A Consumer's Guide to Subgroup Analyses

Andrew D. Oxman, MD, and Gordon H. Guyatt, MD

■ The extent to which a clinician should believe and act on the results of subgroup analyses of data from randomized trials or meta-analyses is controversial. Guidelines are provided in this paper for making these decisions. The strength of inference regarding a proposed difference in treatment effect among subgroups is dependent on the magnitude of the difference, the statistical significance of the difference, whether the hypothesis preceded or followed the analysis, whether the subgroup analysis was one of a small number of hypotheses tested, whether the difference was suggested by comparisons within or between studies, the consistency of the difference, and the existence of indirect evidence that supports the difference. Application of these guidelines will assist clinicians in making decisions regarding whether to base a treatment decision on overall results or on the results of a subgroup analysis.

Annals of Internal Medicine. 1992;**116**:78-84.

From McMaster University Health Sciences Centre, Hamilton, Ontario. For current author addresses, see end of text.

Clinicians faced with a treatment decision about a particular patient are interested in the evidence that pertains most directly to that individual. Thus, it is frequently of interest to examine a particular category of participants in a clinical trial: for example, the women, those in a certain age group, or those with a specific pattern of disease. In observational studies, these examinations, or subgroup analyses, are routine. They are also frequently encountered in reports of clinical trials. In a survey of 45 clinical trials reported in three leading medical journals, Pocock and colleagues (1) found at least one subgroup analysis that compared the response to treatment in different categories of patients in 51% of the reports.

The results of subgroup analyses have had major effects, sometimes harmful, on treatment recommendations. For example, many patients with suspected myocardial infarction who could have benefited from thrombolytic therapy may not have received this treatment as a result of subgroup analyses based on the duration of symptoms before treatment (2) and the conclusion that streptokinase was only effective in patients treated within 6 hours after the onset of pain (3, 4). A later, larger trial showed that streptokinase was effective up to 24 hours after the onset of symptoms (5).

Conclusions based on subgroup analyses can have adverse consequences both when a particular category of patients is denied effective treatment (a "false-negative" conclusion), as in the above example, and when ineffective or even harmful treatment is given to a subgroup of patients (a "false-positive" conclusion). Because of these risks and their frequency, the appropriateness of drawing conclusions from subgroup analyses has been challenged (6, 7), and it has been argued that treatment recommendations based on subgroup analyses may do more harm than good. This hypothesis is currently being tested empirically by comparing treatment recommendations generated from early trials of new treatments based on subgroup analyses with treatment recommendations that would have been made had subgroup analyses been ignored, assessing "whether they lead to more patients receiving treatments that are worthwhile and fewer patients receiving treatments that are not." (Sackett DL. Personal communication.)

Although we agree that subgroup analyses are potentially misleading and that there is a tendency to overemphasize the results of subgroup analyses, in this paper we will present an alternative point of view. The essence of our argument is that subgroup analysis is *both* informative and potentially misleading. Rather than arguing for or against the merits of subgroup analysis, we will present guidelines in this article for deciding how believable the results of subgroup analyses are and, consequently, when to act on recommendations based on subgroup analyses and when to ignore them.

Our discussion will focus on randomized trials and meta-analyses of randomized trials (systematic overviews), although the same principles apply to any other research design. The assumption from which we start in this discussion is that the underlying design of the studies being examined is sound. For treatment trials, sound design involves elements of randomization, masking, completeness of follow-up, and other strategies for minimizing both random error and bias (8, 9). If the study is not sound, the overall conclusion is suspect, let alone conclusions based on subgroup analyses.

Even given a rigorous study design, the extent to which subgroup analyses should be done—or believed—is highly controversial. Although there are those who ignore scientific principles in the subgroup analyses they undertake and report, go on fishing expeditions, and indulge in data-dredging exercises (10, 11), there are also those who mix apples and oranges, drown in the data they pool (12), reach meaningless conclusions about "average" effects (13), and fail to detect clinically important effects because of the heterogeneity of their study groups (14). Although the debate between these two camps is entertaining and can lead to some useful insights, practical advice for assessing the strength of inferences based on subgroup analyses is also important. In providing such advice, we will build on criteria that have been suggested by other authors (15-18).

1. Is the magnitude of the difference clinically important?
2. Was the difference statistically significant?
3. Did the hypothesis precede rather than follow the analysis?
4. Was the subgroup analysis one of a small number of hypotheses tested?
5. Was the difference suggested by comparisons within rather than between studies?
6. Was the difference consistent across studies?
7. Is there indirect evidence that supports the hypothesized difference?

Our criteria are summarized in Table 1 and are described in detail below. An example of a hypothesized difference in subgroup response and the extent to which it meets our proposed criteria is given in Table 2. We will use this example in the text to highlight some of the relevant issues. It should be noted from the outset that our criteria, like any guidelines for making an inference, do not provide hard and fast rules; they simply represent an organized approach to making reasonable judgments.

Guidelines for Deciding whether Apparent Differences in Subgroup Response Are Real

Conceptual Approach Underlying the Guidelines

Subgroup analyses of data from randomized trials or meta-analyses are undertaken to identify "effect modifiers," characteristics of the patients or treatment that modify the effect of the intervention under study. Statistical "interactions" in a set of data are measured to estimate effect modification (an epidemiologic concept) in the population represented by the study sample (19). The term interaction is sometimes (but not in this paper) also used to refer to the concept of synergism or antagonism, a biologic mechanism of action in which the combined effect of two or more factors differs from the sum of their solitary effects (20). In the following discussion, we use the term "interaction" to refer to situations in which the observed effectiveness of an intervention differs across subgroups.

The premise underlying the hypothesis that subgroup analyses do more harm than good is that "unanticipated qualitative interactions" are unusual and, when apparent unanticipated interactions are discovered, they are treatment effects in drugs of a single class; this would suggest that the best estimate of the effect of any one drug is the overall effect of the group of drugs across all methodologically adequate, randomized, controlled trials (21). There is confusion, however, over the fundamental distinction between a "qualitative interaction" and a "quantitative interaction" (22). Although a strict definition of a qualitative interaction would mean that there is a sign reversal (22) (meaning that the treatment is beneficial in one group and harmful in another), it is also used to refer to a substantial quantitative interaction (that is, a difference in the magnitude of effect that is clinically important). From a clinical point of view, it is important to recognize that a substantial quantitative interaction can be as important as a qualitative interaction. For instance, the side effects of a treatment may be such that it is worth administering to patients in whom the magnitude of the treatment effect is large, but not to patients in whom the treatment effect is small or moderate.

Having said this, it is still reasonable to distinguish between interactions that are clinically trivial and those that are clinically important. The former can be ignored, and that is the point at which our guidelines begin. Once the clinician has decided that an interaction, if real, would be important, the subsequent six criteria can be used to help decide on the credibility of the proposed subgroup difference. Three of the criteria (2 to 4) are markers of the potential for random error (that is, mistakes due to chance); one (criterion 5) is a marker of the potential for systematic errors; and the last two address the consistency of the evidence (criterion 6) and its biologic plausibility (criterion 7).

The Guidelines

1. Is the Magnitude of the Difference Clinically Important?

Given the extent of biologic variability, it would be surprising *not* to find interactions between treatment effects and various other factors. Differences in the effect of treatment are likely to be associated with differences in patient characteristics, differences in the administration of the treatment (such as different surgeons or different drug doses), and differences in the primary end point. However, it is only when these

Table 2. An Example of a Hypothesized Difference in Subgroup Response: Digoxin is More Effective in Patients with More Severe Heart Failure

Criterion	Result
1. Magnitude of the difference	Clinically important differentiation between responders and nonresponders.
2. Statistical significance	Yes, *P* values were less than 0.01 in both studies.
3. A priori hypothesis	Yes, the hypothesis was suggested by results of one study and tested in a second study.
4. Small number of hypotheses	If viewed as severity of heart failure, yes. If viewed as components (for example, heart size, third heart sound, ejection fraction), no.
5. Within-study comparisons	Yes, in two crossover trials, comparisons were within studies.
6. Consistency across studies	Yes, in two studies tested. However, it was not tested in other trials, and this is necessary for confirmation.
7. Indirect evidence	Yes, biologically plausible that clinically important response is restricted to those with more severe heart failure.

that is, when they are large enough that they would lead to different clinical decisions for different subgroups—that there is any point in considering them further.

As a rule, the larger the difference between the effect in a particular subgroup (or with a particular drug or dosage of drug) and the overall effect, the more plausible it is that the difference is real. At the same time, as the difference in effect size between the anomalous subgroup and the remainder of the patients becomes larger, the clinical importance of the difference increases.

Unfortunately, if the results of subgroup analysis are only reported for the subgroups within which sizable treatment differences are found, the estimates of the magnitude of the interaction will be biased because only the extreme estimates are reported (23). This is analogous to regression to the mean (the tendency for extreme findings, such as unusually high blood pressure values, to revert toward less extreme values on repeated examination) (24). Moreover, when the overall treatment effect is modest, there is a good chance of finding a "qualitative" interaction even when only two subgroups are examined (17).

When they report the results of subgroup analyses, authors should make clear to readers how many comparisons were made and how it was decided which ones to report. Given current publication practices, however, were the reader simply to conclude that a reported interaction is real just because it is large, he or she would be wrong more often than right. Thus, having determined that an interaction, if real, is large enough to be important, it is essential to consider other criteria.

2. Was the Difference Statistically Significant?

Any large data set has, imbedded within it, a certain number of apparent, but in fact spurious, interactions. Statistical tests of significance can be used to assess the likelihood that a given interaction might have arisen due to chance alone. For example, Yusuf and colleagues (25), in an overview of randomized trials of beta blocker treatment for myocardial infarction, compared agents with and without intrinsic sympathomimetic activity (ISA) and found that the agents without ISA seemed to produce a larger effect than the ones with it. This difference was significant at the 0.01 level, indicating that it was unlikely to have occurred due to chance alone. Yet, two subsequent trials, one of an agent with ISA and one of an agent without ISA, showed the opposite result and, when added to the overview, eliminated the statistical significance of the interaction (26). There are several possible explanations for this, including chance. In other words, although events that occur one out of a hundred times might be considered rare, they do occur. Of course, the lower a P value is, the less likely it is that an observed interaction can be explained by chance alone.

Conversely, just as it is possible to observe spurious interactions, chance is likely to lead to some studies (among a large group) in which even a real interaction is not apparent. This is particularly true if the studies are small and the clinical end points of interest are infrequent. In this case, the power to detect an interaction would be low. Because subgroup analyses always incarry a greater risk for making a type II error—falsely concluding that there is no difference.

Statistical techniques for conducting subgroup analysis include the Breslow-Day technique and regression approaches (27). With the Breslow-Day technique and similar approaches (28), it is possible to use a test for homogeneity to estimate the probability that an observed interaction might have arisen due to chance alone. More commonly, authors simply conduct a number of comparisons for different subgroups and apply chi-square tests or t-tests without formally testing for interactions.

This practice, together with only reporting subgroups within which sizable treatment differences are found, can lead to an overestimate of the significance as well as the size of the difference. One way of adjusting for this bias is to use Bayes or empiric Bayes methods, which shrink the extreme estimates toward the overall estimate of treatment effect (23, 29, 30). Both a point estimate of the magnitude of the difference and a confidence interval can be obtained using these approaches.

Regression models, such as logistic regression (28), can also be used for analysis of interactions if the interactions are modeled by product terms. This approach allows for testing the significance of an interaction while controlling for other factors. If there are many subgroup factors, however, the number of product terms necessary for an adequate modeling of the interactions may be greater than the number of observations; an analysis of the interactions is then impossible. An additional problem with this approach is deciding which of many possible interaction terms to enter into the model as well as the potential for bias in their selection.

Methods for selecting factors to include have been proposed (31) in addition to other approaches to subgroup analysis (15, 18, 23, 27). Although it is not important for clinical readers to understand the details of these approaches, it is important to understand the concepts of statistical significance and power in subgroup analysis. Statistical analysis is a useful tool for assessing whether an observed interaction might have been due to chance alone, but it is not a substitute for clinical judgment.

3. Did the Hypothesis Precede Rather than Follow the Analysis?

Surveying patterns of data that suggest possible interactions may, in fact, prompt the analysis that "confirms" the existence of a possible interaction. As a result, the credibility of any apparent interaction that arises out of post-hoc exploration of a data set is questionable.

An example of this was the apparent finding that aspirin had a beneficial effect in preventing stroke in men with cerebrovascular disease but not in women (32). This interaction, which was "discovered" in the first large trial of aspirin in patients with transient ischemic attacks, was subsequently found, in other studies and in a meta-analysis summarizing these studies (33), to be spurious. This finding, like the streptokinase example, is an example of a "false negative" subgroup analysis. In this instance, many physicians withheld

considerable period.

Whether a hypothesis preceded analysis of a data set is not necessarily a black or white issue. At one extreme, unexpected results might be clearly responsible for generating a new hypothesis. At the other extreme, a subgroup analysis might be clearly planned for in a study protocol to test a hypothesis suggested by previous research. Between these two extremes lie a range of possibilities, and the extent to which a hypothesis arose before, during, or after the data were collected and analyzed is frequently not clear. For example, if data monitoring detects a seeming interaction in a long-term study, it may be possible to state the hypothesis and then test it in future analyses (34). This technique may be most appropriate if additional study patients are still to be accrued.

Although post-hoc analyses will sometimes yield plausible results, they should generally be viewed as hypothesis-generating exercises rather than as hypothesis testing. Decisions about which analyses to do and which ones to report are much more likely to be data driven with post-hoc analyses and thereby more likely to be spurious. On the other hand, when a hypothesis has been clearly and unequivocally suggested by a different data set, it moves from a hypothesis-generating toward a hypothesis-testing framework. In Bayesian terms, the higher prior probability increases the posterior probability (after the subgroup analysis) of an interaction being real (29, 30).

If a hypothesis about an interaction has arisen from exploration of a data set from a study, then an argument can be made for excluding that study from a meta-analysis in which the hypothesis is tested. Certainly, if the hypothesis is confirmed in a meta-analysis that excludes data from the study that originally suggested the interaction, the inference rests on stronger ground. If the statistical significance of the interaction disappears or is substantially weakened when data from the original study are excluded, the strength of inference is reduced.

When considering post-hoc analyses, it should be kept in mind that they are more susceptible to bias as well as to spurious results. The reader should be particularly cautious about analysis of subgroups of patients that are delineated by variables measured after baseline, even if the hypothesis preceded the analysis. If the treatment can influence whether a participant becomes a member of a particular subgroup, the conclusions of the analysis are open to bias. For instance, one might hypothesize that compliers will do better if they are in the treatment group than in the control group but that noncompliers will do equally well in both groups. The reasons for compliance and noncompliance, however, probably differ in the treatment and control groups. As a result, in this comparison, the advantages of randomization (and with it, the validity of the analysis) are lost.

An example of the evolution of a hypothesis concerning responsive subgroups comes from the investigation of the efficacy of digoxin in preventing clinically important exacerbations of heart failure in heart-failure patients in sinus rhythm (*see* Table 2). Lee and colleagues the drug to be effective. They did a regression analysis that suggested that only one factor—the presence of a third heart sound—predicted who would benefit from the drug. Only patients with a third heart sound were better off while taking digoxin. The hypothesis that this might be one of the predictors appears to have preceded the study. Nevertheless, on the basis of the foregoing discussion, the investigators were perhaps too ready to conclude that digoxin use in heart-failure patients in sinus rhythm should be restricted to those with a third heart sound.

4. Was the Subgroup Analysis One of a Small Number of Hypotheses Tested?

Post-hoc hypotheses based on subgroup analysis often arise from exploration of a data set in which many such hypotheses are considered. The greater the number of hypotheses tested, the greater the number of interactions that will be discovered by chance. Even if investigators have clearly specified their hypotheses in advance, the strength of inference associated with the apparent confirmation of any single hypothesis will decrease if it is one of a large number that have been tested. In their regression analysis, Lee and colleagues (35) included 16 variables. This relatively large number increases the level of skepticism with which the presence of a third heart sound as an important predictor of response to digoxin should be viewed.

Unfortunately, as noted above, the reader may not always be sure about the number of possible interactions that were tested. If the investigators chose to withhold this information, despite admonitions not to do so, and reported only those that were "significant," the reader is likely to be misled.

The Beta-Blocker Heart Attack Trial (BHAT) randomized approximately 4000 patients to propranolol or placebo after a myocardial infarction (36). Subsequently, 146 subgroup comparisons were done (37). Although the estimated effects of the treatment clustered around the overall effect, the effect in some small subgroups appeared to be either much more effective or ineffective. The overall pattern, which approximated a "normal" distribution, would suggest that most of the observed difference in effect among the various subgroups was due to sampling error rather than to true interactions.

Another way to consider this is in terms of the effect of multiple comparisons on *P* values. The more hypotheses that are tested, the more likely it is to make a type I error, that is, to reject one of the null hypotheses even if all are actually true. Assuming that no true differences exist, if 100 different comparisons are made, five can be expected to yield a *P* value of 0.05 or less by chance alone. In this situation, a more appropriate analysis would account for the number of subgroups, their relation to other subgroups, and the size of the effect within subgroups and overall (23).

5. Was the Difference Suggested by Comparisons within Rather than between Studies?

Making inferences about different effect sizes in different groups on the basis of between-study differences

entails a high risk compared with inferences made on the basis of within-study differences. For instance, one would be reluctant to conclude that propranolol results in a different magnitude of risk reduction for death after myocardial infarction than does metoprolol on the basis of data from two studies, one that compared propranolol with placebo and another that compared metoprolol with placebo. This could be thought of as an indirect comparison. A direct comparison would involve, in a single study, patients being randomized to receive either placebo, propranolol, or metoprolol. If, in such a direct comparison, clinically important and statistically significant differences in magnitude of effect between the two active treatments were demonstrated, the inference would be quite strong.

An example that illustrates this point comes from an overview examining the effectiveness of prophylaxis for gastrointestinal bleeding in critically ill patients (38). Histamine$_2$- receptor (H_2) antagonists and antacids, when individually compared with placebo, had comparable effects in reducing overt bleeding (common odds ratios of 0.35 in both cases). In contrast, direct comparison from studies in which patients were randomized to receive H_2 antagonists or antacids have shown a statistically significantly greater reduction in bleeding with the latter (common odds ratio, 0.56).

The reason that inference on the basis of between-study differences is so potentially misleading is that there may be a myriad of factors, aside from the most salient difference, which is the basis of the inference being made, that could explain the interaction. For instance, aside from differences in the specific drugs used, different populations (varying in risk for adverse outcomes, for example), varying degrees of co-intervention, or varying criteria for gastrointestinal bleeding each could explain the results. These differences would not be plausible explanations if the inference were based on within-study differences in randomized trials in which populations studied, control of co-intervention, and outcome criteria were all identical.

Stated simply, between-study inferences are based on comparisons between noncomparable groups; even when all of the individual studies were randomized, patients were not randomized to one study or another. Clinical decisions based on between-study comparisons should be made cautiously, if at all. As a rule, inferences based on between-study comparisons should be viewed as preliminary and as requiring confirmation from direct within-study comparison. This is true whether the between-study comparison has to do with different groups or different interventions.

6. Was the Difference Consistent across Studies?

A hypothesis concerning differential response in a subgroup of patients may be generated by examination of data from a single study. The interaction becomes far more credible if it is also found in other studies. The extent to which a comprehensive scientific overview of the relevant literature finds an interaction to be consistently present is probably the best single index as to whether it should be believed.

In other words, the replication of an interaction in independent, unbiased studies provides strong support for its believability. On the other hand, there are two reasons to be cautious in applying this criterion. The first goes back to sample size. Because subgroup analyses often include small numbers of patients, the results tend to be imprecise and the extent to which results from different studies are consistent can be uncertain. The second caution relates to making between-study comparisons. For the same reason that it is risky to base conclusions on between-study differences, it is only reasonable to expect variation in the results of trials of the same therapy, due to differences in the study populations, the interventions, the outcomes, and the study designs, as well as the play of chance. Thus, when assessing the consistency of results, it is important to consider both the power of the comparisons (or their statistical certainty) and other differences between studies that might influence the results.

The hypothesis concerning a third heart sound as a predictor of response to digoxin in heart-failure patients in sinus rhythm was tested in a second crossover, randomized trial (39). The presence of a third heart sound proved a weaker predictor than in the initial study, although its association with response to digoxin did reach conventional levels of statistical significance. However, a number of factors that, like a third heart sound, reflect greater severity of heart failure, were associated with response to digoxin. Thus, support for a more general hypothesis, that response is related to the severity of heart failure, was provided by the second study.

Other studies have examined the efficacy of digoxin in heart-failure patients in sinus rhythm, and these have been summarized in a meta-analysis (40). Unfortunately, none of these studies has conducted subgroup analyses addressing the issue of differential response according to different severity of heart failure. Had these analyses been done in the other studies, the hypothesis would likely have been confirmed or refuted with substantially greater confidence. As it is, we would be inclined to view the conclusion as tentative; the strength of inference is only moderate.

7. Is There Indirect Evidence to Support the Hypothesized Difference?

We are generally more ready to believe a hypothesized interaction if indirect evidence (such as from animal studies or analogous situations in human biology) makes the interaction more plausible. That is, to the extent that a hypothesis is consistent with our current understanding of the biologic mechanisms of disease, we are more likely to believe it. Such understanding comes from three types of indirect evidence: from studies of different populations (including animal studies); from observations of interactions for similar interventions; and from results of studies of other, related outcomes (particularly intermediary outcomes).

The extent to which indirect evidence strengthens an inference about a hypothesized interaction varies substantially. In general, evidence from intermediary outcomes is the strongest type of indirect evidence. Evidence of differences in immune response, for example, can provide strong support for a conclusion that there is an important difference in the clinical effectiveness of a

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.