Journal of the American College of Cardiology
© 2010 by the American College of Cardiology Foundation
Published by Elsevier Inc.

Vol. 55, No. 5, 2010
ISSN 0735-1097/10/$36.00
doi:10.1016/j.jacc.2009.06.065

STATE-OF-THE-ART PAPER AND COMMENTARY

Trial and Error

How to Avoid Commonly Encountered Limitations of Published Clinical Trials

Sanjay Kaul, MD, George A. Diamond, MD

Los Angeles, California

The randomized controlled clinical trial is the gold standard scientific method for the evaluation of diagnostic and treatment interventions. Such trials are cited frequently as the authoritative foundation for evidence-based management policies. Nevertheless, they have a number of limitations that challenge the interpretation of the results. The strength of evidence is often judged by conventional tests that rely heavily on statistical significance. Less attention has been paid to the clinical significance or the practical importance of the treatment effects. One should be cautious that extremely large studies might be more likely to find a formally statistically significant difference for a trivial effect that is not really meaningfully different from the null. Trials often employ composite end points that, although they enable assessment of nonfatal events and improve trial efficiency and statistical precision, entail a number of shortcomings that can potentially undermine the scientific validity of the conclusions drawn from these trials. Finally, clinical trials often employ extensive subgroup analysis. However, lack of attention to proper methods can lead to chance findings that might misinform research and result in suboptimal practice. Accordingly, this review highlights these limitations using numerous examples of published clinical trials and describes ways to overcome these limitations, thereby improving the interpretability of research findings. (J Am Coll Cardiol 2010;55:415–27) © 2010 by the American College of Cardiology Foundation

The randomized controlled trial (RCT) is the apotheosis of scientific progress in clinical medicine. Such trials are key drivers of modern cardiovascular practice, since they are cited frequently as the authoritative foundation for evidence-based management policies. Although the randomization process minimizes the imbalance in measured and unmeasured confounding variables, thereby allowing one to infer causation, not just association, a number of limitations nevertheless serve to challenge the interpretation of the results. Three particular technical limitations are the subject of this review. First, strength of evidence is often judged by conventional tests that rely heavily on statistical significance and estimation of confidence intervals (CIs). Less attention has been paid to the clinical significance or the practical importance of the treatment effects. Second, composite end points are often used to increase the proportion of outcome events and thereby reduce requisite sample size. Although this practice improves trial efficiency and statistical precision, it entails a number of shortcomings that can undermine the scientific validity of the conclusions drawn from these trials. Finally, additional exploratory subgroup analyses are frequently performed without sufficient attention to the reliability of these subordinate analyses. This leads to the reporting of chance findings that encourage suboptimal patterns of practice. In this review, we highlight each of these limitations using numerous examples of published clinical trials and propose practical ways to avoid them and help improve the interpretation of the published findings.

From the Division of Cardiology, Cedars-Sinai Medical Center, and the David Geffen School of Medicine, University of California, Los Angeles, California. Dr. Diamond has served on the Speaker's Panel for Merck and Schering-Plough.

Manuscript received January 25, 2009; revised manuscript received June 1, 2009, accepted June 1, 2009.

Emphasize Clinical Importance Over Statistical Significance

The conventional approach to assessing the strength of the association between an intervention and outcome (the evidence) focuses on p values and CI (1–4). A p value or observed significance level provides a measure of the inconsistency of the data with respect to a specific hypothesis. In clinical trials, investigators pre-specify a significance level (most commonly 0.05) that represents the maximum probability they will tolerate of rejecting a hypothesis when it is in fact true. Some have suggested that p values provide a measure of the strength of the evidence against the null hypothesis; the smaller the p value, the stronger the evidence against the null hypothesis. For example, Sterne and Smith (3) suggest that a p value of 0.05 need not provide strong evidence against the null hypothesis, but it is reasonable to say that a p value <0.001 does. In contrast, others have cautioned that because p values are dependent on sample size, a p value of 0.001 should not be interpreted as

JACC Vol. 55, No. 5, 2010
February 2, 2010:415–27

Abbreviations and Acronyms

ACS = acute coronary syndrome
CABG = coronary artery bypass graft surgery
CI = confidence interval
LMCA = left main coronary artery
MACE = major adverse cardiac event
MCID = minimum clinically important difference
MI = myocardial infarction
NNT = number needed to treat
PCI = percutaneous coronary intervention
RCT = randomized controlled trial
TVR = target vessel revascularization

providing more support for rejecting the null hypothesis than one of 0.05 (4).

In contrast to the well-established standards for decisions regarding statistical significance, no particular guidelines exist for deciding what magnitude of difference is "clinically significant" or "practically important" (1–6). The latter decision depends upon the quantitative magnitude of the treatment effect and the associated context—the seriousness and the frequency of the outcome of interest and the benefit-risk-cost profile. Because of the inherently subjective and context-specific nature of these judgments, investigators have understandably been reluctant to establish fixed boundaries for what constitutes a clinically significant difference.

An unintended consequence of this lack of an established standard has been an erroneous tendency to equate statistical significance with clinical significance. In some instances, statistically significant results may not be clinically important (e.g., small differences in studies with large sample size), and conversely, statistically insignificant results do not completely rule out the possibility of clinically important effects (e.g., large differences in studies with small sample size) (6). Ideally, assessment of both statistical and clinical significance should be used to appraise the strength of the evidence and to aid in optimal utilization of therapeutic interventions in clinical practice.

Consider the TACTICS–TIMI 18 (Treat Angina With Aggrastat and Determine Cost of Therapy With an Invasive or Conservative Therapy–Thrombolysis In Myocardial Ischemia/Infarction 18) trial, a randomized trial of early invasive versus early conservative management of patients with acute coronary syndromes (ACS) (7). In designing the trial, the investigators powered the study to detect a 25% relative risk reduction, presumably representing their estimate of a minimum clinically important difference (MCID) in outcome. Upon conducting this trial, a total of 177 events (15.9%) were observed among 1,114 patients assigned to early invasive management versus 215 events (19.4%) among 1,106 patients assigned to early conservative management (7). The relative risk reduction for this 3.5% absolute difference was 18% (95% CI: 2% to 32%), and this was determined to be statistically significant ($p = 0.028$). The investigators thereby concluded that early invasive management is superior to early conservative management. However, the key question that the thoughtful clinician is interested in is, "What is the probability that early invasive management is associated with a 'clinically important' benefit over early conservative management?" Simply stated, is the 18% risk reduction observed in this study "clinically important"?

Sackett (6) has proposed the use of confidence intervals to answer this question. According to this approach, if the summary treatment effect is large enough to exclude values smaller than the MCID, and not just the null value of zero, then the treatment provides both a statistically and clinically significant benefit. Using this approach, Figure 1A classifies treatment effects into the following categories: "statistically not significant and clinically not important" (example A, where the entire CI lies to the right of the MCID and crosses the null line, thus ruling out any important benefit),

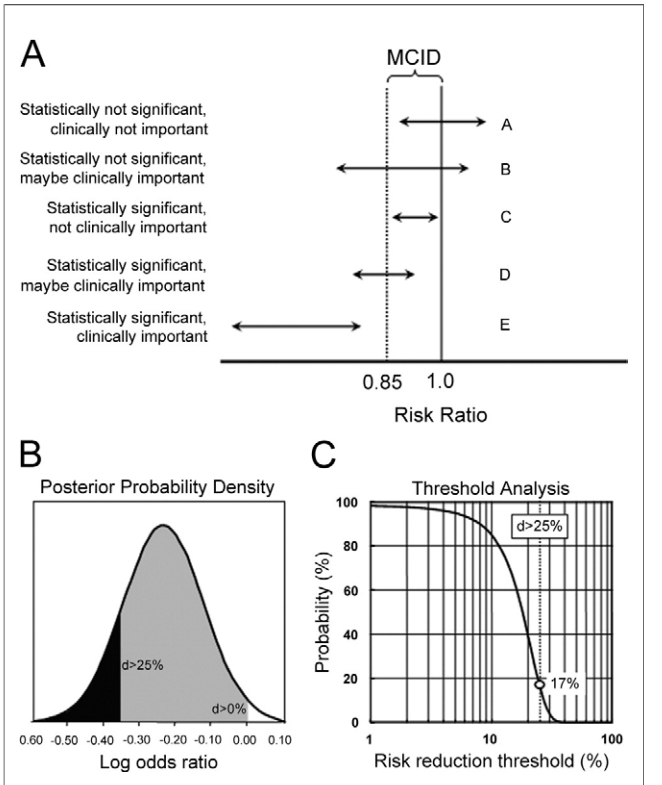

Figure 1 **Statistically Significant and Clinically Important Treatment Benefits, and Bayesian Analysis of Clinical Importance in the TACTICS–TIMI 18 Trial**

(A) Graphic demonstration of statistically significant and clinically important treatment benefits. Five trial results **(A to E)** and their interpretation with reference to zero effect (a risk ratio of 1.0) and a minimal clinically important difference (MCID) of 15% relative risk reduction (corresponding to a risk ratio of 0.85) are shown. Treatment effects **(double arrows)** are expressed as 95% confidence intervals. **(B)** Bayesian analysis of clinical importance in TACTICS–TIMI 18: Probability density plot for the difference in outcomes between the 2 management strategies is shown using the Bayesian approach (a noninformative Gaussian prior probability distribution with mean = 0 and SD = 10). The probability for any given threshold (d) can be computed in terms of the area under the probability density curve. The **light shaded area** to the left of log odds ratio 0 (equivalent to odds ratio of 1) indicates a 98.6% probability of d >0% benefit and the **dark shaded area** to the left of log odds ratio −0.35 (equivalent to odds ratio of 0.71 or relative risk of 0.75) indicates a 17% probability of d >25% benefit. **(C)** Posterior probabilities for a range of threshold of benefit are graphically represented. The probability of d >25% risk reduction is shown as an **open circle** (17%).

"statistically not significant but may be clinically important" (example B, where the CI crosses the MCID as well as the null lines consistent with an indeterminate effect), "statistically significant but not clinically important" (example C, where the entire CI lies to the right of the MCID but does not cross the null line), "statistically significant and may be clinically important," (example D, where the CI is centered to the left of the MCID and does not cross the null), and "statistically significant and clinically important" (example E, where the entire CI lies to the left of the MCID). Most clinical trials are only large enough to generate C- or D-type results, and E-type results are only achieved in meta-analyses of several randomized clinical trials (RCTs) (6). According to this schema, the treatment benefit observed with early invasive versus early conservative management in the TACTICS–TIMI 18 study is statistically significant and may be clinically important because the 95% CI contains the MCID of 25% relative risk reduction (example D), but is not assured, because all values less than the MCID are not excluded with sufficient confidence.

Another index, the number needed to treat (NNT) or the number needed to harm, has also been used to assess clinical importance (8). These indexes and their 95% CIs are calculated as the inverse of the absolute risk differences. The pros and cons of such indexes have been reviewed previously (8). In general, treatment interventions are deemed to be clinically important when the number needed to harm is >>NNT or when the NNT is <50. However, these judgments are context-specific (with respect to disease and outcome severity) and influenced by the benefit-risk-cost profile of intervention and the duration of follow-up.

Alternatively, we have advocated the use of Bayesian analysis to estimate probabilities for a range of clinically important treatment effects (1). Briefly, such inferences are predicated on Bayes' theorem, which postulates that the posterior probability of any given hypothesis is directly related to its prior probability based on previous knowledge and the empirical evidence generated from within the study. Using a noninformative prior, indicating that all treatment effects are equally likely—an essentially flat distribution—permits the posterior to be determined entirely by the study data, as in the classical frequentist analysis. Unlike the frequentist approach, however, the Bayesian approach allows one to specify the probability for any given threshold of clinical importance (1,5,9). Although CI is often interpreted in this manner (as in Sackett's approach [6]), the correct interpretation of a 95% CI from a frequentist perspective is that 95% of all CI limits derived from an unlimited number of repeated experiments would contain the true parameter. It does not actually ascribe any probability to the value of the parameter itself (5).

As illustrated in Figure 1B, Bayesian analysis computes the probability for any given threshold in terms of the area under the probability density curve (1,5,9), and can display this probability graphically across a range of such thresholds as shown in Figure 1C. The results indicate a 98.6% chance that the risk reduction is >0 (1 minus the 1-sided p value of 0.028/2)—but only a 17% chance it is >25% (the MCID threshold). Moreover, the posterior probability for benefit falls as the threshold for benefit increases (85% chance of >10%, 67% chance of >15%, and 40% chance of >20% risk reduction). Thus, although a conventional frequentist analysis shows that early invasive management is associated with a statistically significant reduction in outcome, Bayesian analysis helps clarify whether the benefit is clinically important. Given these posterior probabilities and the important side effects of cost and bleeding, some clinicians might opt not to use an invasive strategy, despite its statistically significant benefit. In contrast, if invasive management were inexpensive and safe, clinicians might still decide to use it, even though there is a <95% chance that it provides an important magnitude of benefit. Thus, Bayesian analysis provides a straightforward, patient-, physician-, and context-specific statement of clinical importance and complements the frequentist analysis in improving the interpretation of the data and informing clinical decision making (1,5,9).

Table 1 compares the statistical significance and clinical importance of treatment interventions for ACS based on the results of RCTs and meta-analyses (10–14). A relative reduction of >15% in recurrent adverse events was considered clinically important (15). Statistically significant differences were observed for 3 of 5 interventions—aspirin being the only treatment intervention providing both statistically significant and clinically important benefits. Two examples highlight a disconnect between statistical significance and clinical importance. Whereas treatment with unfractionated heparin was not statistically significant despite a large risk reduction (owing to relatively small sample size) (11), the probability of >15% risk reduction was 87%, consistent with example B in Figure 1A (statistically not significant but may be clinically important). In contrast, although treatment with platelet glycoprotein IIb/IIIa inhibitor was statistically significant despite a modest risk reduction (due to large sample size) (13), the probability of >15% risk reduction was only 4%, consistent with example C in Figure 1A (statistically significant but not clinically important).

In summary, while statistical significance tells us whether a difference is likely to be real, it does not place that reality into a meaningful clinical context by telling us the difference is small or large, trivial or important. A formal evaluation of clinical importance (using frequentist confidence intervals, the NNT and the number needed to harm indexes, or Bayesian probabilities), given the overall risk-benefit-cost profile of each therapeutic intervention, should be included in the analysis, interpretation, and presentation of the results of clinical trials. To this end, we suggest that explicit standards of evidence be developed that encourage not only robust trial design and statistical methodology but also emphasize the explicit assessment of clinical importance relative to some pre-defined MCID threshold.

JACC Vol. 55, No. 5, 2010
February 2, 2010:415–27

Table 1 Statistical Versus Clinical Significance

Intervention	Control (%)	Rx (%)	Summary Risk Ratio (95% CI)	p Value	NNT (95% CI)	Pr (d ≥MCID)	Interpretation of Confidence Intervals
1. ASA vs. placebo (n = 2,856) (10,11)	12.8	5.5	0.43 (0.33–0.56)	<0.01	14 (11–19)	100%	Statistically significant and clinically important (E)
2. ASA + UFH vs. ASA (n = 1,353) (11)	10.4	7.9	0.67 (0.44–1.02)	0.06	44 (∞–18)	87%	Statistically not significant but may be clinically important (B)
3. ASA + UFH + clopidogrel vs. ASA + UFH (n = 12,562) (12)	11.4	9.3	0.82 (0.74–0.92)	<0.01	54 (35–120)	76%	Statistically significant and may be clinically important (D)
4. ASA + UFH + GPI vs. ASA + UFH (n = 27,051) (13)	11.8	10.5	0.91 (0.86–0.99)	0.012	73 (48–157)	4%	Statistically significant but not clinically important (C)
5. ASA + clopidogrel + UFH vs. ASA + clopidogrel + enoxaparin (n = 10,027) (14)	14.5	14.0	0.96 (0.86–1.06)	0.43	184 (∞–52)	1%	Statistically not significant and clinically not important (A)

Summary risk ratios are derived from random-effects meta-analysis except for #3 and #5, which are based on the CURE trial and the SYNERGY trial, respectively. Posterior probabilities (Pr) are derived from Bayesian analysis, and interpretation of confidence intervals (CIs) is based on schema described in Figure 1A.

ASA = aspirin; GPI = glycoprotein IIb/IIIa inhibitor; MCID = minimum clinically important difference of 15% relative risk difference; NNT = number needed to treat; Rx = treatment; UFH = unfractionated heparin.

Toward this end, it might seem convenient to define this threshold as the minimum detectable difference employed in the design of the trial. Unfortunately, however, the minimum detectable difference is often selected by trialists on purely pragmatic grounds such as financial constraints, restrictions in available candidates, and limitations in follow-up duration—allowing the design of a trial with frugal sample size requirements—and does not necessarily reflect the MCID from the perspective of the practitioner or the patient. The optimal thresholds of clinical importance might well vary from disease to disease, treatment to treatment, outcome to outcome, physician to physician, and patient to patient. Thus, lower thresholds of importance might attach to the reduction in mortality or serious irreversible morbid events such as Q-wave myocardial infarction (MI) or stroke versus higher thresholds for reduction in reversible and less serious morbid events such as recurrent hospitalization, asymptomatic periprocedural troponin elevation, or refractory ischemia. Ideally, assessment of both statistical significance and clinical importance should aid in optimal utilization of therapeutic interventions in clinical practice.

Employ Appropriate Composite End Points

Composite end points are measurable events that lie on a pathophysiologic spectrum and allow for parsimonious summarization of treatment effects, in other words, they can sensibly be added together as being aspects of the same underlying biologic process to quantify the overall treatment effect. Composite end points are frequently used in clinical trials (16–22), with 1 recent survey reporting that 37% of the 1,231 trials published over 7 years used composite outcomes (21). Such use reduces the sample size and cost requirements of clinical trials and is thereby thought to improve trial efficiency and help facilitate the formal evaluation and ultimate availability of effective new treatments.

The typical cardiovascular trial combines "hard" but infrequent end points such as death, Q-wave MI, disabling stroke, and emergency coronary artery bypass graft surgery (CABG) with "soft" but more frequent end points such reintervention, periprocedural MI (e.g., biomarker elevation), recurrent angina, and rehospitalization. Because of their greater frequency, these less important disparate outcomes often drive the effect of therapy on the composite. A systematic review of 114 cardiovascular trials that used composite end points reported a moderate to large gradient in the hierarchy of clinical importance of component events in nearly 40% of the trials (19). Of the 27 trials that reported a statistically significant difference in the composite outcomes, only 7 were driven by hard outcomes.

Major adverse cardiac events (MACE) are arguably the most commonly used composite end point in cardiovascular research (22). There is no consensus definition of MACE, yet its use has become virtually pervasive in cardiovascular research in the last 2 decades. At the broadest level, definitions of MACE in use today include end points that reflect both the safety (death, MI, stroke) and effectiveness (target vessel revascularization [TVR], restenosis, recurrent ischemia, rehospitalization) of various treatment approaches. Three recent literature reviews revealed that although death and MI were included in most definitions of MACE (19,21,22), inclusion of the remaining components was highly variable, thereby contributing to significant heterogeneity across trials. Even the definition of MI was not consistent. Very few trials use the more reliable Q-wave criterion versus the less reliable non–Q-wave and/or cardiac biomarker criterion to define MI (22). Such use opens investigators and sponsors to the charge of "gaming" their trials by inflating the number of (arguably unimportant) outcome events, thereby increasing statistical power (22). Because varying definitions of composites such as MACE lead to substantially different conclusions, some have called for a reappraisal of their use (22).

The construction of the composite end point is generally based on the premise that each component end point is interchangeable. However, for this assumption to be valid, 3

criteria need to be fulfilled (18–21): 1) each component should be of comparable clinical importance; 2) each component should occur with similar frequency; and 3) each component should be similarly sensitive to treatment intervention. All 3 criteria are seldom fulfilled.

Figure 2 illustrates the interpretation of composite outcomes in 3 typical cardiovascular trials. The Stent PAMI (Stent–Primary Angioplasty for Myocardial Infarction) trial was designed to assess whether primary stenting was superior to primary balloon angioplasty in patients with acute ST-segment elevation MI (23). The combined primary end point of death, reinfarction, disabling stroke, or ischemia-driven TVR at 1 year occurred in fewer patients in the stent group than in the angioplasty group (12.6% vs. 20.1%, $p < 0.005$) (23). Examination of the data in Figure 2A reveals that the treatment differences were primarily attributable to differences in TVR ($p < 0.0005$), the most prevalent but the least important component, with little or no impact on reinfarction ($p = 0.7$) or stroke ($p = 0.83$) and death being affected negatively ($p = 0.07$). A formal assessment revealed heterogeneity of treatment effect across components of the composite end point ($p = 0.002$). Thus, there was a large gradient in clinical importance, prevalence, and treatment effect across individual components. This trial highlights why combining end points with varying pathophysiology such as TVR (restenosis) and mortality (restenosis rarely leads to death) may not be advisable.

In contrast, no large gradient in clinical importance or treatment effect across components of the triple composite end point of cardiovascular death, nonfatal MI, or stroke was observed in the HOPE (Heart Outcomes Prevention Evaluation) trial that compared ramipril versus placebo for 9,297 patients at high risk of cardiac events (24) (Fig. 2B), thereby supporting the credibility and validity of the composite end point.

These caveats become even more challenging in instances where efficacy end points are combined with safety end points to assess "net clinical benefit." In such cases, the focus on net clinical benefit has the potential of masking an increase in harmful effect, particularly when the offsetting end points do not have a similar clinical impact. One such example is shown in Figure 2C. In the TRITON–TIMI 38 (Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition With Prasugrel–Thrombolysis In Myocardial Infarction 38) trial, 13,608 patients with ACS were randomly assigned to prasugrel, a new thienopyridine, or to clopidogrel (25). The benefit of prasugrel in the TRITON–TIMI 38 trial was driven by nonfatal MI—nearly one-half of which were periprocedural biomarker elevations of questionable clinical relevance—with little or no impact on all-cause death or nonfatal stroke; however, non-CABG Thrombolysis In Myocardial Infarction (TIMI) major bleeding—arguably more serious than the biomarker elevations—was significantly increased. A more complete accounting of bleeding by including TIMI minor and minimal bleeding in the composite end point would likely have an important influence on the analysis of net clinical benefit (25,26).

The unconventional use of a composite efficacy and safety outcome poses even greater challenges in the assessment of noninferiority. Typically, the noninferiority claim is confined to efficacy alone. Although combining efficacy and safety into 1 composite outcome inflates the event rate and thereby enhances trial feasibility, it can often be misleading because drugs that are relatively ineffective but safer can be made to appear as good as or even better than effective drugs (27). This is illustrated in the REPLACE-2 (Randomized Evaluation in Percutaneous Coronary Intervention Linking Angiomax to Reduced Clinical Events) and the ACUITY (Acute Catheterization and Urgent Intervention Triage Strategy) trials, in which the difference in major bleeding events (statistically significant 43% and 47% relative reductions, respectively, in favor of bivalirudin) exceeded the difference in MI (statistically nonsignificant 13% and 9% relative increase, respectively), thereby biasing the assessment of noninferiority in favor of bivalirudin compared with its active comparator (27).

Table 2 summarizes the results of 6 RCTs in which the primary outcome of interest was a composite end point. In 2 trials, there was no significant heterogeneity in treatment effect across the components (7,28). In contrast, there was significant heterogeneity in 4 trials (25,29–31), with the reduction in the composite end point being typically driven by the most prevalent and arguably the least important component (29–31). In 2 trials, death, the most important end point, was affected adversely (30,31).

Heterogeneity in treatment effects across the component events has important regulatory implications. For example, the composite end point in the evaluation of losartan in the LIFE (Losartan Intervention for End Point Reduction in Hypertension) study was driven by the impact on nonfatal stroke only. Thus, the subsequent regulatory labeling of losartan was restricted to prevention of nonfatal stroke and not the original claim for the triple end point.

Inappropriate use of composite end points sometimes leads to an unfounded illusion of benefit. The use of the MACE composite end point in the bare-metal stent versus sirolimus-eluting stent trial SIRIUS (Sirolimus-Eluting Balloon Expandable Stent in the Treatment of Patients With De Novo Native Coronary Artery Lesions) (31) (Table 2) could erroneously lead one to conclude that the sirolimus-eluting stent is significantly better at reducing death, MI, stent thrombosis, and target lesion revascularization in totality, even though the statistically significant effect on MACE is driven primarily by a reduction in target lesion revascularization alone. The potential for misleading conclusions depending on the study-specific definition of MACE is not trivial (22).

The common statistical approach of using equal weights to combine disparate constituent components of a composite end point is decidedly counterintuitive. A potential solution to this problem is to assign meaningful weights to

JACC Vol. 55, No. 5, 2010
February 2, 2010:415–27

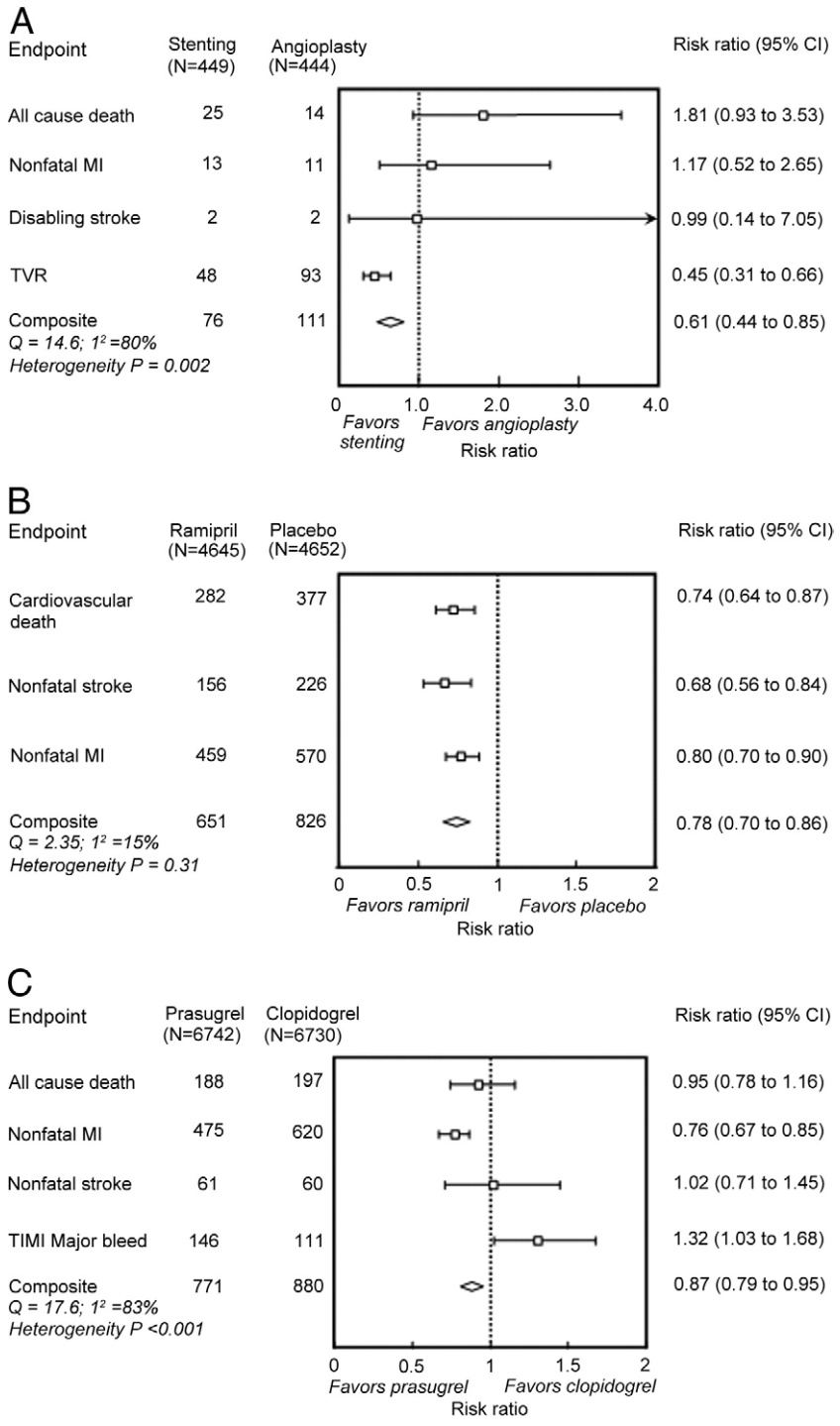

Figure 2 **Composite End Point Analysis for 3 Trials: Stent PAMI, HOPE, and TRITON–TIMI 38**

Data for the composite end point and the individual component end points are shown for 3 trials: **(A)** Stent PAMI, **(B)** HOPE, and **(C)** TRITON–TIMI 38. Heterogeneity is derived from Cochran's Q chi-square test and is implied at $p < 0.10$; I^2 quantifies the impact of heterogeneity. Data for stent PAMI at 1 year of follow-up (23). CI = confidence interval; MI = myocardial infarction; TVR = target vessel revascularization.

the components. Such approaches, however, can be highly subjective, thereby lending themselves to intellectual gerrymandering (19,27,32,33). Thus, these weights need to be prospectively agreed upon based on clinical and statistical considerations that impact the power of the test and consequently the size of the trial.

Table 2 Composite End Point Analysis

	Treatment	Control	OR	95% CI	Heterogeneity p Value	I^2
TACTICS–TIMI 18 (early invasive vs. conservative management in ACS) (7)						
Composite	177/1,114 (15.9%)	215/1,106 (19.4%)	0.78	0.63–0.97	0.34	6%
Death	37/1,114 (3.3%)	39/1,106 (3.5%)	0.94	0.59–1.49		
MI	44/1,114 (3.9%)	66/1,106 (5.9%)	0.65	0.44–0.96		
RI	96/1,114 (8.6%)	110/1,106 (9.9%)	0.85	0.64–1.14		
PRISM-PLUS (tirofiban vs. placebo in ACS) (28)						
Composite	100/773 (12.9%)	143/797 (17.9%)	0.68	0.52–0.90	0.29	17%
Death	15/773 (1.9%)	15/797 (1.9%)	1.03	0.50–2.13		
MI	23/773 (3.0%)	51/797 (6.4%)	0.45	0.27–0.74		
RI	62/773 (8.0%)	77/797 (9.7%)	0.82	0.57–1.16		
LIFE (losartan vs. atenolol) (29)						
Composite	508/4,605 (11.0%)	588/4,588 (12.8%)	0.84	0.74–0.96	0.03	72%
CV death	204/4,605 (4.4%)	234/4,588 (5.1%)	0.86	0.71–1.05		
Stroke	232/4,605 (5.1%)	309/4,588 (6.7%)	0.73	0.62–0.88		
MI	198/4,605 (4.3%)	188/4,588 (4.1%)	1.05	0.86–1.29		
TIME (invasive versus medical therapy for elderly patients with chronic symptomatic CAD) (30)						
Composite	40/153 (26.1%)	96/148 (64.9%)	0.19	0.12–0.31	<0.001	93%
Death	13/153 (8.5%)	6/148 (4.1%)	2.20	0.81–5.94		
MI	12/153 (7.8%)	17/148 (11.5%)	0.66	0.30–1.43		
Hospitalization	15/153 (9.8%)	73/148 (49.3%)	0.11	0.06–0.21		
SIRIUS (SES vs. BMS) (31)						
Composite	46/533 (8.6%)	110/525 (20.9%)	0.42	0.29–0.61	<0.0001	90%
Death	5/533 (0.9%)	3/525 (0.6%)	1.65	0.39–6.93		
MI	15/533 (2.8%)	17/525 (3.2%)	0.87	0.43–1.75		
Target lesion revascularization	22/533 (4.1%)	87/525 (16.6%)	0.22	0.13–0.35		
ST	4/533 (0.8%)	2/533 (0.4%)	2.01	0.37–11.04		
TRITON–TIMI 38 (prasugrel vs. clopidogrel in ACS) (25)						
Composite	644/6,813 (9.5%)	786/6,795 (11.6%)	0.80	0.71–0.89	0.05	67%
CV death	115/6,813 (1.7%)	112/6,795 (1.7%)	1.02	0.79–1.33		
Nonfatal MI	471/6,813 (6.9%)	618/6,795 (9.1%)	0.74	0.66–0.84		
Nonfatal stroke	58/6,813 (0.8%)	56/6,795 (0.8%)	1.03	0.71–1.49		

Heterogeneity is derived from Cochran's Q chi-square test and is implied at $p < 0.10$; I^2 quantifies the impact of heterogeneity. Data for the TRITON–TIMI 38 trial are first events (excludes double counting).
ACS = acute coronary syndromes; BMS = bare-metal stent; CAD = coronary artery disease; CV = cardiovascular; MI = myocardial infarction; RI = refractory ischemia; SES = sirolimus-eluting stent; ST = stent thrombosis; STEMI = ST-segment elevation myocardial infarction.

Such an approach is illustrated in Figure 3 for the primary efficacy end point in the TRITON–TIMI 38 trial. Here, death was given the highest weight of 1.0, followed by an intermediate weight for stroke (0.5). Myocardial infarction was given a weight of 0.06 on the basis of the case fatality rate for MI (74 of a total of 1,609 MIs were fatal) (25). The data show no statistically significant difference between the 2 groups with respect to this weighting. Varying the weight of stroke from 0.1 to 1.0 had no material impact on the results. A sensitivity analysis demonstrates a statistically significant difference in favor of prasugrel only at an MI weight of 0.4 or more—in other words, only when MI is considered nearly one-half as important as death, which is highly unlikely given the case fatality rate of only 6%. Using a similar weighting scheme, we previously demonstrated that a statistically significant difference in favor of drug-eluting stents over bare-metal stents in the SIRIUS trial was observed only at a revascularization weight of 0.5 or more (33)—an arguably implausible conjecture.

It is apparent from these analyses that important information regarding the individual component end points may be obscured by combining them into a composite, which often biases the analysis in favor of the least important component. To avoid this, data on all individual components must be provided, preferably both including and excluding double counting to permit nonhierarchical and hierarchical composite end point analysis, respectively. Individual component analysis should be done to ensure that any 1 or more components meet statistical significance. Ideally, all components should demonstrate evidence of significant benefit (as in the HOPE trial). At the very least, the composite end point should be impacted favorably ($p < 0.05$), and evidence of harm in the most important component such as death or irreversible morbidity (nonfatal MI or stroke) should be excluded. One can put a limit of tolerable inferiority, for example, upper 95% CI not to exceed a risk ratio of 1.2 to 1.3, as in noninferiority trials.

JACC Vol. 55, No. 5, 2010
February 2, 2010:415–27

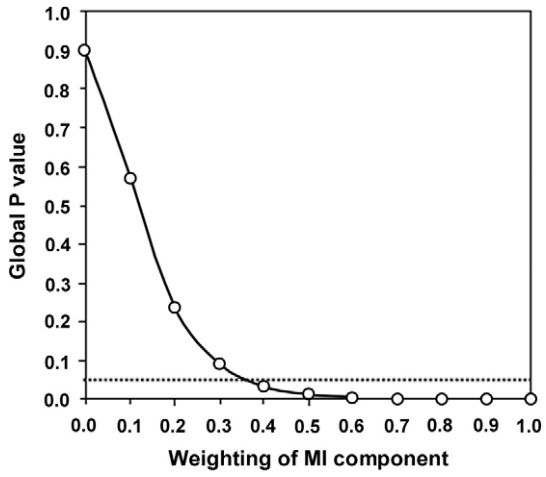

Figure 3 **Weighted Primary Efficacy Composite End Point Analysis for the TRITON–TIMI 38 Trial**

Data used are first events (to exclude double counting) (see Table 2). The p values are derived from a global z statistic under the assumption of no correlation among the components of the composite end point (32,33). A sensitivity analysis across myocardial infarction (MI) weighting is also shown (the weights of death and stroke were fixed at 1.0 and 0.5, respectively). Only MI weights ≥0.4 achieve statistical significance ($p \leq 0.05$, values below the **dotted line**).

However, there are sample size implications, and the strategy should be prospectively defined.

In summary, composite end points can be of value if both their requirements and their limitations are respected, and if the results are reported in a straightforward manner, providing all information necessary for proper interpretation of the trial. Suggested recommendations for their appropriate use in clinical trials are summarized in Table 3.

Perform Judicious Subgroup Analyses

Subgroup analysis means any evaluation of treatment effects (benefit or harm) for a specific end point in subdivisions of the study population defined by various nonrandom baseline characteristics. As the number of RCTs has dramatically increased over the last 3 decades, the exploration of treatment effects in patient subgroups has also simultaneously increased. While such analyses may provide useful information for the care of patients and for future research, they also introduce analytic and interpretive challenges that can lead to overstated results, misleading conclusions, and suboptimal care.

Several reviews have highlighted problems in the reporting of subgroup analyses (34–38). Assmann et al. (35) reported shortcomings of subgroup analyses in 50 trials published in 1997 in 4 leading medical journals. More recently, Parker et al. (36), who reviewed 67 cardiovascular trials published between 1980 and 1997, and Hernández et al. (37), who reviewed 63 cardiovascular trials published in 2002 and 2004, noted the same problems. Chief among them include a lack of pre-specification, and testing of a large number of subgroups without the use of statistically appropriate adjustment for interactions and multiple comparisons. Because a fairly large number of subgroup analyses are often undertaken, the potential for false positive errors is quite common. The collective probability of a false positive error (A) can be computed from the equation: $A = 1 - (1 - a)^X$, where X = number of independent subgroup analyses and a is the false positive error for each individual subgroup analysis (usually 0.05) (39). For example, if 20 subgroup analyses are conducted, the collective probability of at least 1 false positive error is 0.64. Conversely, because these analyses are often underpowered because of small sample size, false negative errors are also common. Finally, these analyses are usually nonrandomized, resulting in imbalances in prognostic factors in subgroups. For these reasons, subgroup analyses should be considered exploratory for informing future research and not conclusive to guide clinical practice.

The results of the CHARISMA (Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance) trial (40), summarized in Figure 4, highlight some of the major caveats that challenge the interpretation of subgroup analyses. In this trial, long-term dual antiplatelet therapy with clopidogrel was compared with placebo for patients without overt atherothrombotic disease but with multiple risk factors (asymptomatic) and for patients with clinically evident atherothrombotic disease (symptomatic) on the background of aspirin treatment. In the overall cohort, the rate of the primary end point was 6.8% with clopidogrel and 7.3% with placebo (relative risk: 0.93, 95% CI: 0.83 to 1.05, $p = 0.22$), the rate of TIMI major bleeding was higher with clopidogrel (2.1% vs. 1.3%, $p < 0.001$), and the rate of death from cardiovascular causes also was higher with clopidogrel (3.9% vs. 2.2%, $p = 0.01$). The rate of the primary end point among asymptomatic patients was 6.6% with clopidogrel and 5.5% with placebo, and it was 6.9% with clopidogrel and 7.9% with placebo among symptomatic patients. An interaction analysis revealed a statistically significant heterogeneity of treatment effects in symptomatic versus asymptomatic subgroup (interaction term 0.73, 95% CI: 0.54 to 1.00, $p = 0.045$). On the basis of these findings, the CHARISMA investigators

Table 3 **Suggested Recommendations for Use of Composite End Points**

Justify the validity of individual components.
Avoid clinically unimportant or uncertain outcomes.
Avoid components unlikely to be impacted by therapy.
Avoid combining efficacy outcomes with safety outcomes.
Report primary composite end point and individual components separately, preferably both hierarchical and nonhierarchical counts.
Examine treatment-by-end-point interaction by a formal assessment of heterogeneity.
Weigh components prospectively relative to their clinical importance.
Conduct and report sensitivity analyses relative to weight of the component driving the composite end point.

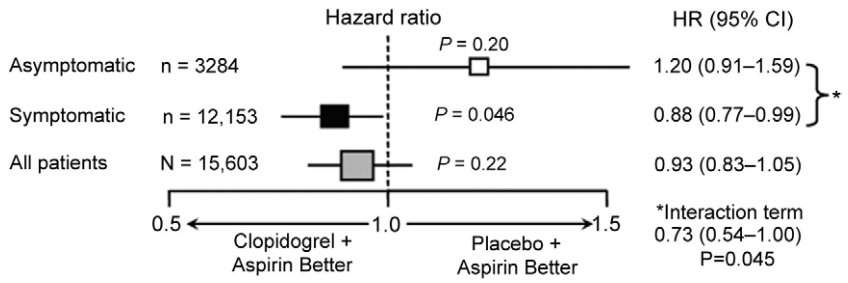

Figure 4 **Subgroup Analysis for the CHARISMA Trial**

The rate of the primary end point among asymptomatic patients (with multiple atherothrombotic risk factors) and symptomatic patients (with documented coronary, cerebrovascular, or peripheral arterial disease) is shown as hazard ratio (HR) and 95% confidence interval (CI). The result of subgroup analysis unadjusted for multiple comparisons is shown as an interaction term with 95% CI and p value. See text for further discussion.

concluded that "a subgroup analysis suggested that clopidogrel was beneficial with respect to the primary efficacy end point in patients who were classified as symptomatic . . ." (40). Are these conclusions justified?

Although the subgroup analysis in the CHARISMA trial was pre-specified, the subgroup classification was biologically plausible—high-risk symptomatic patients being more likely to benefit than low-risk symptomatic patients—and a proper interaction test was conducted, 3 limitations still merit consideration. First, the study failed to reach a statistical significance in the pre-specified overall analysis. In such cases, it is generally not a good idea to perform subgroup analysis because the risk of false positive results increases (you have, so to speak, already used up all of the pre-specified type I error to which you are entitled) (41). The probability of a false positive error is directly related to the number of subgroups tested (39). These caveats apply equally to a post-hoc analysis of the CHARISMA trial that concluded that dual antiplatelet therapy provided significant benefit in patients with documented prior MI, stroke, or symptomatic peripheral arterial disease (the so-called CAPRIE-like cohort) (42). Positive subgroups within negative trials such as the CHARISMA study are virtually always the result of confounding or bias, especially post-hoc defined subgroups (43).

Second, a key principle for interpretation of subgroup results is that quantitative interactions in which 1 treatment is always better than the other, but by various degrees (differences in degree but not direction), are much more credible than qualitative interactions (differences in direction) in which 1 treatment is better than the other for 1 subgroup of patients and worse for the other subgroup of patients (34)—the type of interaction observed in the CHARISMA study. Furthermore, quantitative interactions are likely to be truly present whether or not they are apparent, whereas apparent qualitative interactions should generally be disbelieved as they are seldom replicated consistently (34). Therefore, the overall trial result (a nonsignificant treatment effect in the CHARISMA study) is usually a better guide to the direction of effect in subgroups than the apparent effect observed within a subgroup (significant treatment benefit in symptomatic subgroup in the CHARISMA study).

Third, the subgroup analysis was not adjusted for multiple comparisons. Although a marginal treatment effect for secondary prevention in symptomatic patients was suggested by the subgroup analysis in the CHARISMA study ($p = 0.045$), when adjusted for multiple subgroup analyses, the corrected p value of 0.60 failed to reveal a differential treatment effect in symptomatic versus asymptomatic patients (39). Had the interaction tests been assessed with a criterion of 0.05 divided by 20 (0.0025) to account for the fact that 20 comparisons were conducted (the so-called Bonferroni correction), none would have come close to reaching statistical significance (39). The caveats enunciated in an accompanying editorial (44) and a statistical perspective (39) highlight the problems with uncorrected multiple subgroup comparisons, leading the latter writer to conclude that "overstating the results of subgroup analyses can misinform future research and lead to suboptimal clinical practice."

Consider another example—the TACTICS–TIMI 18 trial, in which the overall results with regard to the primary efficacy triple end point favored early invasive management over early conservative management in patients with non–ST-segment elevation ACS (7). A secondary objective of this study was to test prospectively the validity of the "troponin hypothesis," namely, to determine if benefits of invasive management were limited to troponin-positive patients (7,45). As shown in Figure 5, the treatment by troponin T subgroup interaction test was highly significant ($p = 0.003$), indicating a hazard ratio of 1.13 in the troponin-negative subgroup and 0.61 in the troponin-positive subgroup at 6-month follow-up. With such a highly significant interaction test, the investigators judged the early invasive strategy to be beneficial in the troponin-positive subgroup. Encouraged by these results and in light of similar findings observed in a post-hoc analysis in the

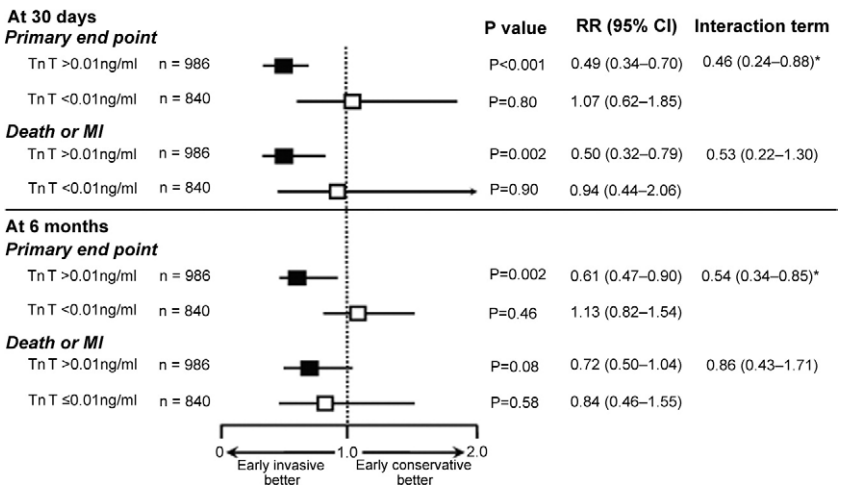

Figure 5 **Subgroup Analysis for the TACTICS–TIMI 18 Trial**

Outcomes at 30 days and at 6 months are stratified by treatment group and cardiac troponin T (TnT) level. The primary end point was death, myocardial infarction (MI), or urgent revascularization. *p = 0.01 at 30 days and p = 0.003 at 6 months for interaction between treatment group and cardiac TnT level with respect to the primary end point. CI = confidence interval; RR = risk ratio.

FRISC-II (Fast Revascularization During Instability in Coronary Artery Disease) trial, the investigators concluded that the "measurement of troponin at the time of presentation is useful in determining the optimal treatment strategy" (7,45). Accordingly, the treatment guidelines were upgraded to reflect these findings, and an early invasive strategy was endorsed as a class I-A recommendation for patients with suspected ACS and elevated troponin level with or without ST-segment depression or other high-risk features (11). An independent confirmation of these findings was conducted in the ICTUS (Invasive Versus Conservative Treatment in Unstable Coronary Syndromes) trial. Although the design of the ICTUS trial (46) was similar to that of the TACTICS–TIMI 18 study (except that enrollment was limited to patients with elevated troponins), the results proved to be different. The overall hazard ratio for the ICTUS trial was 1.07, showing no apparent treatment benefit for troponin-positive patients, thereby calling the strategy of early angiography for all patients with raised troponin into question. No differences in patient baseline characteristics, end point criterion, or concomitant therapy (except for higher statin and clopidogrel use in the ICTUS study) between the TACTICS–TIMI 18 study and the ICTUS trial could explain the discrepant results.

The results of these trials remind us of the need to confirm the findings from subgroup analyses, even if they are pre-specified and the results are as expected. This is particularly true for qualitative interactions—the type reported to be statistically significant in the TACTICS–TIMI 18 study for both troponin T and I (45). These types of interactions are seldom replicable and are likely to be spurious. It is important to note that none of the quantitative interactions reported for the death or nonfatal MI end point at both 30 days and 6 months in the TACTICS–TIMI 18 trial were statistically significant (7,45) (Fig. 5). These results are much more likely to be credible and suggest no treatment by troponin subgroup interaction. The lesson to be learned here is that a trial is typically designed to detect an effect in the whole population and that the most reliable estimate of a subgroup's results is still the overall results, not the estimate of a particular subgroup. The principal value of subgroup analysis is to assess the robustness of the primary conclusions by demonstrating consistency within the subgroups, not to demonstrate inconsistencies in one or another arbitrary subgroup. Subgroup findings should be regarded with suspicion unless they are independently confirmed. Failure to recognize the capriciousness of random variation often leads to premature acceptance of the results, risking the adoption of inferior or unnecessarily costly treatments (16).

The results of 2 recently published large RCTs (25,47) offer yet another example of how an improper interpretation of the results of subgroup analyses can potentially lead to misleading conclusions. The results from the TRITON–TIMI 38 study showed a treatment benefit with prasugrel over clopidogrel with regard to the efficacy end point. However, a bleeding hazard was also observed with prasugrel. A post-hoc exploratory subgroup analysis was conducted to identify patients at higher risk of bleeding (history of prior stroke/transient ischemic attack, body weight <60 kg or age >75 years). Based on the results, the investigators stated that "with exclusion of patients with prior stroke/transient ischemic attack and dose reduction in the elderly (age >75 years) and those with low body weight (<60 kg), the bleeding risk with prasugrel will be minimized" (25). Nevertheless, although the high-risk subgroup was numer-

ically at a higher risk for bleeding with prasugrel (42% vs. 24% relative increase in the non–high-risk subgroup), the between-group difference was not statistically significant (interaction p = 0.64). In contrast, the between-group difference in ischemic events (26% decrease vs. 2% increase in the high-risk subgroup, interaction p = 0.008) and net clinical benefit (20% improvement vs. 7% worsening in the high-risk subgroup; interaction p = 0.006) was significantly in favor of prasugrel in the non–high-risk subgroup. Similarly, the improvement in net clinical benefit in the subgroup without versus with a history of stroke or transient ischemic attack was driven by greater efficacy, not improved safety.

In conclusion, in the TRITON–TIMI 38 trial, exclusion of high-risk characteristics identified patients with improved efficacy, not reduced bleeding risk, associated with prasugrel, thereby questioning the claims of the trial investigators. Combining the efficacy with the safety end point into a "net clinical benefit" end point serves to obscure this subtle, but important, distinction. This has important implications for regulatory approval and clinical practice.

The results of the SYNTAX (Synergy Between Percutaneous Coronary Intervention With Taxus and Cardiac Surgery) trial showed that CABG, as compared with percutaneous coronary intervention (PCI) using the Taxus drug-eluting stent (Boston Scientific, Natick, Massachusetts), is associated with a lower rate of MACE or cerebrovascular events at 1 year among patients with 3-vessel or left main coronary artery (LMCA) disease, or both (47). A post-hoc subgroup analysis showed a trend toward lower events with PCI in cases with anatomically simple LMCA disease (LMCA only and LMCA plus single-vessel disease), compared with CABG-treated patients. These results have prompted many interventional cardiologists to choose PCI with stenting as a good treatment option for patients with LMCA disease (48,49). However, as shown in Figure 6, there was overlap between the treatment effect in the LMCA only cohort (interaction p = 0.51) or LMCA plus single-vessel disease cohort (interaction p = 0.10) and the effect in the overall cohort. Because the primary end point failed to reach a statistical verdict in favor of PCI, the estimate of treatment effect in the overall cohort is the most reliable estimate of treatment effect in the LMCA subgroup. On the basis of these results, it is generally not advisable to draw any inferences regarding PCI being the preferred treatment strategy for patients with LMCA disease (49).

Despite repeated discussion of the potential problems associated with subgroup analysis and published guidelines to improve the quality of subgroup analyses, a recent analysis of 97 trials published in the *New England Journal of Medicine* in 2005 and 2006 showed that problems and ambiguities persist (50). In approximately two-thirds of the published trials, it was unclear whether any of the reported subgroup analyses were pre-specified or post hoc. In more than one-half of the trials, it was unclear whether interaction tests were used, and in approximately one-third of the trials, within-level results were not presented in a consistent way. Recommendations on when and how subgroup analyses should be conducted and reported are shown in Table 4. The goal is to avoid unwarranted data dredging and increase the clarity and completeness of the information reported, thereby improving the interpretability of the findings.

Conclusions

The randomized controlled clinical trial has become the gold standard scientific method for the evaluation of diagnostic and treatment interventions. However, there are a number of limitations that challenge the interpretation of the results of these trials. Careful attention to these caveats

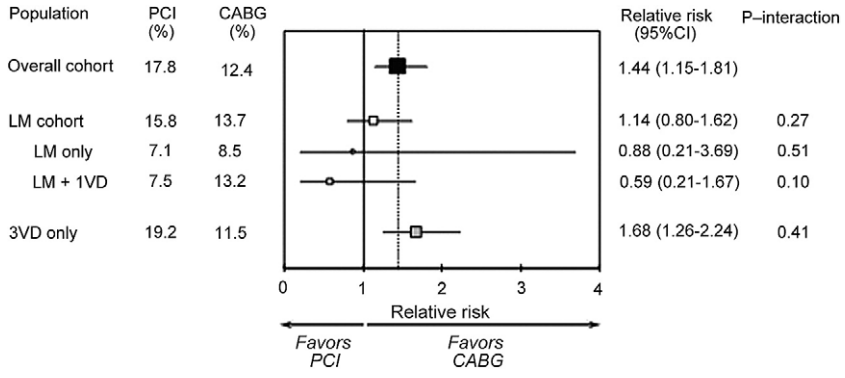

Figure 6 **Subgroup Analysis for the SYNTAX Trial**

The primary outcome (composite of death, nonfatal myocardial infarction [MI], nonfatal stroke, and repeat revascularization) at 12 months is stratified by mode of revascularization (percutaneous coronary intervention [PCI] vs. coronary artery bypass graft surgery [CABG]) and the anatomic subset of coronary artery disease: overall cohort, which includes left main (LM) plus 3-vessel disease (3VD), LM cohort subdivided into isolated LM and LM plus 1-vessel disease (1VD), and 3VD (48). Interaction p value unadjusted for multiple comparisons is shown; the **dotted line** represents the overlap in treatment effect. CI = confidence interval.

JACC Vol. 55, No. 5, 2010
February 2, 2010:415–27

Table 4 Suggested Guidelines for Subgroup Analysis

Prospectively define hypothesis.
Limit analyses to biologically plausible subgroups based on prior evidence.
Limit analyses to statistically significant treatment effects in pre-specified overall analysis.
Identify statistically significant interaction of treatment with the subgroup variable.
Perform adjustments for multiple comparisons.
Report results of subgroup analyses as exploratory and requiring independent confirmation.
Avoid overinterpretation of subgroup differences.

is not only key for critical evaluation of the literature, but it also has implications for the care and treatment of patients, and for the development and implementation of practice guidelines and reimbursement policy. Misinterpreting the results of trials can misinform future research and lead to suboptimal clinical practice.

Reprint requests and correspondence: Dr. Sanjay Kaul, Division of Cardiology, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, South Professional Tower, Room 5536, Los Angeles, California 90048-1804. E-mail: kaul@cshs.org.

REFERENCES

1. Diamond GA, Kaul S. Bayesian approaches to the analysis and interpretation of clinical megatrials. J Am Coll Cardiol 2004;43:1929–39.
2. Goodman SN. Toward evidence-based medical statistics. 1: the p value fallacy. Ann Intern Med 1999;130:995–1004.
3. Sterne JAC, Smith GD. Sifting the evidence—what's wrong with significance tests? BMJ 2001;322:226–31.
4. Panagiotakos DB. The value of p value in biomedical research. Open Cardiovasc Med J 2008;2:97–9.
5. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. J Epidemiol Comm Health 1998;52:318–23.
6. Sackett DL. Superiority, equivalence and noninferiority trials. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, editors. The Principles Behind the Tactics of Performing Therapeutic Trials. Clinical Epidemiology: How to Do Clinical Practice Research. 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins, 2006:193–6.
7. Cannon CP, Weintraub WS, Demopoulos LA, et al. Comparison of early invasive and conservative strategies in patients with unstable coronary syndromes treated with the glycoprotein IIb/IIIa inhibitor tirofiban. N Engl J Med 2001;344:1879–87.
8. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452–4.
9. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. JAMA 1995;273:871–5.
10. Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients (erratum in: BMJ 2002;324:141). BMJ 2002;324:71–86.
11. Braunwald E, Antman EM, Beasley JW, et al. ACC/AHA guideline update for the management of patients with unstable angina and non-ST-segment elevation myocardial infarction—summary article: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on the Management of Patients With Unstable Angina). J Am Coll Cardiol 2002;40:1366–74.
12. The Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial Investigators. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation (erratum in: N Engl J Med 2001;345:1506,1716). N Engl J Med 2001;345:494–502.
13. Boersma E, Harrington RA, Moliterno DJ, et al. Platelet glycoprotein IIb/IIIa inhibitors in acute coronary syndromes: a meta-analysis of all major randomised clinical trials. Lancet 2002;359:189–98.
14. Ferguson JJ, Califf RM, Antman EM, et al. Enoxaparin versus unfractionated heparin in high-risk patients with non-ST-segment elevation acute coronary syndromes managed with an intended early invasive strategy: primary results of the SYNERGY randomized trial. JAMA 2004;292:45–54.
15. Califf RM, DeMets DL. Principles from clinical trials relevant to clinical practice: part I. Circulation 2002;106:1015–21.
16. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials (part I). Circulation 2002;106:746–51.
17. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 2003;289:2554–9.
18. Montori V, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse J, Pachero-Huergo V, Bryant D. Validity of composite end points in clinical trials. BMJ 2005;330:594–6.
19. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007;334:786.
20. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, et al. Methodologic discussions for using and interpreting composite end points are limited, but still identify major concerns. J Clin Epidemiol 2007;60:651–7.
21. Lim E, Brown A, Helmy A, Mussa S, Altman D. Composite outcomes in cardiovascular research: a survey of randomized trials. Ann Intern Med 2008;149:612–7.
22. Kip K, Hollabaugh K, Marroquin O, Williams D. The problem with composite end points in cardiovascular studies. J Am Coll Cardiol 2008;51:701–7.
23. Mattos LA, Grines CL, Sousa JE, et al. One-year follow-up after primary coronary intervention for acute myocardial infarction in diabetic patients. A substudy of the STENT PAMI trial. Arq Bras Cardiol 2001;77:549–61.
24. The Heart Outcomes Prevention Evaluation Study Investigators. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. N Engl J Med 2000;342:145–53.
25. Wiviott SD, Braunwald E, McCabe CH, et al., for the TRITON-TIMI 38 Investigators. Prasugrel versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2007;357:2001–15.
26. Wiviott SD, Braunwald E, Murphy SA, Antman EM. A perspective on the efficacy and safety of intensive antiplatelet therapy in the Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition With Prasugrel–Thrombolysis in Myocardial Infarction 38. Am J Cardiol 2008;101:1367–70.
27. Kaul S, Diamond GA. Making sense of noninferiority: a clinical and statistical perspective on its application to cardiovascular clinical trials. Prog Cardiovasc Dis 2007;49:284–99.
28. The Platelet Receptor Inhibition for Ischemic Syndrome Management in Patients Limited by Unstable Signs and Symptoms (PRISM-PLUS) Trial Investigators. Inhibition of the platelet glycoprotein IIb/IIIa receptor with tirofiban in unstable angina and non-Q-wave myocardial infarction. N Engl J Med 1998;338:1488–97.
29. Dahlof B, Devereux RB, Kjeldsen SE, et al., for the LIFE Study Group. Cardiovascular morbidity and mortality in the Losartan Intervention for End Point Reduction in Hypertension Study (LIFE): a randomized trial against atenolol. Lancet 2002;359:995–1003.
30. The TIME Investigators. Trial of Invasive Versus Medical Therapy in Elderly Patients With Chronic Symptomatic Coronary-Artery Disease (TIME): a randomised trial. Lancet 2001;358:951–7.
31. Moses JW, Leon MB, Popma JJ, et al. Sirolimus-eluting stents versus standard stents in patients with stenosis in a native coronary artery. N Engl J Med 2003;349:1315–23.
32. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple end points in clinical trials. Biometrics 1987;43:487–98.
33. Tung R, Kaul S, Diamond GA, Shah PK. Narrative review: drug-eluting stents for the management of restenosis. A critical appraisal of the evidence. Ann Intern Med 2006;144:913–9.

34. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93–8.
35. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355: 1064–9.
36. Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000;139:952–61.
37. Hernández A, Boersma E, Murray G, Habbema J, Steyerberg E. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J 2006;151:257–64.
38. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86.
39. Lagakos SW. The challenge of subgroup analyses—reporting without distorting (erratum in N Engl J Med 2006;355:533). N Engl J Med 2006;354:1667–9.
40. Bhatt DL, Fox KAA, Hacke W, et al. Clopidogrel and aspirin versus aspirin alone for the prevention of atherothrombotic events. N Engl J Med 2006;354:1706–17.
41. Moyé LA. Statistical Reasoning in Medicine: The Intuitive P-Value Primer. New York, NY: Springer-Verlag, 2000.
42. Bhatt DL, Flather, MD, Hacke W, et al. Patients with prior myocardial infarction, stroke, or symptomatic peripheral arterial disease in the CHARISMA trial. J Am Coll Cardiol 2007;49:1982–8.
43. Gebel JM, Jr. The CAPRIE-like subgroups of CHARISMA: a CAPRIEciously biased analysis of an unCHARISMAtic truth. J Am Coll Cardiol 2007;50:1704.
44. Pfeffer MA, Jarcho JA. The charisma of subgroups and the subgroups of CHARISMA. N Engl J Med 2006;354:1744–6.
45. Morrow DA, Cannon CP, Rifai N, et al. Ability of minor elevations of troponin I and T to identify patients with unstable angina and non-ST elevation myocardial infarction who benefit from an early invasive strategy: results from a prospective, randomized trial. JAMA 2002;286:2405–12.
46. de Winter RJ, Windhausen F, Cornel JH, et al., for the Invasive Versus Conservative Treatment in Unstable Coronary Syndromes (ICTUS) Investigators. Early invasive versus selectively invasive management for acute coronary syndromes. N Engl J Med 2005;353:1095–104.
47. Serruys PW, Morice M-C, Kappetein AP, et al., for the SYNTAX Investigators. PCI vs CABG for severe coronary artery disease. N Engl J Med 2009;360:961–72.
48. Park S-J, Park D-W. Percutaneous coronary intervention with stent implantation versus coronary artery bypass surgery for treatment of left main coronary artery disease: is it time to change guidelines? Circ Cardiovasc Intervent 2009;2:59–68.
49. Teirstein PS. Percutaneous revascularization is the preferred strategy for patients with significant left main coronary stenosis. Circulation 2009;119;1021–33.
50. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357:2189–94.

Key Words: results ■ megatrials ■ interpret.