

# Reporting Clinical Trial Results To Inform Providers, Payers, And Consumers

Conventional analyses of clinical trials can underestimate potential risks and benefits to patients.

**by Rodney A. Hayward, David M. Kent, Sandeep Vijan, and Timothy P. Hofer**

**ABSTRACT:** Results of randomized clinical trials are the preferred “evidence” for establishing the benefits and safety of medical treatments. We present evidence suggesting that the conventional approach to reporting clinical trials has fundamental flaws that can result in overlooking identifiable subgroups harmed by a treatment while underestimating benefits to others. A risk-stratified approach can dramatically reduce the chances of such errors. Since professional and economic incentives reward advocating treatments for as broad a patient population as possible, we suggest that payers and regulatory bodies might need to act to motivate prompt, routine adoption of risk-stratified assessments of medical treatments’ safety and benefits.

Everything should be made as simple as possible, but not one bit simpler.  
—Albert Einstein

CLINICIANS, POLICYMAKERS, and governmental regulatory bodies rely on the randomized controlled trial (RCT) as the preferred “evidence” for establishing the benefits and safety of medical treatments. Improving the quality of this evidence base has become a target of international health policy with the recent proposal for mandatory registration of clinical trials in an effort to eliminate post hoc decisions not to publish trial results when they reflect unfavorably on the treatment being studied.<sup>1</sup> We present another problem with the base of

---

*Rodney Hayward (rhayward@umich.edu) is director of the Department of Veterans Affairs (VA) Center for Practice Management and Outcomes Research, VA Ann Arbor Healthcare System, and a professor of medicine and public health at the University of Michigan. David Kent is an assistant professor of medicine at Tufts University School of Medicine and a clinical investigator at the Institute for Clinical Research and Health Policy Studies at Tufts–New England Medical Center in Boston. Sandeep Vijan is a research scientist at the VA Ann Arbor Healthcare System and an assistant professor in the Department of Internal Medicine, University of Michigan School of Medicine. Timothy Hofer is research scientist at the VA Ann Arbor Healthcare System and an associate professor in the Department of Internal Medicine, University of Michigan School of Medicine.*

evidence on the safety and benefits of medical treatments, along with a proposed solution.

The main results of clinical trials are presented as the average benefit across all people in the trial. To aid decisions in applying the trial results to individuals, researchers commonly conduct subgroup analyses to identify specific groups of patients who might benefit more or less than average. Such analyses typically compare groups that differ in a single attribute (such as age or sex), and such “one-variable-at-a-time” analyses often yield little useful information. Researchers have proposed that evaluating overall risk using multivariable prediction tools (hereafter referred to as “risk-stratified” analysis) could overcome some of the limitations inherent in conventional approaches to analyzing and reporting clinical trials.<sup>2</sup>

To examine this issue, we evaluated clinical trials using conventional analyses and compared these results with those obtained using risk-stratified analyses that examine benefits for patients who were at lower versus higher risk. We conclude that the conventional approach has fundamental flaws that can result in overlooking identifiable subgroups harmed by a treatment while greatly underestimating benefits to others. Therefore, the “medical evidence” used to make critical policy decisions (for example, on drug safety, insurance benefit packages, and performance measures) might be systematically misleading or incomplete. We further conclude that although a risk-stratified approach could better elucidate how the safety and benefits of treatments vary across the population, it also runs counter to current professional and economic incentives to promote treatments to as broad a patient population as possible. Therefore, the adoption of risk-stratified assessments of the safety and benefits of medical treatments might require the active intervention of payers and regulatory bodies.

### **How The Average Results Of A Trial Can Be Misleading**

The average benefits observed in a clinical trial often do not reflect the benefits observed in all, or even most, patients in a clinical trial.<sup>3</sup> For example, Exhibit 1 shows results for a hypothetical treatment that reduces the risk of a bad outcome by 17 percent when the benefit is averaged across all study subjects. In this example, the 17 percent relative risk reduction (RRR) means that one bad outcome is prevented for every ninety-eight people treated (the number needed to treat, or NNT, is a measure of absolute treatment benefit; the lower the NNT, the more effective the treatment). However, focusing on the average benefit for the entire population (the main finding conventionally reported in clinical trials) ignores the fact that higher-risk subjects are six times more likely to benefit from treatment (NNT = 39) than lower-risk subjects (NNT = 238).

Further, when there is an appreciable risk of treatment-related adverse events, reporting only the average result of a clinical trial might obscure a group that is harmed by treatment.<sup>4</sup> Consider a hypothetical treatment that decreases the baseline risk of patients suffering a bad outcome by 30 percent over five years, but at a

**EXHIBIT 1**  
**How The Average Result From A Clinical Trial Can Underestimate The Benefit In High-Risk Study Subjects And Overestimate Benefit In Lower-Risk Subjects**

**Assumption: the overall relative reduction in overall risk of bad outcomes is 17% for all study subjects**

**Rates of bad outcomes over 5 years of treatment**

Risk group (percent of study population)	Control event rate (CER) <sup>a</sup>	Experimental event rate (EER) <sup>a</sup>	Relative risk reduction (RRR) <sup>a</sup>	Absolute risk reduction (ARR) <sup>a</sup>	Number needed to treat (NNT) <sup>a</sup>
Overall (100%)	6.0%	4.98%	17.0%	1.02%	98
Higher risk (20%)	15.0	12.45	17.0	2.55	39
Moderate risk (40%)	5.0	4.15	17.0	0.85	118
Lower risk (40%)	2.5	2.08	17.0	0.42	238

**Assumption: Treatment reduces events by 30% for all risk groups, at a cost of serious harm of 2 treatment-related events per year for every 1,000 treated<sup>b</sup>**

Overall (100%)	6.0	5.20%	13.3%	0.80%	125
Higher risk (20%)	15.0	11.50	23.3	3.50	29
Moderate risk (40%)	5.0	4.50	10.0	0.50	200
Lower risk (40%)	2.5	2.75	-10.0	-0.25	-400

**SOURCE:** Authors' statistical simulation of hypothetical study results.

**NOTE:** Minus signs denote that there was net harm, rather than benefit.

<sup>a</sup> CER is a measure of baseline risk (that is, outcome rates in the absence of treatment), EER is a measure of outcome rates for those who received the treatment, RRR is a measure of the percent reduction in risk associated with treatment (RRR is derived by dividing the magnitude of the difference between EER and CER by CER), ARR is the difference in event rates in the two groups (EER - CER), and NNT is the number of people you need to treat to prevent one bad outcome (1/ARR, with lower positive numbers indicating a more effective treatment).

<sup>b</sup> EER = (CER × 0.7) + (0.2 × 5).

price of two treatment-related severe adverse events every year for every thousand patients treated. In this instance, the average benefit (NNT= 125) greatly underestimates the benefit for high-risk patients (NNT = 29) but overestimates the benefit for the median patient (NNT = 200) (Exhibit 1). Even more of concern, the average benefit obscures the fact that this treatment increases the risk of bad outcomes in low-risk patients (40 percent of all subjects). In short, if patients are at very low risk of a bad outcome in the absence of treatment, even a small risk of harm from treatment can result in the treatment doing more harm than good.

Heterogeneity of baseline risk (and benefits), like that demonstrated in Exhibit 1, is common in large clinical trials. Indeed, many researchers have proposed that heterogeneity in the study population is advantageous because it makes the trial's results more applicable for usual clinical practice. In this paper we demonstrate that conventional analysis of clinical trials does not adequately address the benefits and safety of medical treatments in heterogeneous patient populations, but that risk-stratified analysis can often detect such differences, yet is rarely done.

## Conventional Analysis Versus Risk Stratification

■ **Theory and calculations.** When situations such as that shown in Exhibit 1 (bottom panel) occur, how likely is it that conventional analysis will clearly identify those who receive little to no benefit? Exhibit 2 presents a hypothetical circumstance under which conventional analysis would be expected to perform well: (1) overall statistical power is good, (2) individual risk factors have relatively large independent effects (odds ratios = 2), (3) treatment benefits vary greatly between higher- and lower-risk patients, and (4) the risk factor being examined is common.<sup>5</sup> Yet even under such favorable circumstances, there is less than a one in five chance that “one-variable-at-a-time” analysis will find that those with higher risk received more relative benefit (statistical power = 19 percent). Thus, conventional subgroup analysis will usually misleadingly portray a consistent treatment benefit across all subgroups and encourage us to conclude that all people similar to those in the study should receive this treatment.

Exhibit 2 also shows just how wrong such a conclusion would be. Let’s reevaluate this study’s results using a risk-stratified approach, which takes into account the additive effect of people having multiple risk factors simultaneously.<sup>6</sup> In the second panel of Exhibit 2 we see that by using a simple risk index, we can detect that the relative risk of treatment actually varies profoundly, from a 49 percent increase in harm for those with no risk factors to a 40 percent reduction in harm for those with four or more risk factors. With this dramatic variation in treatment benefit (and harm), it is particularly noteworthy that the conventional approach

### EXHIBIT 2 The Ineffectiveness Of Conventional Subgroup Analysis For Distinguishing Differences In The Relative Treatment Benefit In Lower- Versus Higher-Risk Patients

Probability of identifying subgroup effects if the average baseline risk in the study population is 5.6%				
For 5-year follow-up				
	True control event rate (CER)	True relative risk reduction (RRR)	True number needed to treat (NNT)	Likelihood of a statistically significant finding ( $p < .05$ )
Overall results	5.6%	21%	85	0.75 <sup>a</sup>
Conventional subgroup comparison				
Risk factor present (40% of subjects: n = 3,520)	7.8	28	46	0.19 <sup>b</sup>
Risk factor absent (60% of subjects: n = 5,280)	4.1	12	198	
Risk index <sup>c</sup>				
0 risk factors (n = 1,505)	1.6	-41	-152	0.68 <sup>b</sup>
1 risk factor (n = 3,170)	3.2	3	1,042	
2 risk factors (n = 2,673)	6.1	23	71	
3 risk factors (n = 1,152)	11.7	35	24	
4+ risk factors (n = 300)	21.9	36	13	

**EXHIBIT 2**  
**The Ineffectiveness Of Conventional Subgroup Analysis For Distinguishing Differences In The Relative Treatment Benefit In Lower- Versus Higher-Risk Patients (cont.)**

Probability of identifying subgroup effects if the average baseline risk in the study population is 3.0%				
For 5-year follow-up				
	True control event rate (CER)	True relative risk reduction (RRR)	True number needed to treat (NNT)	Likelihood of a statistically significant finding ( $p < .05$ )
Overall results	3.1%	0%	No benefit	0.034 <sup>a</sup>
Conventional subgroup comparison				
Risk factor present (40% of subjects: n = 3,520)	4.4	14	162	0.26 <sup>b</sup>
Risk factor absent (60% of subjects: n = 5,280)	2.3	-17	-263	
Risk index <sup>c</sup>				
0 risk factors (n = 1,505)	0.9	-117	-95	0.80 <sup>b</sup>
1 risk factor (n = 3,170)	1.7	-35	-168	
2 risk factors (n = 2,673)	3.4	4	735	
3 risk factors (n = 1,152)	6.6	26	58	
4+ risk factors (n = 300)	12.9	32	24	

**SOURCE:** Authors' statistical simulation of a study that has 80% statistical power to identify a relative risk reduction of >25 percent (two-tailed test;  $\alpha = .05$ ), with six independent risk factors for the primary outcome (each having an independent effect on baseline risk; odds ratio = 2); treatment is assumed to decrease the risk of major bad outcomes by 50% over a five-year period, but at a cost of three treatment-related serious bad outcomes per year per 1,000 patients treated (0.3% percent per year).

**NOTE:** Minus signs denote that treatment had net harm, rather than benefit.

<sup>a</sup> For the overall results, the statistical comparison is the probability that a study will suggest that the treatment confers benefits ( $p < .05$ ).

<sup>b</sup> For the subgroup comparisons, the statistical comparison tests whether the subgroup with the risk factor receives more or less benefit (two-tailed testing) than the subgroup without the risk factor (testing for an interaction between the risk factor and intervention [treatment versus control] in a logistic regression model). In the second panel, the conventional subgroup comparison for the risk factor with a 40 percent prevalence had a statistical power of 19 percent for detecting that those with the risk factor had a greater relative benefit from treatment than those without the risk factor. Although not shown in the exhibit, conventional comparisons for the four risk factors with a prevalence of 25 percent had statistical power of 15 percent, and the conventional comparison for the one risk factor with a prevalence of 10 percent had a statistical power of 10 percent.

<sup>c</sup> Area Under the Receiver Operator Characteristic (AUROC) curve for the Risk Index was 0.67 for the first example and 0.65 for the second example.

fails to identify subgroup differences. Conversely, in some situations in which there is no average benefit across the entire population (a “negative trial”), the poor statistical power of conventional analysis can obscure the fact that many patients get substantial benefit from treatment (Exhibit 2, second panel).

■ **Real-world examples from the medical literature.** Having demonstrated the statistical utility of this approach, let us move to asking if it really matters in some important clinical trials. The original analysis of the European Carotid Surgery Trial (ECST) found that carotid endarterectomy (CEA), a surgical procedure designed to relieve blockages of arteries in the neck, reduced the absolute risk of major stroke or death by almost 12 percent (NNT = 9). Conventional “one-variable-at-a-

time” subgroup analysis failed to identify any patient subgroup that would not benefit from this surgical procedure (in agreement with a previous study), so the authors endorsed that CEA should be recommended “for most patients with a recent nondisabling carotid TIA when the symptomatic stenosis is greater than 80%.”<sup>7</sup> However, in a landmark study, Peter Rothwell and Charles Warlow reanalyzed the ECST using a risk prediction tool. Upon reanalysis, patients with a higher risk score (baseline five-year stroke risk = 40%) received dramatic benefits from surgery (NNT = 3), but the typical patient in the study (baseline stroke risk ≈ 12%) received no benefit, and surgery resulted in net harm for many patients with low baseline risk.<sup>8</sup> Overall, they found that 16 percent of the patients, whose risk was more than three times greater than that of the remaining 84 percent of study subjects, accounted for almost all of the benefit seen in the “average” result.

The Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) trial presents another dramatic example. GUSTO found a significant decrease in mortality for acute heart attack patients who were treated with a new “clot-busting” medication (a thrombolytic medication called tPA) when compared with results for those treated with an older, less expensive “clotbuster” (streptokinase). However, risk-stratified analyses of GUSTO found dramatic variations in benefit that once again could not be identified using conventional subgroup analysis.<sup>9</sup> For example, David Kent and colleagues divided patients based on an externally validated risk/benefit stratification model that predicted (1) risk of death due to heart attack, (2) risk for brain hemorrhage (a known complication of clotbusters), and (3) relative benefit from treatment (determined by time from symptom onset to time of clot-buster administration).<sup>10</sup> They found that 25 percent of GUSTO subjects accounted for more than 60 percent of the total benefit. However, they found that half of the GUSTO population received little to no net benefit from tPA, and they also identified a group in which the risk of tPA-related brain hemorrhage exceeded tPA’s benefits.

Yet another example, similar to the “negative” trial presented in Exhibit 2, concerns the Alteplase Thrombolysis for Acute Noninterventional Therapy in Ischemic Stroke (ATLANTIS B) study. Conventional analyses for patients with ischemic stroke treated three to five hours after symptom onset could not identify any subgroup that benefited from treatment with a clot-buster medication. However, risk-stratified analyses identified a sizable subgroup of patients (at low risk for treatment-related brain hemorrhage) who received significant benefit from treatment.<sup>11</sup>

In each of the cases above, conventional subgroup analysis was unable to accurately detect variations in benefit and safety that were clinically important and readily identifiable using a risk-stratified approach. In the first two examples, interventions were promoted for people who receive little or no benefit and potential harms to identifiable low-risk subgroups were ignored, while in the latter case, the potential benefit in a sizable patient subgroup was completely missed.

These examples are particularly compelling in that they relate to treatments that are expensive and commonly used. It is noteworthy, therefore, that despite these important findings, including uncovering safety problems that were not identified in conventional subgroup analysis, subsequent clinical trials in these clinical areas (and there have been many) have generally not reported risk-stratified analysis.

■ **Current approach to reporting RCT results.** We reviewed clinical trials published in the *Journal of the American Medical Association*, the *Lancet*, or the *New England Journal of Medicine* during 2001 and identified 108 clinical trials that reported results on major patient outcomes, such as mortality or major morbidity.<sup>12</sup> Of the 108 eligible trials, 42 (39 percent) reported no subgroup analysis. Nearly all subgroup analyses reported on single patient attributes in isolation. Only four studies (4 percent) reported treatment benefit for lower- versus higher-risk patients, and only one of these studies used a robust statistical method.

## Discussion And Implications

It is well recognized that investigators almost always have a perceptual bias toward viewing their results positively. Strong professional, political, and financial incentives often amplify this predisposition. This study addresses a variant of this bias: the desire to promote a beneficial treatment for use in as many people as possible.

If we as a society are to make the best use of our health care dollars, we need to know who truly benefits from increasingly costly interventions. The current conventions for analyzing and reporting the results of clinical trials fail to provide policymakers with essential information for making such decisions. However, the evidence base for informing providers, payers, and consumers could be dramatically improved by one simple addition to conventional reporting of clinical trials: Whenever a multivariable prediction tool is available, the observed relative and absolute risk reduction for subjects with higher versus lower predicted net benefit should be reported using risk-stratified analysis. When using a validated prediction tool and robust statistical methods, this approach can represent a single a priori statistical comparison, thereby avoiding the high risk of false positive and false negative results inherent in multiple “one-variable-at-a-time” subgroup analysis. Even for small studies that have marginal statistical power, risk-stratified analysis will still be valuable for comparing results between different studies or conducting meta-analyses. When possible, prediction tools should be externally developed and validated and should be part of the prespecified a priori analysis plan. Certainly no analytic technique can fully account for all important factors (such as basic design or sample-size limitations); however, our results clearly suggest that risk-stratified analysis can detect safety problems and identify high-benefit subgroups that cannot be detected by conventional methods.

■ **Proposals for change.** Given the bias against publishing negative results, investigators might be more likely to conduct robust risk-stratified analysis when the

*“Organizations representing purchasers and consumers could help advocate for more complete reporting of medical evidence.”*

.....

overall study results show no benefit (a negative trial) than when the trial's average result is positive. For example, we found only one clinical trial published in 2001 that used an analytic approach similar to what we propose. This study, which examined a treatment for unstable coronary syndrome, used a multivariable risk-stratified analysis and found that low-risk patients did not receive substantial benefit from treatment.<sup>13</sup> It is therefore interesting that subsequent positive clinical trials examining treatments for acute coronary syndromes (including a study by the same investigator published in the same journal three years later) have not used this risk-stratified approach.<sup>14</sup> Given current incentives, we think it unlikely that most researchers will voluntarily conduct and report analyses evaluating whether low-risk subgroups do not benefit from a treatment. Journals also have an understandable bias toward reporting more positive and easy-to-understand results, which might in part explain why editorial boards have not required risk-stratified analysis. Pressure from organizations representing purchasers' and consumers' interests (such as the Leapfrog Group, National Committee for Quality Assurance, and others) and those setting guidelines for clinical trial reporting (such as the Consolidated Standards for Reporting of Trials, or CONSORT) could help advocate for more complete reporting of medical evidence.<sup>15</sup>

Given the obvious economic incentives for industry (and researchers with strong financial connections to industry) to get treatments approved for as broad a population as possible, regulatory agencies such as the U.S. Food and Drug Administration (FDA) and the U.K. National Institute for Clinical Excellence (NICE) should consider requiring risk-stratified analysis, since, as discussed above, safety problems in identifiable subgroups can be missed if we continue to rely on conventional reporting. There is at least one precedent for having the FDA require risk-stratified analysis. In 2001 a clinical trial demonstrated the efficacy of a new and expensive treatment (drotrecogin—about \$10,000–\$16,000 per patient) for people with severe life-threatening infections (sepsis). However, the FDA advisory board noted that a measure of disease severity (Acute Physiology and Chronic Health Evaluation, or APACHE, score) was collected in this study but not considered in the published analyses. When a risk-stratified analysis was later required, it was found that the 50 percent of patients with lower mortality risk (APACHE II scores less than 25) did not benefit from treatment (relative risk = 0.99 [0.75, 1.30]). As a result, this treatment was approved for use only in those with higher APACHE scores.<sup>16</sup> Criticisms of this decision often focused on the post hoc nature of this analysis, whereas it could be more appropriate to wonder why the most logical and important subanalysis was not part of the a priori analysis plan.

■ **Caveats and future study.** Although a multivariable approach is an important advance in reporting medical evidence, it has limitations. Individual risk factors can have complex and linked effects on both the benefits and risks of treatment, which calls for great care in model development and validation.<sup>17</sup> One particularly challenging question is how best to coordinate validating and updating population-specific prediction tools and facilitating their optimal use in day-to-day clinical practice. However, better information technology, especially Internet and handheld device applications, could greatly aid this effort. In addition, clinical practice and informed patient decision making can be greatly improved by even a qualitative understanding, without any mathematical calculations, that benefit is highly dependent on baseline risk.

Still, reality can present unwanted complexity. It can be far easier to deal with simple averages and artificial dichotomies. Thus, we predict that the most difficult challenge for risk-stratified analysis will come from the ways in which this approach will inevitably make decision making more challenging and nuanced. Risk-stratified analysis will make explicit that the amount of expected benefit for individuals almost always exists along a continuum, with some people residing in a range in which statistical certainty ranges from “no benefit” to “moderate benefit.” The false dichotomies of the current paradigm might be erroneous, but they are also often much more congenial for provider, patient, and policymaker alike, since they are congruent with binary decision making (to treat or not to treat).

Therefore, risk-stratified reporting could be more accurate in estimating an individual’s risks and benefits of treatment, but it also runs the risk of inducing policy paralysis by illuminating the arbitrary nature of any prespecified treatment threshold. We propose that instead of retreating to the comfort of false dichotomies, we consider intermediate steps for our policy decisions. For example, instead of deciding whether coverage of a treatment should be zero or 100 percent, we could adjust copayments based on the amount of expected benefit (instead of basing copayments on the cost of treatment alone).<sup>18</sup> Similarly, instead of considering performance measures as met or not met, we might need to consider the degree of importance of the deviation from recommended care.<sup>19</sup> Certainly, allowance for patients’ preferences should become even more important when there is greater uncertainty regarding the likely risks and benefits of treatment.

**U**NDERSTANDING HOW TREATMENT BENEFITS vary between lower- versus higher-risk patients is fundamental to optimal policy decision making, but adoption of this approach will often run counter to the incentives faced by those funding, conducting, and reporting clinical trials. The public may best be served by proactive policies advancing risk-stratified analysis rather than simply expecting researchers to adopt this approach voluntarily. Regulatory bodies and payers have a particularly strong interest in advancing risk-stratified assessments of medical evidence, to decrease the chances that incomplete analyses

mislead us to extending expensive and burdensome treatments to those who may derive no benefit—or worse, harm.

.....  
*The authors thank Adam Tremblay for conducting the literature review and Joel Howell, Joy Pritts, and two anonymous reviewers for their comments on earlier drafts. This work was supported in part by Veterans Affairs (VA) Research and Development, the VA Health Services Research and Development Service (Grant no. QUERI DIB 98-001); and the VA Cooperative Studies Program (Grant no. CSP #465), with additional support being provided by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (Grant no. P60 DK-20572). David Kent is supported by a Career Development Award from the National Institute for Neurological Disorders and Stroke (Grant no. K23 NS44929-01). The above views and opinions are those of the authors and do not necessarily reflect those of the U.S. Department of Veterans Affairs or the University of Michigan.*

## NOTES

1. R. Steinbrook, "Registration of Clinical Trials—Voluntary or Mandatory?" *New England Journal of Medicine* 351, no. 18 (2004): 1820–1822.
2. Multivariable risk prediction models/tools use statistical formulas that estimate individual risk based on combining information from multiple independent risk factors. A risk prediction model or tool could estimate risk in the absence of treatment (baseline disease risk), risk of adverse effects of treatment (treatment risk), or modifiers of treatment effects (differential relative treatment benefit). Prediction models can be presented as a full regression prediction model (such as predicted probability of death using APACHE), as a simple risk index (such as predicting birth outcomes using a ten-point APGAR score), or condensed into risk categories (low, medium, high perioperative risk). With the common use of computers and handheld devices in medical practice, even mathematically complex risk models can be made transparent and user-friendly for clinicians. Risk-stratified subgroup analysis is an approach to subgroup analysis of clinical trials that examines variation in a treatment's relative and absolute benefit (or harm) across the study population as a function of predicted risks and benefits (using a multivariate risk prediction tool).
3. W.S. Browner, "Willy Sutton and the Number Needed to Treat," *American Journal of Medicine* 116, no. 8 (2004): 564–565; and D.K. Owens et al., "Effect of Risk Stratification on Cost-Effectiveness of the Implantable Cardioverter Defibrillator," *American Heart Journal* 144, no. 3 (2002): 440–448.
4. E.S. Fisher and H.G. Welch, "Could More Health Care Lead to Worse Health?" *Hospital Practice* 34, no. 12 (1999): 15–25.
5. The calculations of the reliability and accuracy of subgroup analyses in Exhibit 2 are based on standard simulation techniques for calculating statistical power. For each iteration, a random sample was generated using the specified study characteristics shown. Two thousand iterations of the study sample size ( $N = 8,800$ ) were conducted for each scenario to achieve precise estimates. For each of the 2,000 randomly generated study samples, we conducted logistic regression analyses, with occurrence of bad outcome ( $1 = \text{yes}$ ,  $0 = \text{no}$ ) as the dependent variable. The overall treatment effect was tested using treatment arm ( $2 = \text{treated}$ ,  $1 = \text{control}$ ) as the independent variable. Significant subgroup effects were tested by examining interaction effects between treatment and a single risk factor ( $2 = \text{yes}$ ,  $1 = \text{no}$ ) for the conventional analysis and testing interactions between treatment and the full risk index ( $0-6$ ) for the multivariable risk-stratified analyses. The results reported in Exhibit 2 are the average results for the 2,000 iterations.
6. P.M. Rothwell, "Can Overall Results of Clinical Trials Be Applied to All Patients?" *Lancet* 345, no. 8965 (1995): 1616–1619; and S. Vijan, D.M. Kent, and R.A. Hayward, "Are Randomized Controlled Trials Sufficient Evidence to Guide Clinical Practice in Type II (Non-Insulin-Dependent) Diabetes Mellitus?" *Diabetologia* 43, no. 1 (2000): 125–130.
7. European Carotid Surgery Trialists' Collaborative Group, "Randomised Trial of Endarterectomy for Recently Symptomatic Carotid Stenosis: Final Results of the MRC European Carotid Surgery Trial (ECST)," *Lancet* 351, no. 9113 (1998): 1379–1387.
8. P.M. Rothwell and C.P. Warlow, "Prediction of Benefit from Carotid Endarterectomy in Individual Patients: A Risk-Modelling Study," *Lancet* 353, no. 9170 (1999): 2105–2110.
9. D.M. Kent et al., "An Independently Derived and Validated Predictive Model for Selecting Patients with

- Myocardial Infarction Who Are Likely to Benefit from Tissue Plasminogen Activator Compared with Streptokinase,” *American Journal of Medicine* 113, no. 2 (2002): 104–111; and R.M. Califf et al., “Selection of Thrombolytic Therapy for Individual Patients: Development of a Clinical Model,” *American Heart Journal* 133, no. 6 (1997): 630–639.
10. Kent et al., “An Independently Derived and Validated Predictive Model.”
  11. D.M. Kent, R. Ruthazer, and H.P. Selker, “Are Some Patients Likely to Benefit from Recombinant Tissue-Type Plasminogen Activator for Acute Ischemic Stroke Even Beyond Three Hours from Symptom Onset?” *Stroke* 34, no. 2 (2003): 464–467.
  12. To examine the current approach to reporting clinical trial results, we reviewed trials published in the three journals during 2001. In a PubMed search, 263 articles were identified. Review of abstracts identified 155 studies as being ineligible: 28 were not true experiments, and 127 did not include mortality or major morbidity as a main outcome. The initial review of eligibility was blinded to the presence and results of subgroup analysis. A 10 percent hand review of hard-copy journals found no additional clinical trials unidentified by the PubMed search. The reviewers read through the entire methods and results sections of each clinical trial and answered questions related to the analysis and reporting of main results and subgroup comparisons. The primary reviewer knew that 20 percent of studies would undergo duplicate reviews but was blinded with regard to which specific studies had been selected for duplicate review.
  13. C.P. Cannon et al., “Comparison of Early Invasive and Conservative Strategies in Patients with Unstable Coronary Syndromes Treated with the Glycoprotein IIb/IIIa Inhibitor Tirofiban,” *New England Journal of Medicine* 344, no. 25 (2001): 1879–1887.
  14. C.P. Cannon et al., “Intensive versus Moderate Lipid Lowering with Statins after Acute Coronary Syndromes,” *New England Journal of Medicine* 350, no. 15 (2004): 1495–1504.
  15. D. Moher, K.F. Shulz, and D.G. Altman, “Revised Recommendations for Improving the Quality of Reports of Parallel Group Randomized Trials 2001,” April 2001, [www.consort-statement.org/Statement/revised\\_statement.htm](http://www.consort-statement.org/Statement/revised_statement.htm) (19 October 2004).
  16. U.S. Food and Drug Administration, “Drotrecogin alfa (Activated),” September 2003, [www.fda.gov/cder/biologics/products/drotelil12101.htm](http://www.fda.gov/cder/biologics/products/drotelil12101.htm) (1 September 2005).
  17. D.G. Altman and P. Royston, “What Do We Mean by Validating a Prognostic Model?” *Statistics in Medicine* 19, no. 4 (2000): 453–473.
  18. A.M. Fendrick et al., “A Benefit-based Copay for Prescription Drugs: Patient Contribution Based on Total Benefits, Not Drug Acquisition Cost,” *American Journal of Managed Care* 7, no. 9 (2001): 861–867.
  19. R.A. Hayward et al., “Quality Improvement Initiatives: Issues in Moving from Diabetes Guidelines to Policy,” *Diabetes Care* 27, no. 2 Supp. (2004): B54–B60.