Are Unadjusted Analyses of Clinical Trials Inappropriately Biased Toward the Null?

David M. Kent, MD, MS; Thomas A. Trikalinos, MD, PhD; Michael D. Hill, MD, MS

See related article, pages 888–894.

One of the delights of clinical practice has proven to be a major nuisance for clinical research: patients are nonidentical. Indeed, patients have multiple characteristics that influence the likelihood of the outcome of a disease, which can make it difficult in the extreme to accurately discern the effects of therapy from casual clinical experience or even careful observational studies. Randomization, a process by which patients are assigned to a treatment arm by chance rather than by choice, was a brilliant innovation that has made possible causal inferences regarding a treatment’s effect. Although randomization is not perfect in practice, it is remarkably effective at ensuring the comparability between treatment groups, so much so that it has almost tricked us into thinking that patient differences in outcome risks have been rendered irrelevant in the context of clinical trials.

However, randomization only ensures similarity of the outcome risks between treatment groups; it does nothing to mitigate the between-patient differences in outcome risks within treatment groups. These differences can lead to clinically important differences in treatment effects across patients such that the between-patient differences in outcome risks between treatment groups; it does nothing to mitigate the likelihood of the outcome of a disease, which can make it difficult in the extreme to accurately discern the effects of therapy from casual clinical experience or even careful observational studies. Randomization, a process by which patients are assigned to a treatment arm by chance rather than by choice, was a brilliant innovation that has made possible causal inferences regarding a treatment’s effect. Although randomization is not perfect in practice, it is remarkably effective at ensuring the comparability between treatment groups, so much so that it has almost tricked us into thinking that patient differences in outcome risks have been rendered irrelevant in the context of clinical trials.

However, randomization only ensures similarity of the outcome risks between treatment groups; it does nothing to mitigate the between-patient differences in outcome risks within treatment groups. These differences can lead to clinically important differences in treatment effects across patients such that the summary results of a trial may not apply to all, or even most, patients in the trial. Heterogeneity of risk can create an even more fundamental, and even less appreciated, problem with the summary results of clinical trials: Even in the absence of any heterogeneity of treatment effect (ie, when all patients get an identical treatment benefit), and in the absence of confounding and bias, risk heterogeneity can still play a very mischievous role such that an unadjusted (crude) analysis may be both inefficient and yield an inaccurate estimate of the summary treatment effect.

In this issue of Stroke, Gray et al perform one of the more comprehensive analyses of the effects of risk adjustment on statistical power and sample size requirements using the unique Optimizing Acute Stroke Trials (OAST) database. Using 23 different trials that provide data on baseline characteristics, and have a nonneutral treatment effect, they find a consistent increase in the statistical power or (alternatively) a consistent decrease in the sample size required comparing risk-adjusted analysis with conventional (unadjusted) analysis.

This study adds to the growing literature showing that risk-adjusted analyses can make trials more efficient, reducing the required sample size on the order of 15% to 30%. This effect is not widely understood and has been attributed to an increase in “precision” or a reduction in variance. However, across these studies, there has also been a consistent change in the magnitude of the estimated treatment effect in the risk-adjusted compared with the crude analysis; the risk-adjusted OR, hazard ratio, or hazard ratio, always shows a larger treatment effect than the crude analysis.

Given these results, one might expect that routine risk adjustment of clinical trial results will be taken up immediately, because pharmaceutical companies (and even academics) are not exactly well known for ignoring trial costs and even less for biasing their trials toward the null. Yet, there remain barriers to routine risk-adjusted analyses, which are more complex and less transparent. A crude analysis relies simply on counting those with and without the outcome in each arm of the study and reporting the ratio; how could this be biased? On the other hand, the results of a risk-adjusted analysis are conditional on the selected covariates; how can we trust study results when the outcome of the analysis depends on the particular variables the investigators decide to control for? Surely this must increase, not decrease, the opportunity for bias.

In fact, reporting a crude OR or a crude hazard ratio to summarize a treatment effect is arguably inappropriate, because these measures have a property referred to as noncollapsibility. That is, the OR for the total cohort will not be a weighted average of the stratum-specific ORs. This is true even in the simplest example when all patients, regardless of risk, experience a consistent treatment effect. Such an example is shown in the Table, which depicts a trial enrolling patients who belong to 4 different risk/strata for whom treatment yields a consistent improvement of their odds of a good outcome by 50% (ie, a uniform OR of 1.5 for all risk strata). Surprisingly, if one calculates the OR for the overall results, one would find that treatment increased the odds of a good outcome by only 38%. This represents an underestimation of the within-stratum treatment effect of almost 25%.

The property of noncollapsibility is related to the nonlinearity of the OR. It has been shown mathematically that, in the presence of heterogeneity, the crude OR will always be more conservative (ie, closer to 1) than the within strata OR. If, for example, the outcome rates in the Table represented bad outcomes instead of good outcomes, and the within-strata treatment effect was 0.75 (a 25% reduction in the odds of the outcome), the crude OR would be 0.79 (a 21% reduction in the odds of the outcome). Risk adjustment corrects this “foreshortening” of the crude OR-based treatment effect by comparing like to like.

When is noncollapsibility important to consider? First, it is only an issue when the measure of effect is nonlinear like with OR and hazard ratio; thus, when relative risk is the effect
measure, it is not an issue. Second, when outcome rates within all strata are low, the effect will be negligible. However, when the outcome rates within one or more strata are high, as is typical in stroke, this effect can be substantial. Outcome rates within strata as high as those shown in the Table are not unusual for stroke. According to the Stroke-Thrombolytic Predictive Instrument (TPI), the probability of a good outcome (modified Rankin Scale ≤1) would be approximately 80% for patients who are male and 60 years old with a National Institutes of Health Stroke Severity score of 5 or 6. Indeed, among patients enrolled in thrombolytic trials, the expected control outcome rate in the quintile of patients with the best prognosis is approximately 70%. Still, even in trials in which the average outcome rates are relatively low, as in many cardiovascular trials, the presence of a group at high risk for the outcome can cause the crude OR to be conservative compared with the risk-adjusted OR.

When the crude OR differs from the risk-adjusted OR, which one should be preferred? From the perspective of trial efficiency, it has been demonstrated consistently that risk adjustment leads to diminished sample size requirements. In terms of transparency, some might argue that using the group results is the most simple and understandable approach. However, this point is certainly arguable, because the treatment effect estimated based on group averages is conditional on the degree of heterogeneity in the sample. Even where all patients get the same treatment effect (as defined by the OR), a large, simple, broadly inclusive clinical trial will counterintuitively yield a more modest treatment effect estimate than the average of several clinical trials targeted to specific risk groups. The risk-adjusted analysis also has the advantage of estimating the more clinically relevant patient-level effect size; in simulations, which permit one to specify a “true” treatment effect, results show that a crude analysis will consistently underestimate this “true” effect, whereas adjusted analyses are more accurate.

On the other hand, real life is more complicated than our example and other simulations, and one cannot automatically ascribe discrepancies between the risk-adjusted and crude effects to noncollapsibility alone. After all, despite randomization, residual imbalances across treatment arms may persist for both observed and unknown factors alike. While risk adjusting rebalances for the known factors, its effect in any given trial on the myriad unknown factors that influence outcomes remains beyond scrutiny as does the influence of these unknown factors on the treatment effect. Although this should not systematically introduce new biases, it would seem that using nonlinear measures of effect, which avoids the issues of noncollapsibility and thus the need to risk-adjust, may be the best choice, but this may not always be an available option, especially in time-to-event analyses.

The problem of noncollapsibility in effect measures used in clinical trials remains underappreciated, its causes buried rather deeply in the literature. We have found that even experienced statisticians are frequently surprised and delighted when confronted with the paradox shown in the Table. However, can this statistical parlor trick really have important consequences in the outcome of clinical trials; can it explain in part the lack of progress in stroke trials? Although poor translation in stroke therapeutics clearly has other important causes, it is becoming apparent that inefficiencies in effect measures may not be trivial and that failure to account for risk heterogeneity when using nonlinear effect measures may be important.

### Source of Funding
Dr. Kent and Trikalinos are partially supported by a grant of the National Institute of Health (NIH/NCRR 1UL1 RR025752).

### Disclosures
None.

### References

**Key Words:** clinical trials  statistical analysis

---

### Table. Unadjusted Summary Results Underestimate Stratum-Specific Odds Ratios Even When Treatment Benefit Is Uniform

<table>
<thead>
<tr>
<th>Risk/Severity Strata</th>
<th>Control Rate</th>
<th>Treatment Rate</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe stroke</td>
<td>20.0%</td>
<td>27.3%</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>40.0%</td>
<td>50.0%</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>60.0%</td>
<td>69.2%</td>
<td>1.5</td>
</tr>
<tr>
<td>Mild stroke</td>
<td>80.0%</td>
<td>85.7%</td>
<td>1.5</td>
</tr>
<tr>
<td>Summary of trial result</td>
<td>50.0%</td>
<td>58.1%</td>
<td>1.38</td>
</tr>
</tbody>
</table>

**Summary of trial result**

50.0% 58.1% 1.38

**NOTE:** The problem of noncollapsibility in effect measures used in clinical trials remains underappreciated, its causes buried rather deeply in the literature. We have found that even experienced statisticians are frequently surprised and delighted when confronted with the paradox shown in the Table. However, can this statistical parlor trick really have important consequences in the outcome of clinical trials; can it explain in part the lack of progress in stroke trials? Although poor translation in stroke therapeutics clearly has other important causes, it is becoming apparent that inefficiencies in effect measures may not be trivial and that failure to account for risk heterogeneity when using nonlinear effect measures may be important.

### Source of Funding
Dr. Kent and Trikalinos are partially supported by a grant of the National Institute of Health (NIH/NCRR 1UL1 RR025752).

### Disclosures
None.

### References

Are Unadjusted Analyses of Clinical Trials Inappropriately Biased Toward the Null?
David M. Kent, Thomas A. Trikalinos and Michael D. Hill

*Stroke*. 2009;40:672-673; originally published online January 22, 2009;
doi: 10.1161/STROKEAHA.108.532051

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2009 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://stroke.ahajournals.org/content/40/3/672