Subgroup Analyses in Nephrology Clinical Trials

Steven Fishbane,* Hitesh H. Shah,* Ashish Kataria,* Shayan Shirazian,† and Rajiv Agarwal‡

Introduction
There are various experimental designs available for evaluating the effectiveness of medical interventions. Although none are completely without limitations, the randomized controlled trial (RCT) is generally accepted as providing the most reliable evidence in the hierarchy of methodological options (1). The strength of well-designed RCTs derives from limitation of bias through random allocation of patients to treatment groups and greater validity of statistical tests (2,3).

A properly designed RCT defines one or more outcomes of interest. A sample size is determined that is predicted to provide sufficient statistical power to detect clinically relevant differences in outcomes between study groups. When the study is completed, the major focus should be on reporting results of the outcomes for the entire study population (intervention versus control group). Because the study is planned and powered for this purpose, the greatest evidence resides here. However, it is a frequent practice to go beyond the primary analyses and perform analyses of the effect of the intervention in subgroups of patients as well. This practice is done to determine whether the results of the trial apply equally among participants of diverse characteristics.

A subgroup analysis may be defined as an evaluation of treatment effect in subgroups of patients defined by baseline characteristics (4). For example, do the results of a trial apply equally to men and women? Analyses of patient subgroups in RCTs help to provide a better understanding of treatment effect heterogeneity (5–7). This understanding could improve therapeutic targeting and allow for subsequent studies to test treatment effects in selected patient subsets. However, the results of subgroup analyses are subject to misinterpretation and do not have the strength of evidence of outcomes tested in the complete randomized population (8). Accordingly, results of subgroup analyses should be viewed with caution and used primarily for the purpose of generating hypotheses that may be tested in subsequent trials (9–11). This caution is especially true when the subgroup analyses have not been pre-specified. In this commentary, we will review basic concepts on performance and reporting of subgroup analyses, and we will discuss how these analyses are used in nephrology trials. It should be noted that similar considerations apply to observational studies, although these studies will not be the focus of the current article.

Statistical Considerations in Subgroup Analyses
A cornerstone of study design is to ensure sufficient statistical power to detect clinically relevant differences between study groups (12). A common problem is to study too few patients, limiting the study’s statistical power. A smaller number of patients increases the risk of not observing a true treatment effect. The latter, an example of false-negative results, is also known as a type II statistical error. In an interesting study, the work by Freiman et al. (13) analyzed 71 published studies that did not show a significant difference between study groups. Remarkably, 71.4% of these studies were underpowered, with greater than 10% risk of not detecting a 50% difference between groups. Because subgroup analyses reduce the sample size even more, most are underpowered. False-negative findings in subgroups are, therefore, an ever-present risk.

An even greater risk for subgroup analyses is the finding of a treatment effect when none truly exists. This finding is called false discovery, an example of a type I statistical error. The problem, termed multiplicity, occurs when multiple analyses are performed. To maximize use of data from a costly RCT, the temptation is often to do several subgroup analyses. The greater the number of analyses performed, the greater the likelihood of a positive result finding by chance. To understand how many subgroup analyses have been conducted, the number of outcomes tested is multiplied by the number of patient characteristics. If two patient outcomes are tested in 10 subgroups, then there are 20 different subgroup analyses. In one recent nephrology study, there were 25 subgroup analyses conducted (14). As noted in the work by Lagakos (8), when 10 independent tests are conducted, each with a $P$ value of 0.05 for significance, the likelihood of at least one false-positive result is 40%. The most effective way to manage this problem is to limit the number of analyses conducted, which is discussed below. Alternatively, because of the risk for false-positive results, a more stringent $P$ value (lower than the conventional <0.05) could be set for statistical significance. To limit the probability of a false-positive result to 5%, each test should use a criterion of $(1–0.95)^{1/x}$ for significance, where $x$ is the number of analyses conducted (9). This method is the traditional Bonferroni correction, which many statisticians consider to be too conservative. Other methods have evolved to replace the conservative Bonferroni. The Holm–Bonferroni procedure is an example of controlling...
the false discovery rate (15). Other methods have been developed as well to control for the false discovery rate in exploratory studies (16–18).

Selection of Subgroup Analyses
To the greatest extent possible, subgroup analyses should be prespecified. This specification should occur during the study design phase and then be reported in the methods section of the study manuscript. Failure to predefine subgroup analyses diminishes the credibility of results (19). When subgroup analysis results are reported that were not prespecified, it may be unclear how many analyses were actually performed. This finding suggests the possibility that analyses were conducted post hoc based on inspection of collected data, a practice sometimes referred to as fishing. In contrast, analyses that are prespecified, based on biologically plausible hypotheses, suggest a thoughtful, credible process (4). The process of subgroup analysis selection should begin with a consideration of biologic plausibility (where is it likely that differential effects of an intervention would be found). Each subgroup analysis should be prespecified, and the reason for its selection should be explained in the methods section of the publication. In addition, the direction of expected effect should be stated. These actions help to increase the credibility of analyses and interpretation of results.

An example of subgroup analysis selection from nephrology comes from the Dialysis Clinical Outcomes Revisited (DCOR) Trial (20). Hemodialysis patients were randomized to treatment with either sevelamer or calcium-based phosphate binders for up to 45 months. The primary endpoint, all-cause mortality, was not found to be significantly different between the two treatment groups. A subgroup analysis was conducted for patients who remained on treatment for at least 2 years. The results indicated a survival benefit in this subgroup for sevelamer-treated patients. This analysis was not prespecified in the methods section of the publication, and there is no assertion in presenting the results that it was prespecified. This finding raises the question of whether inspection of the data motivated presentation of results for this subgroup. Again, this finding would be problematic; (1) performing multiple analyses greatly increases the risk of false-positive findings, and (2) the failure to prespecify suggests the possibility that many analyses could have been conducted looking for any positive finding. To the authors’ credit, the work by Suki et al. (20) appropriately refrained from making claims about these findings in their discussion.

Analysis and Reporting of Subgroups
Subgroups require a different method of statistical analysis than primary analyses. An intuitive but incorrect approach is to report on a treatment effect in one subgroup (example: men) and then in another subgroup (example: women) and claim a differential effect based on these separate analyses. This method is not appropriate for ascertaining heterogeneity in treatment effects. The correct statistical analysis is a test for interaction (21,22). Despite the use of this statistical approach, one must realize that most trials lack the statistical power to detect subgroup heterogeneity. Therefore, the failure to find differential treatment effects in subgroups does not rule out that heterogeneity actually exists. For this reason, authors should avoid conclusive statements as to treatment effects in subgroups.

The study publication should list all prespecified subgroup analyses in the methods section. Specifically, authors have a responsibility to report on the total number of subgroup analyses performed. This reporting means both prespecified analyses and analyses that may have been conducted on a post hoc basis. If subgroups analyses are performed post hoc (not optimal), then it is critically important to report how many analyses were conducted and not just positive findings. Failure to do so suggests the possibility that numerous analyses may have been conducted in an effort to find a positive result. In addition, for continuous variables, specific cutoff values (e.g., age<40 years versus age>40 years) should be prespecified. Without this specification, it is not possible to know whether multiple cutoff values were tested in a search for significance.

One effective approach is to present the subgroup results by first listing the subgroups studied. For each subgroup, list the number of patients in the subgroup and the treatment effect size in each group. We prefer presentation of confidence intervals for each subgroup but not P values for individual subgroups. By focusing on potential effect size rather than statistical significance, problems of interpretation related to multiplicity and underpowering can be minimized.

Because of the limitations noted, discussion of results of subgroup analyses should be appropriately conservative. Claims should be limited to discussion of the results and if appropriate, consideration of a biologic explanation for observed heterogeneity. Reference to other studies that have conducted similar analyses is helpful and may strengthen subgroup findings. Discussion should avoid treatment recommendations based on subgroup results and should warn readers that results require confirmation in studies specifically powered for the purpose. An example of how subgroup analyses might affect clinical treatment was from an RCT that found aspirin to be effective in the secondary prevention of stroke. Despite the overall beneficial effect, a subgroup analysis from this study found that the treatment was not effective in women (23). Several years later, the initial subgroup analysis was refuted; a meta-analysis found that aspirin was actually highly effective in women (24). In the intervening period, in clinical practice, aspirin may have been underused in women. Examples like this one are a reminder that subgroups analyses must be performed with care and that their results must interpreted and discussed with great care. Other recommendations for performing and reporting on subgroup analyses are found in Table 1.

Even carefully performed subgroup analyses need to be interpreted with caution. The Prospective Randomized Amlodipine Survival Evaluation Study compared amlodipine with placebo for patients with advanced congestive heart failure (CHF) (25). The primary endpoint showed no significant benefit for amlodipine. There were a number of prespecified subgroup analyses. One focused on cause of CHF: ischemic versus nonischemic. Despite the lack of efficacy by amlodipine in the primary analysis, in the nonischemic subgroup, amlodipine was found to reduce the risk of endpoints by 31%. This finding went counter to the hypothesis: “Yet, some caution is warranted, since our a priori expectation was that amlodipine would be
more beneficial in patients with ischemic heart disease” (25). The work by O’Connor et al. (25) recommended that a second trial be conducted in the nonischemic CHF population. This work is an example of subgroup reporting at its best—a clear expression of caution and the need for confirming studies. In fact, the subsequent study found no benefit of amiodipine in nonischemic CHF (26).

As this study indicates, when heterogeneity in treatment effect is discovered, the most rigorous confirmation is performance of a new RCT in the subgroup of interest. In fact, this continued testing is rarely done. When subsequent RCTs have been performed in subgroups, it is unusual for them to confirm the discovered heterogeneity in treatment effect. Therefore, subgroup analyses are most helpful when consistency in treatment effect is proven, lending strength to findings in the whole study population. In contrast, differential effects in subgroups should always be viewed with caution.

An example of reporting problems for subgroup analyses also comes from the DCOR Study (discussed above) (18). A well publicized subgroup analysis in DCOR was of patient age. Among patients greater than 65 years old, there was a survival advantage for sevelamer-treated patients. The work by Suki et al. (20) stated in the results section that this analysis was prespecified. Several criticisms apply. First, this analysis should have been mentioned in the methods section of the publication with an explanation of why the analysis was conducted and the expected directionality of effect. Second, because randomization was stratified by age above or below 55 years, it is unclear why the subgroup cutoff was 65 years. Third, it seems from the reporting that multiple subgroups were examined, because the work mentions race, sex, diabetes status, cause of kidney disease, and dialysis duration (as well as patients on treatment over 2 years) (20). As discussed above, multiple comparisons increase the risk for a false-positive finding. Fourth, the work by Suki et al. (20) devotes an entire paragraph in the discussion section to the age subgroup analysis. This discussion should have noted the limitations of subgroup analyses and the hypothesis-generating nature of the findings, but the work by Suki et al. (20) should be complimented for relating the subgroup results to other studies for external validity. Finally, the age ≥ 65 years subgroup analysis was reported in the abstract (“Only in patients over 65 years of age was there a significant effect of sevelamer in lowering the mortality rate”) (20). Because of the limitations and risk for misinterpretation of subgroup analyses, results should not be reported in abstracts.

### Subgroup Analyses in Nephrology Studies

To better understand how subgroup analyses are conducted and reported in nephrology trials, we performed an analysis of nephrology trials published between July 1, 2010 and June 30, 2011. The literature search consisted of a review of nephrology and transplantation journals and four general medical journals (New England Journal of Medicine, Lancet, Annals of Internal Medicine, and Archives of Internal Medicine). Potential articles were found through both a Medline search using Medical Subject Headings terms and free text as well as complete review of the tables of contents of all selected journals. All RCTs were potentially eligible; articles were excluded if they represented long-term follow-up of previously reported RCTs, post hoc secondary analyses of previously reported RCTs, or phase 1 and pharmacokinetic studies.

Table 2 summarizes the characteristics of the trials. We found that approximately one-third (37.3%) of published nephrology randomized controlled trials reported subgroup analyses (Table 3). This finding is less than the finding reported for subgroup analyses in high-impact medical journals (60%) (4,10). However, we found deficiencies in reporting in the nephrology trials. The majority failed to prespecify that subgroup analyses would be performed (77.4%). When articles did report prespecification, there was an almost uniform failure to list all subgroup variables analyzed, explain why the subgroup was chosen for analysis, and report the expected directionality of treatment effect. Statistical testing was inappropriate, with a test of interaction reported in only 35.3% of subgroup analyses. Instead, it was more common to claim treatment effects based on separate tests for each level of subgroup. In addition, when subgroup analyses were discussed, there was a frequent failure to note the limitations of the analyses and discuss that the results should be viewed with caution. We found positive claims to be common in discussion sections, with general overinterpretation of the importance of the subgroup results.

An example of inappropriate subgroup analyses in nephrology comes from an excellent but small (n = 32) RCT of patients with Wegener’s Granulomatosis randomized to immunosuppressive therapy with or without added plasma exchange (PE) (27). The primary study results were that the addition of PE was found to be beneficial in the overall study population. A subgroup analysis was performed to determine the effect of the treatment in patients with lower or higher baseline serum creatinine. The analysis is mentioned in the methods section, but it is not clear whether

---

**Table 1. Recommendations for reporting of subgroup analyses**

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Plan subgroup analyses during the study’s design phase</td>
</tr>
<tr>
<td>2</td>
<td>Limit analyses to variables where a plausible hypothesis for treatment heterogeneity exists</td>
</tr>
<tr>
<td>3</td>
<td>In the methods section of the publication, list the prespecified analyses and explain the rationale and expected direction of effect for each analysis; if additional unplanned post hoc analyses are conducted (not recommended), the number performed and reported on should be clearly listed</td>
</tr>
<tr>
<td>4</td>
<td>Do not report results of subgroup analyses in the abstract</td>
</tr>
<tr>
<td>5</td>
<td>Use statistical tests for interaction and report effect estimates</td>
</tr>
<tr>
<td>6</td>
<td>When discussing findings of subgroup analyses, provide context to prior studies, if available, avoid overinterpretation, explain the hypothesis-generating nature of the findings, and emphasize the overall study results, with far less emphasis on subgroup results</td>
</tr>
</tbody>
</table>

Modified from references 1 and 9, with permission.
therapy in WG patients at creatinine levels >250 μmol/L.” These statements are not supported by the subgroup analyses. In our opinion, they overinterpret results that prematurely provoke a claim to immediate applicability to clinical practice (27).

In conclusion, although subgroup analyses may be helpful for suggesting heterogeneous effects of treatments, the results may be misleading and overinterpreted. In nephrology trials, there are important deficits in the quality of subgroup reporting. To avoid misinterpretation and any adverse effect on clinical practice, standards are needed, and adherence is critical. With proper planning, conduct, and reporting, subgroup analyses can be a helpful feature of clinical trials, including those trials in nephrology.

Disclosures
None.

References

Published online ahead of print. Publication date available at www.cjasn.org.