drug-drug interactions, having an impact on therapeutic selection. Lastly, a patient's risk tolerance, or receptivity to being exposed to treatment risks, may also change over time, especially after therapeutic failures or as the underlying condition and treatment-related consequences are understood and accepted.

## Differences Within Studies

The intent of study-related inclusion and exclusion criteria is to define the population of interest and, simultaneously, reduce the variability across eligible patients. Despite these criteria, wide variation may still exist within a study or when extrapolating study results to a plan's population based on factors such as degree of disease severity, ages of patients enrolled, comorbid diseases, genotype, phenotype, and renal and hepatic status. Consequently, the relationship between these factors must be evaluated to understand how they may result in disparate or inconsistent findings among patients within a study.

Heterogeneity may exist within a given study when there are both treatment responders and nonresponders or when the level of benefit varies widely. Examining underlying baseline characteristics or risk profiles of subgroups of study subjects (e.g., high risk vs. low risk) may help reveal reasons for such differences in treatment effects. This article provides an overview of the commonly used approaches to assist readers in evaluating the clinical importance and relevance of differences between patient subgroups. The authors encourage readers to access other resources on more advanced methodological and analytical approaches.[3,12,13]

Within a given treated population, multiple subgroups may exist that have differing responses to treatment. One way to identify who may experience the greatest benefit or risk is to identify subgroups of patients with common observable characteristics. For example, characteristics such as increased age, alcohol intake, hypertension, and presence of atrial fibrillation are associated with a greater risk for stroke. Different subpopulations of patients possessing the above-mentioned characteristics may exist within a study (e.g., high-risk vs. low-risk patients) or among studies (e.g., different inclusion or exclusion criteria). For clinicians and decision makers, study results may not be generalizable to a real-world patient population with different age distributions, comorbidities, or treatment adherence.

Subgroups are useful in identifying not only those with the largest potential benefit, but also individuals who may be at increased risk of harm. A recent article in the *New England Journal of Medicine* examining rivaroxaban in the treatment of atrial fibrillation is a good example of how to prespecify subgroups in the analysis plan and report the findings.[14] The appendices, available in the online supplemental materials, provide details on response-to-therapy by varying comorbidity status and concomitant medications.

While subgroup analyses may aid in identifying patients more likely to benefit or experience harm from treatment, there are important considerations to keep in mind when interpreting the findings. First, subgroup analyses require sufficient sample size to avoid committing a Type II error, stating that no differences exist when, in fact, there are differences, but the sample size was too small to detect them. Second, subgroup risk factors are not necessarily independent from each other and may occur simultaneously. For example, age and presence of hypertension are both positively correlated with stroke. Third, subgroups should be identified *a priori* (before analyzing the data) to avoid making a Type I error and stating that a difference exists when, in fact, there are no differences. Examples of such an error occur when the statistical differences are spurious or by chance as a result of conducting multiple analyses on the same data. Furthermore, it is important to adjust for multiple comparisons and note when analyses are hypothesis generating or confirmatory.[15] Despite these caveats, appropriate subgroup analyses offer a critical first step in predicting if or when patients respond.

Kent et al. (2010) proposed a number of recommendations for conducting and reporting subgroup analysis.[15] These recommendations were adapted and are presented in a checklist format (Table 1) to assist managed care professionals in interpreting subgroup findings from clinical trials. This checklist highlights aspects for readers to consider as they evaluate the literature to assess whether differences in treatment response were appropriately analyzed, reported, and interpreted. Important considerations include demonstrating variation in risk with risk prediction models, prespecifying and justifying primary subgroup analyses, clearly identifying exploratory analyses, and using appropriate statistical methods to test for heterogeneity of treatment effects.[15] When treatment differences are clinically relevant and conducted via confirmatory methods, subgroups should be considered in clinical care algorithms and coverage policies to ensure that both patient and population health are optimized.

## Prediction Approaches

While subgroup analyses can be useful, in some situations a single factor may not sufficiently explain differences in observed treatment effect. Prediction models permit the simultaneous evaluation of multiple factors using regression models. This analytic approach takes multiple factors into account simultaneously and allows for interactions between the risk factors and thus may be more powerful than subgroup analyses. For many health conditions, validated risk-adjustment prediction models exist to predict mortality, hospitalization, and most commonly billed health care costs. Prediction models allow the analyst to generate a summary score, many times summing the values for individual risk factors or using assigned or empirical weights. For example, the Framingham score predicts the risk of a first major coronary event[16]; the Gail model predicts incident breast cancer[17]; the APACHE II (Acute Physiology and Chronic Health Evaluation II) score predicts intensive care unit survival[18]; and the Charlson Comorbidity Index uses administrative data to predict mortality.[19] Numerous other models exist to predict clinical outcomes for a wide range of common diseases.[15]

There are a number of limitations to using prediction models. First, they always incorporate some degree of error. The extent of the inaccuracy of a model is a function of how closely

**TABLE 1** Checklist for Reporting on Subgroup Analyses and Heterogeneity of Treatment Effects

| Questions | Your Answers |
|---|---|
| 1. Do the authors demonstrate variation in risk using a risk prediction model or index in (a) overall study population and (b) separate treatment arms?<br>• Reports how predicted risk (or risk score) varies (a) within the study population and (b) by treatment arm.<br>• Displays variance of study population graphically (e.g., histograms or box and whiskers plots) or reports the mean, standard deviation, median, and interquartile ranges. | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |
| 2. Are primary subgroup analyses risk-stratified with relative and absolute risk reductions?<br>• Risk prediction model is prespecified (i.e., fully specified before any analysis of treatment effect has begun) and preferably externally developed.<br>• Reports both absolute and relative risk reductions. | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |
| 3. Are additional primary subgroup analyses prespecified and limited to patient attributes with strong a priori justification?<br>• Justifies all primary subgroup analyses based upon strong pathophysiological or empirical evidence that such factors influence treatment effects. | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |
| 4. Are secondary subgroup analyses reported separately from primary subgroup comparisons?<br>• Clearly labels secondary subgroup analyses as exploratory (i.e., potentially useful for hypothesis generation and informing future research, but having little or no immediate relevance to patient care). | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |
| 5. Do the authors (a) report all subgroup analyses conducted, (b) use appropriate statistical methods to test heterogeneity of treatment effects (e.g., interaction terms), and (c) avoid overinterpretation?<br>• Limits comparisons to statistical significance of treatment heterogeneity between subgroups using interaction terms.[a] | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |

*Adapted from Kent et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal.*[15]
[a]*Testing for the significance of a treatment effect within a subgroup is inappropriate because of poor statistical power.*

the patients included in the prediction sample match those individuals who were used to develop the prediction model. Large sample sizes are required to help offset the potential inaccuracies. Second, unlike subgroup analyses, prediction models can account for risk factors that are not independent of other risk factors by including statistical interaction terms. In many situations, the assumption that risk factors are independent (i.e., no interaction is present—meaning that the results are not affected by the values of 2 independent variables) is violated because the presence of 1 condition is commonly accompanied by other comorbidities. For example, individuals with type 2 diabetes may also have hypertension.[20] Third, summary scores or prediction models are typically limited to predicting a single outcome for certain types of patients. For example, the CHADS2 score described earlier only predicts the risk of stroke, while the HEMORR2HAGES,[21] HAS-BLED,[22] and ATRIA[23] models only predict the risk of bleeding, despite the relevance of both stroke and bleeding outcomes to atrial fibrillation. Some risk models, such as the UKPDS (United Kingdom Prospective Diabetes Study) risk engine, predict the risk of cardiovascular complications in type 2 diabetes, yet few models exist to predict cardiovascular risk for patients with type 1 diabetes.[24] Fourth, the number of factors incorporated into the prediction model must be considered. Some summary scores are derived from many characteristics, but not all characteristics are consistently and reliably captured in routine clinical care (e.g., smoking). Thus, use of a risk tool may be precluded because of lack of sufficient data to populate the model. Despite these limitations, summary scores and predication models are useful for understanding differences observed across patients within a study. They also are a useful measure for evaluating differences among studies (e.g., low-risk patients vs. high-risk patients). The authors encourage readers to consult Iezzoni's book, *Risk Adjustment for Measuring Health Care Outcomes*,[25] to learn more about risk prediction model development and use; the book contains many examples of how to use summary scores as well as an in-depth discussion of their limitations.

## Differences Among Studies

Systematic reviews of studies, with or without meta-analysis, are increasingly important as more evidence is published and with the recent emphasis on CER. Assessing differences among studies is essential for appropriate conduct and interpretation of systematic reviews. For example, researchers should determine whether or not they can reasonably combine the studies into a quantitative meta-analysis. Assessing differences among studies is essential as well for clinicians and decision makers evaluating the quality of the systematic review. We address these issues and identify tools to aid decision makers in the following section.

There are several reasons for conducting a systematic review with meta-analysis, also known as quantitative evidence syn-

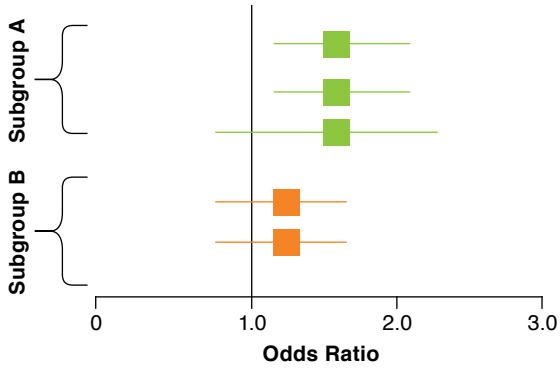
thesis. First, a systematic review with meta-analysis provides a more robust estimate of a treatment effect (how well a medication works) because it includes multiple studies, resulting in a larger overall sample size and greater statistical power. Numerous small studies may provide inconclusive evidence, but their combined evidence may show a clear treatment advantage. For example, early studies evaluating the use of aspirin in preventing death post-MI involved less than 2,000 patients.[26,27] Taken individually, the confidence intervals were not statistically different from the placebo or no treatment groups. This lack of significance was caused by the studies having relatively small sample sizes that were evaluating a rare outcome (death). However, when the evidence was combined, aspirin clearly had a significant impact on survival.[28] This finding was confirmed in the ISIS-2 study, involving more than 10,000 patients.[29] Combining data across studies may lead to statistical significance—but one should always keep in mind that statistical significance is not necessarily clinically meaningful. In the example given, death is clearly a clinically significant endpoint, but not all meta-analyses will have such an important outcome. Second, systematic review with meta-analysis can account for moderator variables, such as improvements in treatments over time. Third, such review may permit investigation of new questions not yet addressed by the individual studies. Finally, systematic review with meta-analysis can also facilitate clinical guideline development and coverage and reimbursement decisions. Information derived from such evidence synthesis is often more robust and reduces uncertainty associated with a single study or less rigorous attempts to decipher the results from many studies.

Regarding systematic reviews including a meta-analysis, there are 3 types of heterogeneity to consider: clinical heterogeneity (variability in patients, interventions, and outcomes studied); methodological heterogeneity (variability in study design and risk of bias); and statistical heterogeneity.[13] The former 2 contribute to the latter. For example, studies evaluating aspirin use in preventing death post-MI differed in the doses studied, length of long-term follow-up, and the ages of subjects included in the studies. This variability may or may not be important to interpreting the findings. Figure 1 shows 3 different scenarios with 2 subgroups (A and B) each: (a) clinical heterogeneity is present but has minimal impact on the treatment effect; (b) clinical heterogeneity is present, but the relevance of the impact has to be determined on clinical grounds; and (c) clinical heterogeneity is present and leads to an impact on the treatment effect.[13]

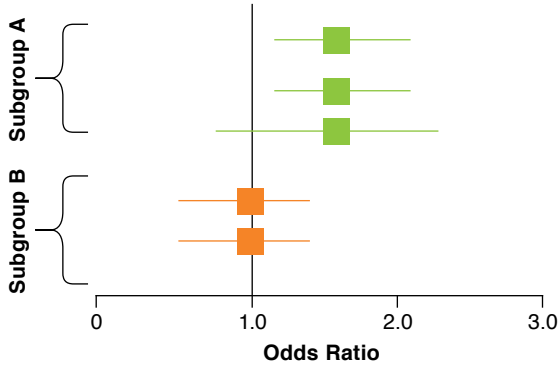
When there are differences in the reported effects that extend beyond what is expected from random error alone, statistical heterogeneity is present. The individual studies included in a systematic review may differ from the population

**FIGURE 1** Detecting Clinical Heterogeneity

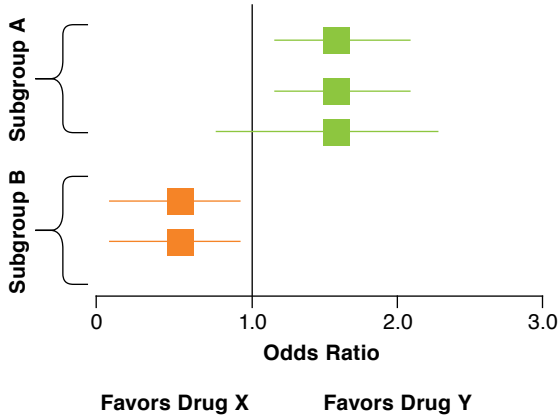
**A**. Clinical heterogeneity is present but has a minimal impact on the treatment effect because the point estimates and corresponding 95% confidence intervals for the subgroups overlap.




**B**. Clinical heterogeneity is present, but the relevance has to be determined on clinical grounds because subgroup B has shifted to the left, though the magnitude of the shift still shows overlap of the confidence intervals.

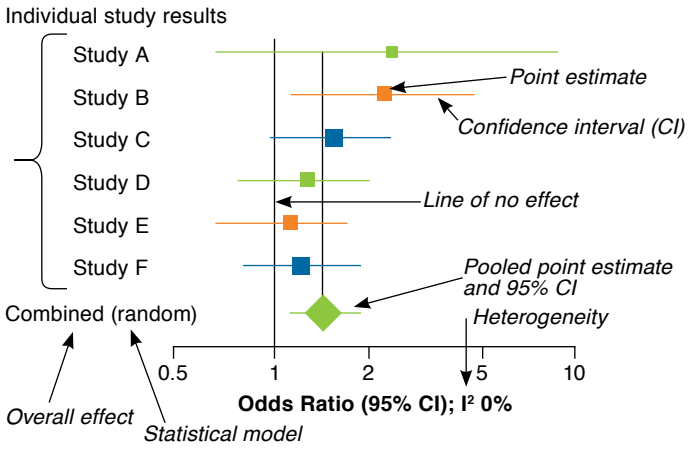



**C**. Clinical heterogeneity is present and leads to a relevant impact on the treatment effect because subgroup B has an effect that is opposite subgroup A, lying to the left of the odds ratio value of 1.0.




*Source: West SL, Gartlehner G, Mansfield AJ, et al. Comparative effectiveness review methods: clinical heterogeneity.*[13]

**FIGURE 2** Interpreting a Forest Plot




*Adapted from Gartlehner et al. Second-generation antidepressants in the pharmacologic treatment of adult depression: an update of the 2007 comparative effectiveness review.*[30]

studied (e.g., primary vs. secondary fracture prevention); time of study conduct (e.g., when background standard of care may be different); study location (e.g., United States vs. global trial); doses (e.g., therapeutic vs. sub- or supratherapeutic); or length of follow-up (e.g., 18 months vs. 2 years). These may result in clinical, methodological, and/or statistical heterogeneity.

The methods used for meta-analysis are beyond the scope of this primer; however, an overview of how to interpret the findings, focusing specifically on heterogeneity and graphical results, is included here. A "typical" forest plot from a meta-analysis is displayed in Figure 2.[30] The horizontal axis is the scale of measurement for the meta-analysis. For dichotomous outcomes (e.g., presence or absence of an MI) the horizontal axis is commonly represented as the odds ratio. Values greater than 1 indicate increasing odds of experiencing the event, and values less than 1 indicate decreasing odds of the event relative to a comparison group. Individual studies are listed on the left side with overall results shown at the bottom. Studies can be listed in chronological order or by some other attribute, such as first author's last name. To the right of the study identifier is a box that represents the point estimate of intervention effect relative to the comparison group. For each study, the box size gives a sense of the total number of subjects (including both intervention and comparison groups). The "whiskers" extending from either side of the box typically represent the 95% confidence intervals for that study. In general, but not always, smaller studies will have wider whiskers (larger confidence intervals) than larger studies. Typically, at the bottom of the forest plot is a diamond with whiskers that represents the summary of the effect across all studies. The width of the 95% confidence interval for the summary measure is usually narrower than the individual studies above because more observations are reflected in the estimate; however, the confidence interval may be wider if there are conflicting findings. Also, the authors of the analysis should report the degree of statistical heterogeneity present in the analysis.

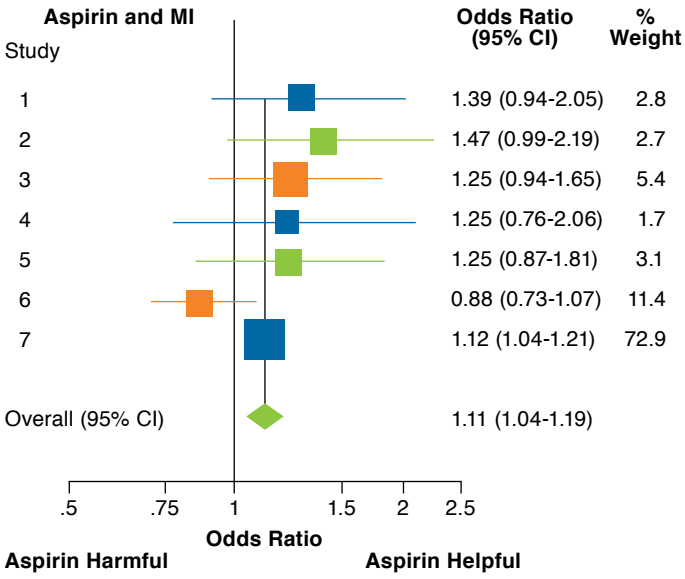
Cochran's Q and the $I^2$ statistic are measures of statistical heterogeneity that are often reported with systematic reviews. Cochran's Q is calculated by finding the difference between each study's effect size and the overall effect size (e.g., the magnitude of benefit in the study vs. the full meta-analysis), squaring each result, and summing across all studies, then dividing by the sum of the weights (1/variance). The resulting value is similar to a chi-square statistic, and the number of observation values allowed to change (e.g., degrees of freedom) is equal to the number of studies minus 1. However, interpreting Cochran's Q is different because, unlike with other statistical tests, it is *desirable* to have a *P* value greater than the Type I error rate, typically set to 0.05. A small Q value (closer to 1; higher *P* value) suggests the lack of statistical heterogeneity, an indication that combining studies is statistically appropriate. However, Cochran's Q can be misleading because it has less power to detect the presence of heterogeneity with fewer studies and excessive power to detect trivial heterogeneity with a large number of studies.

The $I^2$ statistic is now considered the primary approach for evaluating statistical heterogeneity. $I^2$ is calculated by subtracting the degrees of freedom from Cochran's Q and then dividing by Cochran's Q and multiplying the result by 100%. The $I^2$ statistic quantifies heterogeneity; it describes the percentage of the variability in the effect estimate that is caused by heterogeneity rather than chance.[31] Most analysts use $I^2$ values greater than 50% as the threshold to indicate heterogeneity. $I^2$ is not subject to the same bias when evaluating fewer studies that affects the Cochran's Q statistic.

When no heterogeneity exists, the preferred approach is to use the "fixed" effects model to report the findings of a meta-analysis. This approach assumes that all studies are measuring the same construct using similar patients and interventions. With a fixed effects approach, any error observed between an individual study and the overall mean effect is presumed as random.

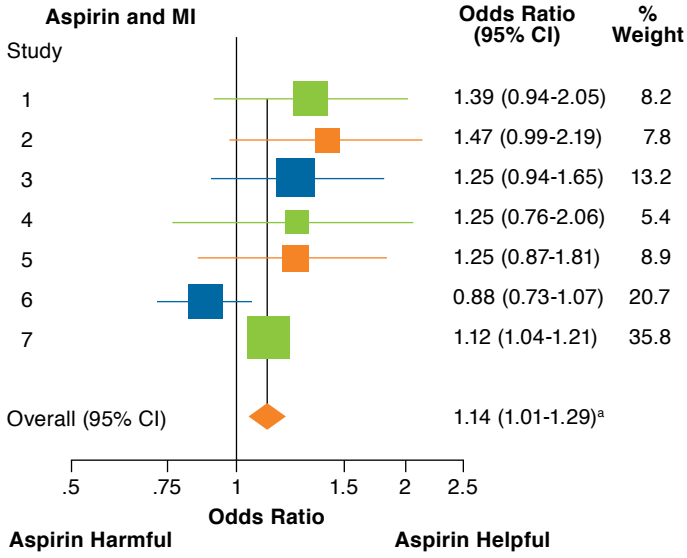
However, when statistical heterogeneity is present, the assumption that all studies are the same is violated. In this case, the analyst should use a "random" effects model that accounts for 2 sources of error: (a) between study variation, and (b) within study variation. The random effects approach relaxes the assumption that all studies are the same. Figures 3 and 4, respectively, display the fixed and random effects analyses (survival) of aspirin following a MI. Note that the overall

**FIGURE 3** Fixed Effects Forest Plot




*CI = confidence interval; MI = myocardial infarction.*

**FIGURE 4** Random Effects Forest Plot




*[a]Notice wider confidence interval vs. fixed effects (1.04-1.19).*
*CI = confidence interval; MI = myocardial infarction.*

point estimates for the 2 methods are similar but not exact. The point estimate for the random effects is slightly higher (1.14) as compared with the fixed effect (1.11). Also, note that the width of the confidence interval is greater for the random effects approach (95% CI = 1.01-1.29), which is more conservative than the fixed effects model (95% CI = 1.04-1.19). When digging a little deeper into the studies included in the analysis, one would find different doses of aspirin used, different ages of patients included in the study, and differing lengths of follow-up. All of these factors would contribute more variance to the analysis. However, because the random effects approach takes both within-study variance and between-study variance into account, the results are more robust than the findings from the fixed effect.

There are several tools available to evaluate the quality of systematic reviews, with or without meta-anlayses.[32-34] Table 2 provides a 4-item tool for assessing heterogeneity across studies. This tool was adapted from an 11-item instrument developed and validated for assessing the methodological quality of systematic reviews, called the "assessment of multiple systematic reviews" or AMSTAR.[32] To assess heterogeneity across studies, items address (a) graphical representation of studies to visualize heterogeneity, (b) statistical heterogeneity and appropriate methods to combine the studies, (c) assessment of publication bias, and (d) description of included study characteristics.

## Publication Bias

A common criticism of evidence synthesis relates to the issue of not including unpublished studies that show inverse effects (from the anticipated) or no effect at all.[35] Publication bias may take many forms, such as publishing only positive studies or studies being published in English. All systematic reviewers need to be concerned about missing data because the results can be biased to the degree that missing information is not included. Thus, heterogeneity may exist with respect to what literature is available, and there is substantial evidence that publication bias exists.[36] To help address this issue, the clinicaltrials.gov website provides the analyst and reader with information regarding planned (and perhaps conducted) interventional studies. However, not all trials may be listed in clinicaltrials.gov.

Researchers have long recognized the issue of publication bias and have created several approaches to identify its possible presence. If studies are missing at random, then the effect of publication bias will be less. However, it is important that authors of meta-analyses evaluate the studies identified for possible publication bias. Examining consistency in the forest plot and creating a funnel plot are 2 visual approaches for evaluating publication bias.[37] A funnel plot is an upside-down funnel that has larger studies near the top and smaller studies towards the bottom.[38] The effect on the x-axis and 1/variance on the y-axis are used to create a scatterplot of the studies included in a meta-analysis. A plot that has symmetrical distribution of studies suggests no publication bias. On the other hand, an asymmetrical plot, especially with small studies, is more suggestive that publication bias exists.

**TABLE 2** Checklist to Assess Heterogeneity Across Studies

| Questions | Your Answers |
|---|---|
| 1. Were the studies presented graphically to allow readers to visualize heterogeneity among the study results?<br>• For example, a forest plot showing each study's point estimate and the pooled point estimate, with confidence intervals.<br>• For the forest plot, do the studies look the same (eyeball test) and do the confidence intervals overlap? | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |
| 2. Were the methods used to combine the findings of studies appropriate?<br>• Tested pooled results to ensure the studies were combinable, to assess their homogeneity (i.e., chi-squared test for homogeneity, $I^2$).<br>• Uses random effects model if heterogeneity exists, and/or justifies the clinical appropriateness of combining (i.e., is it sensible to combine?). | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |
| 3. Was the likelihood of publication bias assessed?<br>• Includes a combination of graphical aids (e.g., funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test). | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |
| 4. Were the characteristics of the included studies provided?<br>• Aggregates data from the original studies (e.g., table) on the participants, interventions, and outcomes.<br>• Analyzes and reports the ranges of characteristics in all studies (e.g., age, race, sex, relevant socioeconomic data, disease status, duration, severity, comorbidities). | ❑ Yes<br>❑ No<br>❑ Can't evaluate<br>❑ Not applicable |

*Adapted from Shea et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews.*[32]

There are statistical methods that can assist in quantifying the degree of publication bias. One of these tests, Egger's, conducts a linear regression of the effect against the standard error, weighted by the inverse of the variance.[39] A straight-line relationship between the intervention effect and the standard error that is vertical indicates no bias.

## Discussion

The growing interest in CER will expand the availability of new information, comparing treatment alternatives in real-world patients. The goal is that this evidence will improve the health of both individual patients and populations of patients. Not all patients respond in the same way because of different clinical characteristics or preferences; therefore, practitioners should understand that a treatment that works on "average" for a population of patients may have less-than-optimal results for individuals or subpopulations of patients.

When reviewing literature for a P&T committee meeting, clinicians should consider that individual patient variability, variability within the populations studied (e.g., with or without particular risk factors), and variability between clinical studies may result in clinically meaningful differences in treatment response. As the number of CER studies increases, health plans must recognize that what works best for most plan members may provide suboptimal treatment for an individual member. Therefore, as richer clinical information (such as test results or clinical severity) becomes available through electronic medical records, more sophisticated benefit designs may help guide patients to their optimal treatment through logistical methods (e.g., utilization management techniques) or financial incentives (e.g., tiers). For example, differences in treatment response identified in the literature may guide clinical pathways and coverage policies for subpopulations of patients. Referring back to the atrial fibrillation and risk of stroke scenario, patients who develop diabetes may require different care management based on a higher CHADS2 risk score.

Typically, P&T committees can more efficiently use findings from rigorous systematic reviews rather than conduct such analyses themselves. However, as decision makers rely more heavily on systematic reviews to evaluate comparative effectiveness, the potential for inadequate evaluation of differences between studies may lead to drawing inaccurate conclusions. When summarizing clinical trial results among numerous studies or critically evaluating systematic reviews, it is imperative to evaluate for heterogeneity between the studies to avoid inappropriate comparisons.

## Conclusion

An understanding of the sources of heterogeneity and how to recognize and evaluate these differences will assist pharmacy and medical managers in assessing the relevance and impact of differences between individuals, populations, and clinical studies. Over time, our understanding of biological differences (pharmacogenetic, pharmacodynamic, and pharmacokinetic) will evolve, and the methods used to predict response to treatment will improve, leading to better decision making. In the interim, gaining a better understanding and appreciation of the differences among patients, subpopulations, and studies will serve to optimize health for the population as well as for individual patients.

## Authors

*DANIEL C. MALONE, RPh, PhD, is Professor, Department of Pharmacy Practice and Science, and LISA E. HINES, PharmD, is Research Clinical Pharmacist, College of Pharmacy, The University of Arizona, Tucson, Arizona. JENNIFER S. GRAFF, PharmD, is Director, Comparative Effectiveness Research, National Pharmaceutical Council, Washington, DC.*

*AUTHOR CORRESPONDENCE: Daniel C. Malone, RPh, PhD, Professor, Department of Pharmacy Practice and Science, College of Pharmacy, The University of Arizona, 1295 N. Martin, Dachman Hall B307F, Tucson, AZ 85721-0202. Tel.: 520.626.3532; Fax: 520.626.7355; E-mail: malone@pharmacy.arizona.edu.*

## DISCLOSURES

All authors participated in concept and design, writing, and revision of the manuscript. Graff is an employee of the National Pharmaceutical Council, which provided funding for this article. Malone served as a consultant to the National Pharmaceutical Council in 2011.

## REFERENCES

1. Lin JH, Lu AY. Interindividual variability in inhibition and induction of cytochrome P450 enzymes. *Annu Rev Pharmacol Toxicol.* 2001;41:535-67.

2. The Cochrane Collaboration. Glossary. Available at: http://www.cochrane.org/glossary. Accessed February 22, 2014.

3. Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *J Clin Epidemiol.* 2013;66(8):818-25.

4. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q.* 2004;82(4):661-87.

5. AMCP Format Executive Committee. A format for the submission of clinical and economic evidence of pharmaceuticals in support of formulary consideration. Version 3.1. December 2012. Available at: http://amcp.org/practice-resources/amcp-format-formulary-submisions.pdf.

6. Greenfield S, Kravitz R, Duan N, Kaplan SH. Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment. *Am J Med.* 2007;120(4 Suppl 1):S3-S9.

7. McLaughlin MJ, and HTE Policy Roundtable Panel. Healthcare policy implications of heterogeneity of treatment effects. *Am J Med.* 2007;120 (4 Suppl 1):S32-S35.

8. Eichler HG, Abadie E, Breckenridge A, et al. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nat Rev Drug Discov.* 2011;10(7):495-506.

9. Kaplan SH, Billimek J, Sorkin DH, Ngo-Metzger Q, Greenfield S. Who can respond to treatment? Identifying patient characteristics related to heterogeneity of treatment effects. *Med Care.* 2010;48(Suppl 6):S9-S16.

10. You JJ, Singer DE, Howard PA, et al. Antithrombotic therapy for atrial fibrillation: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest.* Feb 2012;141(Suppl 2):e531S-75S.

11. Gage BF, van Walraven C, Pearce L, et al. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. *Circulation.* 2004;110(16):2287-92.

12. Willke RJ, Zheng Z, Subedi P, Althin R, Mullins CD. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Med Res Methodol.* 2012;12:185.

13. West SL, Gartlehner G, Mansfield AJ, et al. Comparative effectiveness review methods: clinical heterogeneity. Methods Research Reports. September 2010. AHRQ Publication No. 10-EHC070-EF. Available at: http://effectivehealthcare.ahrq.gov/ehc/products/93/533/Clinical_Heterogeneity_Revised_Report.pdf. Accessed March 24, 2014.

14. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med.* 2011;365(10):883-91.

15. 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.

16. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998;97(18):1837-47.

17. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst.* 1989;81(24):1879-86.

18. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med.* 1985;13(10):818-29.

19. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40(5):373-83.

20. Long AN, Dagogo-Jack S. Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. *J Clin Hypertens (Greenwich).* 2011;13(4):244-51.

21. Gage BF, Yan Y, Milligan PE, et al. Clinical classification schemes for predicting hemorrhage: results from the National Registry of Atrial Fibrillation (NRAF). *Am Heart J.* 2006;151(3):713-19.

22. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJ, Lip GY. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. *Chest.* 2010;138(5):1093-100.

23. Fang MC, Go AS, Chang Y, et al. A new risk scheme to predict warfarin-associated hemorrhage: the ATRIA (Anticoagulation and Risk Factors in Atrial Fibrillation) Study. *J Am Coll Cardiol.* 2011;58(4):395-401.

24. Zgibor JC, Piatt GA, Ruppert K, Orchard TJ, Roberts MS. Deficiencies of cardiovascular risk prediction models for type 1 diabetes. *Diabetes Care.* 2006;29(8):1860-65.

25. Iezzoni LI. *Risk Adjustment for Measuring Health Care Outcomes.* 4th ed. Arlington, VA: Health Administration Press, AUPHA Press; 2013.

26. The Coronary Drug Project Research Group. Aspirin in coronary heart disease. *J Chronic Dis.* 1976;29(10):625-42.

27. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA.* 1980;243(7):661-69.

28. Draper D, Gaver DP, Goel PK, et al. *Combining Information: Statistical Issues and Opportunities for Research.* Alexandria, VA: American Statistical Association; 1993.

29. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet.* 1988;2(8607):349-60.

30. Gartlehner G, Hansen RA, Morgan LC, et al. Second-generation antidepressants in the pharmacologic treatment of adult depression: an update of the 2007 comparative effectiveness review. (Prepared by the RTI International–University of North Carolina Evidence-based Practice Center, Contract No. 290-2007-10056-I.) AHRQ Publication No. 12-EHC012-EF. Rockville, MD: Agency for Healthcare Research and Quality. December 2011. Available at: http://effectivehealthcare.ahrq.gov/ehc/products/210/863/CER46_Antidepressants-update_20111206.pdf. Accessed March 22, 2014.

31. Deeks JJ, Higgins JPT, Altman DG, eds. Chapter 9: analysing data and undertaking meta-analyses. In: Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions.* Version 5.1.0. Updated March 2011. Available at: http://www.cochrane-handbook.org. Accessed February 22, 2014.

32. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol.* 2007;7:10.

33. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol.* 2009;62(10):e1-34.

34. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7: a reviewer's checklist. *Med Decis Making.* 2013;33(5):679-91.

35. Light RJ, Pillemer DB. *Summing Up: The Science of Reviewing Research.* Cambridge, MA: Harvard University Press; 1984.

36. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev.* 2009;(1):MR000006.

37. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-analysis.* Chichester, UK: John Wiley & Sons; 2009.

38. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ.* 2011;343:d4002.

39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* 1997;315(7109):629-34.