View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by LSHTM Research Online

Pocock, SJ; Stone, GW (2016) The Primary Outcome Fails - What Next? The New England journal of medicine, 375 (9). pp. 861-70. ISSN 0028-4793 DOI: https://doi.org/10.1056/NEJMra1510064

Downloaded from: http://researchonline.lshtm.ac.uk/2822454/

DOI: 10.1056/NEJMra1510064

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/

REVIEW ARTICLE

THE CHANGING FACE OF CLINICAL TRIALS

Jeffrey M. Drazen, M.D., David P. Harrington, Ph.D., John J.V. McMurray, M.D., James H. Ware, Ph.D., and Janet Woodcock, M.D., *Editors*

The Primary Outcome Fails — What Next?

Stuart J. Pocock, Ph.D., and Gregg W. Stone, M.D.

A well-designed trial derives its credibility from the inclusion of a prespecified, a priori hypothesis that helps its authors avoid making potentially false positive claims on the basis of an exploratory analysis of the data. Nevertheless, an unreasonable yet widespread practice is the labeling of all randomized trials as either positive or negative on the basis of whether the P value for the primary outcome is less than 0.05. This view is overly simplistic. P values should be interpreted as a continuum wherein the smaller the P value, the greater the strength of the evidence for a real treatment effect.[1-3] Confidence intervals are also useful in indicating the range of uncertainty around the estimated treatment effect. Moreover, the interpretation of any trial should depend on the totality of the evidence (i.e., the primary, secondary, and safety outcomes), not just a single end point.

Herein we outline the thought processes to follow when the primary outcome of a trial is first perceived as negative. In a subsequent issue of the *Journal*, we will review the questions to ask when the outcomes of a trial appear to be positive. We illustrate our points with examples from trials — primarily from cardiovascular studies, which constitute our sphere of expertise — but the underlying issues apply to the whole of medicine.

From the Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London (S.J.P.); and Columbia University Medical Center, New York Presbyterian Hospital, and the Cardiovascular Research Foundation — all in New York (G.W.S.). Address reprint requests to Dr. Pocock at the Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel St., London WC1E 7HT, United Kingdom; or at Stuart.Pocock@lshtm.ac.uk.

N Engl J Med 2016;375:861-70.
DOI: 10.1056/NEJMra1510064
Copyright © 2016 Massachusetts Medical Society.

KEY QUESTIONS WHEN THE PRIMARY OUTCOME FAILS

The failure to achieve the 5% level of significance is certainly not promising for the test treatment. Typical first reactions include the following: What went wrong? Is the treatment truly ineffective? Are there glimmers of hope? What next? Addressing the 12 questions below provides a path forward (Table 1).

IS THERE SOME INDICATION OF POTENTIAL BENEFIT?

Whether a signal of treatment benefit (a "trend") should be inferred from a P value greater than 0.05 requires thoughtful consideration. When the primary findings of a trial are completely neutral, interpretation is straightforward. For instance, when the PERFORM trial of terutroban versus aspirin in patients with ischemic stroke (see box for a list of the complete names of all trials mentioned in this article)[4] showed no significant between-group difference with respect to the composite primary outcome — ischemic stroke, myocardial infarction, or other vascular cause of death (hazard ratio, 1.02; 95% confidence interval [CI], 0.94 to 1.12) — the trial was stopped early because of futility, and no safety advantages were detected for terutroban. These findings support the interpretation of a "negative trial."

In contrast, in the TORCH trial,[5] in which the effects of salmeterol plus flutica-

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

Table 1. Questions to Ask When the Primary Outcome Fails.

Is there some indication of potential benefit?
Was the trial underpowered?
Was the primary outcome appropriate (or accurately defined)?
Was the population appropriate?
Was the treatment regimen appropriate?
Were there deficiencies in trial conduct?
Is a claim of noninferiority of value?
Do subgroup findings elicit positive signals?
Do secondary outcomes reveal positive findings?
Can alternative analyses help?
Does more positive external evidence exist?
Is there a strong biologic rationale that favors the treatment?

Trial Names.

ASCOT: Anglo-Scandinavian Cardiac Outcomes Trial
ASPEN: Atorvastatin Study for Prevention of Coronary Heart Disease Endpoints in Non–Insulin-Dependent Diabetes Mellitus
BARI 2D: Bypass Angioplasty Revascularization Investigation 2 Diabetes
BEAUTIFUL: Ivabradine for Patients with Stable Coronary Artery Disease and Left-Ventricular Systolic Dysfunction
CAPRICORN: Carvedilol Postinfarction Survival Control in Left Ventricular Dysfunction
CARDS: Collaborative Atorvastatin Diabetes Study
CHAMPION trials: Cangrelor versus Standard Therapy to Achieve Optimal Management of Platelet Inhibition
CHARM-Preserved: Candesartan in Heart Failure Assessment of Reduction in Mortality and Morbidity
CIBIS II: Cardiac Insufficiency Bisoprolol Study II
EXCEL: Evaluation of XIENCE PRIME Everolimus Eluting Stent System [EECSS] or XIENCE V EECSS versus Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization
MATRIX: Minimizing Adverse Hemorrhagic Events by Transradial Access Site and Systemic Implementation of Angiox
MOXCON: Sustained Release Moxonidine for Congestive Heart Failure
PEGASUS-TIMI 54: Prevention of Cardiovascular Events in Patients with Prior Heart Attack Using Ticagrelor Compared to Placebo on a Background of Aspirin–Thrombolysis in Myocardial Infarction 54
PERFORM: Terutroban versus Aspirin in Patients with Cerebral Ischemic Events
PROactive: Prospective Pioglitazone Clinical Trial in Macrovascular Events
SHIFT: Systolic Heart Failure Treatment with the I_f Inhibitor Ivabradine Trial
SIGNIFY: Study Assessing the Morbidity–Mortality Benefits of the I_f Inhibitor Ivabradine in Patients with Coronary Artery Disease
SPARCL: Stroke Prevention by Aggressive Reduction in Cholesterol Levels
STICH: Comparison of Surgical and Medical Treatment for Congestive Heart Failure and Coronary Artery Disease
SYMPLICITY HTN-3: Renal Denervation in Patients with Uncontrolled Hypertension
SYNTAX: Synergy between PCI with Taxus and Cardiac Surgery
TARGET: Do Tirofiban and ReoPro Give Similar Efficacy Trial
TOPCAT: Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist
TORCH: Towards a Revolution in COPD Health
VALIANT: Valsartan in Acute Myocardial Infarction

sone propionate versus placebo were assessed in patients with chronic obstructive pulmonary disease (COPD), the P value for the primary outcome of death from any cause was 0.052, and "significant benefits in all other outcomes" (e.g., COPD exacerbations and health status) were found. Consequently, the findings of this trial merited a more constructive interpretation than that of "negative trial."

WAS THE TRIAL UNDERPOWERED?

The inclusion of too few patients in a study increases the risk that a significant treatment benefit will not be shown, even if such an effect exists (a type 2 error). For instance, in a trial of bisoprolol versus placebo in patients with systolic heart failure,[6] the hazard ratio for the primary outcome, death from any cause, was 0.80 (95% CI, 0.56 to 1.15; P=0.22). However, with only 621 patients, the trial was underpowered. Fortunately, the sponsors persisted, and the subsequent CIBIS II trial,[7] which included 2647 patients, showed that mortality was lower among those who received bisoprolol than among those who received placebo (hazard ratio, 0.66; 95% CI, 0.54 to 0.81; P<0.0001). Note that the estimated 34% lower mortality with bisoprolol in this larger trial was within the 95% confidence interval reported in the first study.

In general, when a trial is too small to detect modest treatment effects, it is appropriate to describe the findings as inconclusive rather than negative. An adequately powered study requires the accrual of a sufficient number of primary-outcome events, which can be achieved by recruiting more patients, enrolling patients at higher risk, prolonging follow-up, specifying an outcome that occurs more frequently (including the use of composite outcomes), or a combination thereof.

WAS THE PRIMARY OUTCOME APPROPRIATE (OR ACCURATELY DEFINED)?

The use of a composite outcome increases the number of primary events but does not necessarily increase statistical power. For instance, in the PROactive trial,[8] in which pioglitazone was compared with placebo in patients with type 2 diabetes, the composite primary outcome was death, myocardial infarction, stroke, acute coronary syndrome, endovascular surgery, or leg amputation. With 514 primary events in the pioglitazone

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

group versus 572 primary events in the placebo group, the P value was 0.08. For the more conventional composite outcome of death, myocardial infarction, or stroke, there were 301 events in the pioglitazone group versus 358 events in the placebo group (P=0.03). Thus, the addition of the extra components merely contributed random noise, thereby diluting a potentially real effect into nonsignificance.

Trial success may hinge on definitions of the outcomes and on the methods used for their adjudication. For example, the CHAMPION PLATFORM trial of cangrelor versus clopidogrel in patients undergoing percutaneous coronary intervention (PCI)[9] was stopped early for futility, since cangrelor was not shown to be beneficial with respect to the primary outcome (death, myocardial infarction, or ischemia-driven revascularization within 48 hours). However, the definition of periprocedural myocardial infarction did not effectively identify infarctions that occurred soon after PCI among patients with biomarker-positive acute coronary syndrome; a more precise definition of myocardial infarction might have contributed to a positive result.[10] Thus, in a subsequent trial, CHAMPION PHOENIX,[11] the rise and fall of biomarkers and clinical events were more carefully adjudicated to better discriminate periprocedural myocardial infarctions. A 22% lower rate of the 48-hour primary outcome (death, myocardial infarction, stent thrombosis, or ischemia-driven revascularization) with cangrelor than with clopidogrel was found (P=0.005) and resulted in U.S. and European regulatory approval.

WAS THE POPULATION APPROPRIATE?

An apt question to ask when a new treatment fails is whether the wrong patient population was studied. For instance, two large trials of ivabradine involving patients with stable coronary disease — BEAUTIFUL[12] and SIGNIFY[13] — failed to show any treatment benefit. However, in the SHIFT trial, which involved patients with chronic heart failure,[14] the incidence of the primary outcome, cardiovascular death or hospitalization for heart failure, was 26% lower with ivabradine than with placebo (P<0.0001). Selection of the appropriate population on the basis of mechanistic effects and preliminary studies is essential for pivotal trial success.

WAS THE TREATMENT REGIMEN APPROPRIATE?

Determination of the dosage regimen for a new drug in a pivotal trial can be challenging. In hindsight, the failures of tirofiban for the prevention of ischemic events associated with percutaneous coronary revascularization in the TARGET trial[15] and of moxonidine in the MOXCON trial[16] may well have been the result of dose selection — too low in the case of tirofiban and too high in the case of moxonidine. However, such musings, even if they are based on reexamination of the in vitro or phase 2 dose-ranging data, rarely lead to a subsequent trial in which the potentially more appropriate dose is tested. Some pivotal trials minimize this risk by having a three-group design that includes two dosage regimens for the new drug; one example is PEGASUS-TIMI 54,[17] a study in which the lower, 60-mg dose of ticagrelor bested both a 90-mg dose and placebo for long-term use beyond 1 year after a myocardial infarction.

WERE THERE DEFICIENCIES IN TRIAL CONDUCT?

A true treatment effect may be diluted, or disappear entirely, if there is poor adherence to the study protocol. For instance, in the TOPCAT trial,[18] a six-country study of spironolactone versus placebo that involved patients who had heart failure with preserved left ventricular ejection fraction, the composite outcome (cardiovascular death, cardiac arrest, or hospitalization for heart failure) showed only a nonsignificant trend in favor of spironolactone (hazard ratio 0.89; 95% CI, 0.77 to 1.04; P=0.14). But patients in Russia and Georgia had very few primary outcome events,[19] which suggests that there was some failure in study conduct in those countries or the enrollment of atypical patients. Confining analysis to only the other four countries yielded a significant treatment benefit (hazard ratio, 0.82; 95% CI, 0.69 to 0.98; P=0.026). There has been debate as to whether this post hoc evidence is convincing enough to recommend spironolactone for patients who have heart failure with preserved left ventricular ejection fraction.

IS A CLAIM OF NONINFERIORITY OF VALUE?

When a new treatment fails to show superiority to an active control, can noninferiority be claimed? Such a claim can be desirable if the new treatment has other advantages (e.g., it is less invasive

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

or has fewer side effects), but in most cases it is appropriate to make that claim only if the noninferiority hypothesis was prespecified. For instance, in the VALIANT trial,[20] in which patients with complicated myocardial infarction received valsartan, captopril, or both, no benefit was shown for valsartan with regard to the primary outcome — death from any cause (hazard ratio, 1.00; 97.5% CI, 0.90 to 1.11; P=0.98). However, this confidence interval excluded the prespecified noninferiority margin of 1.13, which allowed investigators to conclude that valsartan was noninferior to captopril. Valsartan is thus an acceptable alternative for patients who cannot take captopril because of unacceptable side effects (e.g., cough, a taste disturbance, or rash).

DO SUBGROUP FINDINGS ELICIT POSITIVE SIGNALS?

Although it is appropriate to consider subgroup findings in any major trial, for a trial in which the overall result for the primary outcome is neutral or negative, such considerations are often misleading, since the potential for harm is often implied for the partner subgroups. Such qualitative interactions are rarely plausible (unless a strong mechanistic underpinning is present), and the analyses are typically not adjusted for multiple comparisons; even if the findings from statistical tests of interaction are significant, such findings should usually be perceived as useful for generating hypotheses at best.[21,22] Indeed, we find it hard to think of an example in which an apparent benefit in a subgroup in a trial with a negative outcome has led to a confirmation in a subsequent trial."

Nevertheless, such a scenario has motivated a large-scale, international trial of coronary revascularization strategies. The SYNTAX trial[23] of PCI versus coronary-artery bypass grafting (CABG) in patients with three-vessel or left main coronary artery disease yielded overall superior results with CABG. But for the subgroup with left main coronary artery disease (further excluding patients with high anatomical complexity), PCI appeared to be an acceptable (possibly superior) alternative to CABG. This post hoc subgroup analysis served as the motivation for the ongoing EXCEL trial[24] of PCI versus CABG in patients with left main coronary artery disease and low-to-moderate anatomical complexity, the results of which are expected in the fall of 2016.

DO SECONDARY OUTCOMES REVEAL POSITIVE FINDINGS?

If the primary outcome is negative, positive findings for secondary outcomes are usually considered to be hypothesis-generating. Certainly, regulatory approval of a new drug is unlikely to follow. However, in some instances, secondary findings are compelling enough to affect guidelines and practice. For instance, in the ASCOT trial of amlodipine versus atenolol for hypertension,[25] the hazard ratio for the composite primary outcome of nonfatal myocardial infarction or fatal coronary heart disease was 0.90 (95% CI, 0.79 to 1.02; P=0.11). However, the data supporting evidence of the superiority of amlodipine with respect to stroke, total cardiovascular events, death from any cause, and new-onset diabetes were overwhelming (P<0.001, P<0.0001, P=0.02, and P<0.0001, respectively) (Fig. 1). In hindsight, the primary outcome was an odd choice: the decision not to include stroke in a hypertension trial is unconventional. These results support recommendations against the use of atenolol as a first-line or second-line antihypertensive agent.

Few studies are appropriately powered to assess effects on mortality. Proper interpretation can thus be challenging when a large trial shows a reduction in all-cause mortality, which is plausible but not prespecified — especially if the primary outcome was negative. For example, in the MATRIX trial,[26] patients with an acute coronary syndrome who underwent PCI were randomly assigned to receive procedural anticoagulation with bivalirudin or unfractionated heparin. There was no significant difference in the 30-day composite primary outcome of death, myocardial infarction, or stroke (relative risk, 0.94; 95% CI, 0.81 to 1.09; P=0.44). However, bivalirudin was associated with a markedly lower incidence of major bleeding as well as lower all-cause mortality (relative risk 0.71; 95% CI, 0.51 to 0.99; P=0.04), a result also observed in some previous studies.[27] This finding of reduced mortality with bivalirudin, although mechanistically plausible, ideally requires an additional adequately powered trial for resolution.

CAN ALTERNATIVE ANALYSES HELP?

Covariate Adjustment

Covariate-adjusted analysis that includes baseline variables strongly related to the primary outcome will result in slightly greater statistical power

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

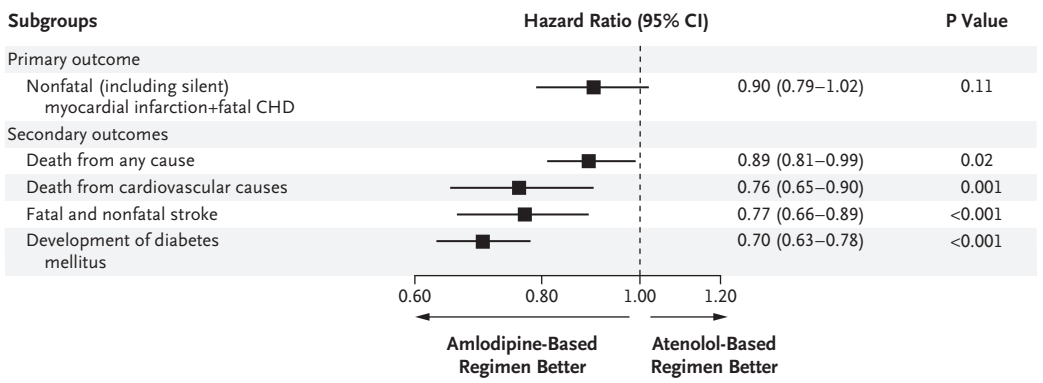

Figure 1. Major Results from ASCOT.

Although the primary outcome did not reach statistical significance, numerous secondary outcomes were positive. Given the biologic plausibility of these findings and their consistency with the results from previous trials, the results of this study provide meaningful data that can be used to inform decisions regarding treatments for hypertension. Adapted from Dahlöf et al.[25] CHD denotes coronary heart disease.

than a crude unadjusted analysis.[22] However, if the covariates were not precisely prespecified or the adjusted analysis was not predeclared as primary, the finding will be perceived as interesting and exploratory rather than one that affects the main conclusions of the trial.

For example, in the SPARCL trial of atorvastatin versus placebo after stroke or transient ischemic attack,[28] an unadjusted analysis yielded a borderline result in favor of atorvastatin for the primary outcome of recurrent stroke (P=0.05). A prespecified, covariate-adjusted analysis that accounted for geographic region, entry event and duration, age, and sex yielded a hazard ratio of 0.84 (95% CI, 0.71 to 0.99; P=0.03). It is not clear which was the prespecified primary analysis. Under the dubious premise that a significance level of 5% should be of paramount importance, one might debate whether the trial is "positive." A more reasonable verdict is that overall there is modest evidence of a treatment benefit.

As-Treated or Per-Protocol Analyses

Analysis conducted according to the intention-to-treat principle[29] is the main method used to make a valid comparison between two treatment strategies according to the treatments that were actually delivered to all patients who underwent randomization. When an intention-to-treat analysis fails to reach statistical significance, arguments are advanced that nonadherence and treatment crossovers may have masked real treatment effects and that as-treated or per-protocol analyses may get closer to the truth. Unfortunately, the use of as-treated or per-protocol populations introduces selection bias, because patients who do not adhere to the treatment regimen and those who cross over to the other treatment strategy may have a different prognosis that is unrelated to actual treatment. Hence, such analyses rarely influence conclusions regarding treatment efficacy that are based on the intention-to-treat principle. However, on-treatment analyses may be considered appropriate when safety issues are examined.

In the STICH trial[30] of CABG versus medical therapy in patients with left ventricular dysfunction (Fig. 2), in the intention-to-treat analysis, the hazard ratio for the primary outcome of death from any cause at a median follow-up of 4 years was 0.86 (95% CI, 0.72 to 1.04; P=0.12). Both an as-treated analysis (in which all patients who received CABG in the first year, including patients who crossed over to CABG, were compared with those who received medical therapy alone) and a per-protocol analysis (in which data from any patients who crossed over within the first year were excluded) revealed lower mortality with CABG (P<0.001 and P=0.005, respectively). Nonetheless, the principal conclusion remained "no significant difference between medical therapy and CABG with respect to the primary outcome." In the intention-to-treat population, other

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

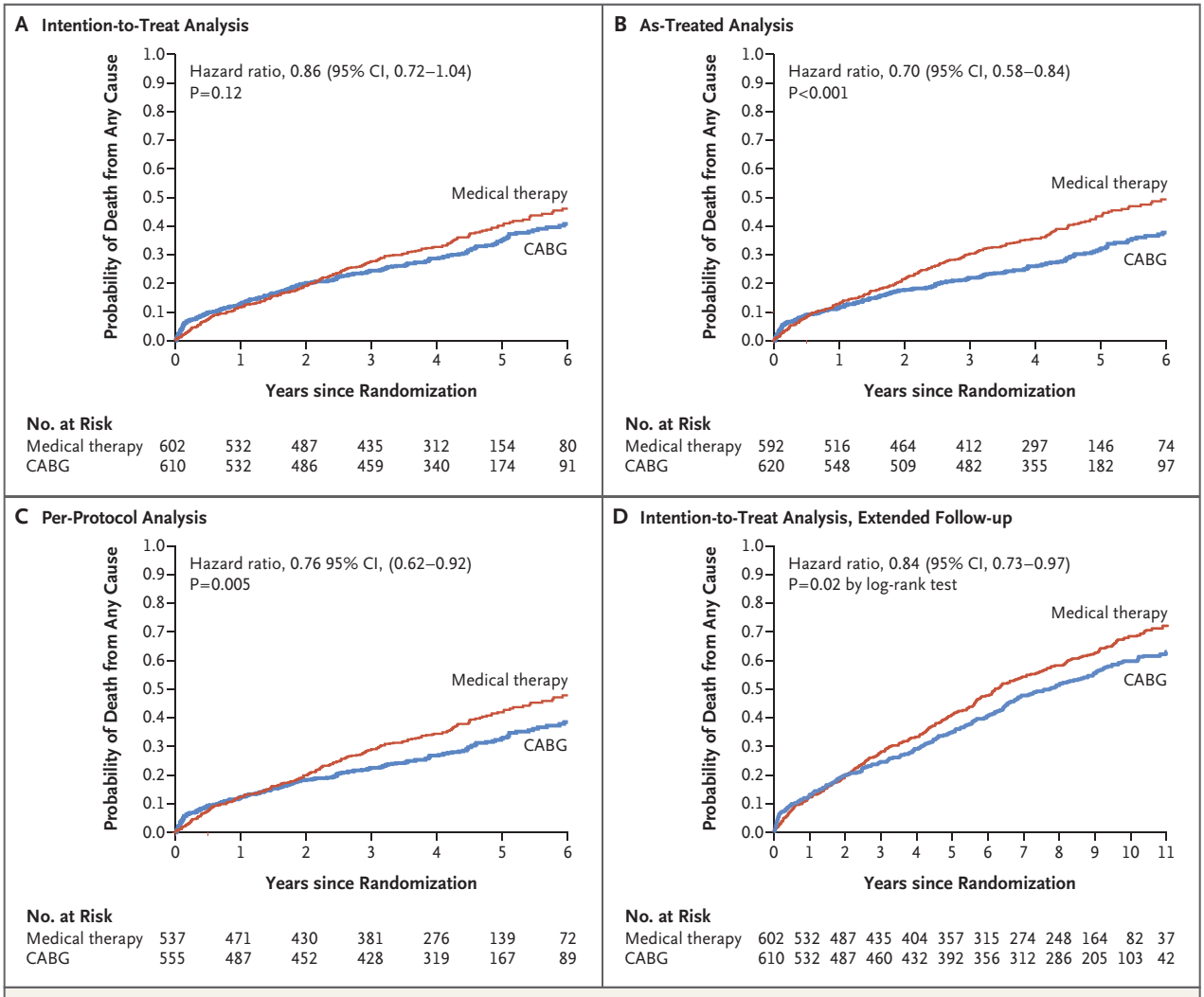

Figure 2. Results from the Primary Intention-to-Treat Population and the Alternative As-Treated and Per-Protocol Populations in the STICH Trial and the Potential Effect of Extended Follow-up.

Among 1212 patients with coronary artery disease and left ventricular ejection fraction of less than 35%, a total of 610 patients were randomly assigned to coronary-artery bypass grafting (CABG) and 602 patients to medical therapy. There was a nonsignificant difference in all-cause mortality in the prespecified, intention-to-treat analysis (Panel A). However, CABG was superior to medical therapy in an as-treated analysis in which the 592 patients who were treated with medical therapy throughout the first year after randomization were compared with the 620 patients who underwent CABG during initial assignment or as a result of crossover (Panel B). Similarly, CABG was superior to medical therapy in a per-protocol analysis in which the 537 patients who were randomly assigned to medical therapy who did not cross over to CABG during the first year of follow-up were compared with the 555 patients who were randomly assigned to and underwent CABG (Panel C). During an extended 10-year follow-up period, a significant mortality benefit with CABG emerged from an intention-to-treat analysis (Panel D). Adapted from Velazquez et al.[30,31]

benefits related to the outcomes of death from cardiovascular causes and the composite of death or hospitalization for cardiovascular disease were noted. Moreover, 10-year follow-up data in the STICH study has shown lower mortality with CABG than with medical therapy alone in the intention-to-treat population (hazard ratio, 0.84; 95% CI, 0.73 to 0.97; P=0.02).[31] Thus, the totality of the evidence supports an important role for CABG in patients with left ventricular dysfunction.

A related issue is the question of how to interpret trials that have high rates of crossover. For example, in the BARI 2D trial[32] of prompt coronary

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

revascularization versus intensive medical therapy among patients with type 2 diabetes, there were no significant differences in the 5-year coprimary outcomes of death and major cardiovascular events (P=0.97 and P=0.70, respectively). However, 42% of patients in the medical-therapy group had undergone clinically indicated revascularization, which raises questions about the value of medical therapy alone. Although such crossovers are an integral component of the initial conservative approach to treatment (and allowed revascularization to be avoided in the majority of patients), when crossovers occur frequently, it is fair to ask whether an adequate distinction may be drawn between the alternative strategies.

Analyses of Repeat Events

In studies of chronic diseases such as heart failure, conventional composite outcome analyses concentrate on the time to the first event and ignore any repeat events that occur subsequently. This approach can lead to serious loss of statistical power and underestimation of a treatment effect.

For instance, in the CHARM-Preserved trial[33] of candesartan versus placebo in patients with heart failure and preserved left ventricular ejection fraction, the hazard ratio for the composite primary end point of time to first unplanned admission to the hospital for the management of worsening heart failure or cardiovascular death was 0.89 (95% CI, 0.77 to 1.03; P=0.12). A subsequent analysis of all heart-failure–associated hospitalizations, including repeat hospitalizations, showed a rate ratio of 0.75 (95% CI, 0.62 to 0.91; P=0.003) (Fig. 3). The authors concluded that "recurrent events should be routinely incorporated into the analysis of future clinical trials in heart failure."

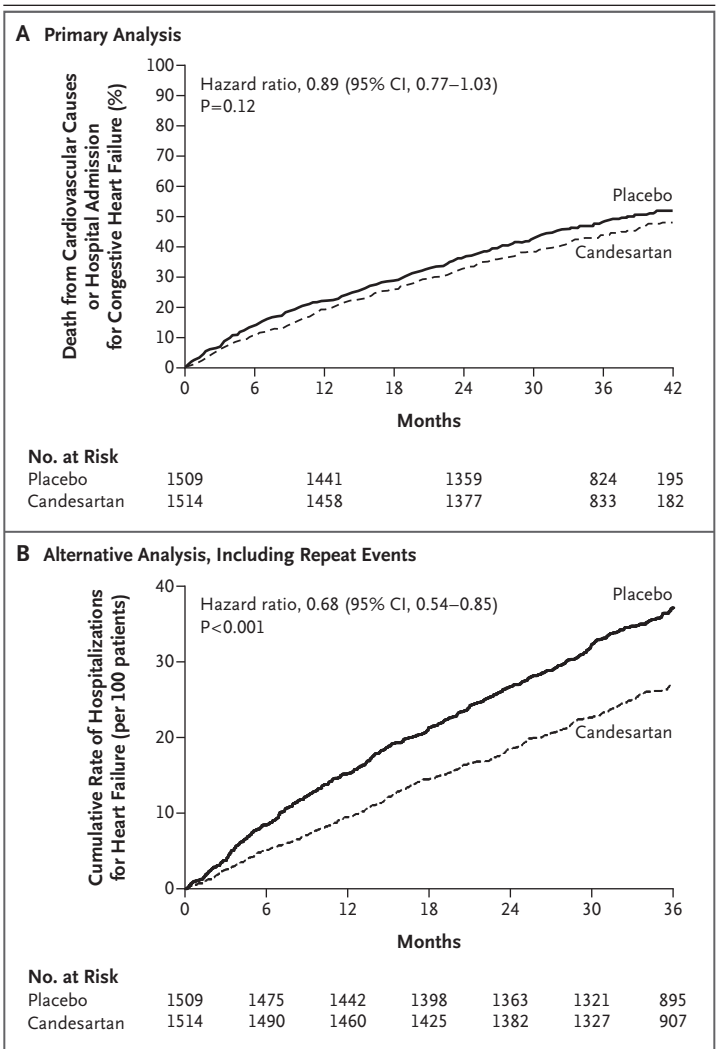

Figure 3. Positive Findings Achieved by Means of an Alternative Analysis in the CHARM-Preserved Trial.

Among 3023 patients with heart failure and preserved left ventricular ejection fraction who underwent randomization, no significant difference was observed between the candesartan group and the placebo group with respect to the primary outcome of cardiovascular death or hospitalization for heart failure in the prespecified time-to-first-event analysis (Panel A). However, when all events were accounted for in an alternative analysis, the data provided strong evidence for the effectiveness of candesartan (Panel B). Adapted from Rogers et al.[33]

DOES MORE POSITIVE EXTERNAL EVIDENCE EXIST?

When a negative primary outcome in an adequately powered trial seems surprising given preexisting evidence, the strength and quality of prior studies must be scrutinized. First, nonrandomized comparisons and surrogate end points from prior trials are not strong evidence. Evidence from analogous trials or meta-analyses involving similar types of patients, treatments, and outcomes are more valuable.

For instance, in the ASPEN trial,[34] in which the use of atorvastatin versus placebo was assessed in patients with type 2 diabetes, the hazard ratio for the composite primary outcome, a mix of cardiovascular events, was 0.90 (95% CI, 0.73 to 1.12; P=0.34). Given the positive outcomes associated with statins in other trials conducted in various patient populations, the results of the ASPEN trial were disappointing. In the larger

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

Table 2. Examples of Trials with Positive Claims Despite the Failure of the Primary Outcome.

ASCOT[25] and CAPRICORN[37]: Data from secondary outcomes provided strong evidence of superiority
TOPCAT[18,19]: Findings were positive after the exclusion of outlier countries
SYNTAX[23]: Data from a study subgroup provided justification for another trial
STICH[30,31]: Data from the as-treated and per-protocol analyses and from extended follow-up provided support for the primary outcome
CHARM-Preserved[33]: Data supporting the study drug were strong after recurrent events were incorporated into the analysis

CARDS trial,[35] which also involved the comparison of atorvastatin and placebo in patients with type 2 diabetes, the hazard ratio for the composite primary outcome (which was similar to that in the ASPEN trial) was 0.63 (95% CI, 0.48 to 0.83; P=0.001), and a meta-analysis of the two trials also produced a positive conclusion. The apparent inconsistency is not great (note the overlapping confidence intervals), so perhaps ASPEN was just the "unlucky" statin trial in which there was random variation away from a true treatment effect.

Nonetheless, favorable findings from meta-analyses should be interpreted cautiously, given the variations across trials in patient selection, the actual treatments studied, and definitions of outcomes and other differences in trial design and conduct. In general, evidence from one large, adequately powered randomized trial is preferred to that from a meta-analysis of smaller studies. Discrepancies between a large trial and a prior meta-analysis warrant further studies to resolve these inconsistencies.

IS THERE A STRONG BIOLOGIC RATIONALE THAT FAVORS THE TREATMENT?

One needs to be wary of arguments regarding biologic rationale. Almost any new treatment in a phase 3 trial has a plethora of supportive scientific data from animal studies and early-phase trials. Nonetheless, history is filled with records of many large pivotal trials that failed to exhibit any signs of efficacy (or that revealed heretofore unanticipated safety issues). For instance, the hypothesis that raising high-density lipoprotein cholesterol levels could be a new means of reducing cardiovascular events looked promising, but no trial of inhibitors of cholesteryl ester transfer protein has fulfilled that promise.[36] Nature often overcomes our best efforts to interrupt the order of things. Thus, if methodologic flaws in a trial are not the cause of treatment failure, it is usually time to "move on," while trying to understand the biologic reasons for the failure.

DISCUSSION

The 12 points explained above can be used to provide assistance in deciding what to do next when a trial fails to produce a positive finding for its primary outcome. Certainly one needs to be circumspect. Researchers may opt to move in one of three directions.

DECLARE THAT THE TRIAL IS POSITIVE

Remarkable circumstances are usually required to report that a trial is positive even though the results of the primary outcome were not statistically significant at the prespecified level. The descriptions of the findings for the five trials listed in Table 2 provide a framework for debate as to whether each contained positive findings of clinical importance despite the negative primary outcome. However, although such considerations may inform guidelines committees, regulators are rarely swayed by such secondary analyses.

One notable exception was the CAPRICORN trial, which assessed the effects of carvedilol versus placebo after myocardial infarction in patients with left ventricular dysfunction.[37] The composite primary outcome — death or hospitalization from any cause — failed to reach significance (hazard ratio, 0.92; 95% CI, 0.80 to 1.07; P=0.30). But all-cause mortality alone did provide some evidence of benefit (hazard ratio, 0.77; 95% CI, 0.60 to 0.98; P=0.03) and, after much debate, led to FDA approval, perhaps because all-cause mortality had been the original primary outcome (an unfortunate switch was made by the investigators midtrial), and external evidence existed as to the effectiveness of beta-blockers in this population of patients.

IMPROVE THE DESIGN OF FUTURE TRIALS

Trialists and sponsors usually have strong mechanistic support and background evidence that justifies the conduct of a major randomized trial. Hence, after a disappointing result, explanations are sought to guide the effort of designing a potential new trial. Aspects to consider include adjusting the treatment regimen, altering the study population, modifying the primary out-

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

come, increasing the sample size, and improving other aspects of the trial that affected its quality. Such difficult and costly decisions should be based on realistic expectations rather than naive optimism.

For example, after numerous open-label studies were conducted with highly positive results, renal denervation failed to substantially reduce blood pressure in patients with refractory hypertension in the sham-controlled SYMPLICITY HTN-3 trial.[38] Proposed explanations for this finding (which few expected) include an unfavorable mix of patients (some of whom had hypertension with an underlying cause that made a response to renal denervation unlikely), inadequate delivery of radiofrequency energy, changes in the drug treatment, and the failure to control for regression to the mean. Blinded mechanistic trials are ongoing in patients with hypertension who are not taking any antihypertensive medication to determine whether renal denervation does indeed "work," before additional, large-scale trials are conducted.

ABANDON THE TREATMENT AS INEFFECTIVE

The purpose of randomized trials is to distinguish between treatments that are effective and those that are not. Unfortunately, many innovations land in the second category. Hence, if the overall results of a trial show little or no evidence of treatment efficacy, and especially if there are safety issues, it may be wise to desist from further investigation. Such a conclusion may finally have been reached for thrombus aspiration in patients with acute myocardial infarction. After many years of mixed results from smaller trials, two large randomized trials[39,40] have now convincingly shown that routine thrombus aspiration has no benefit.

CONCLUSIONS

When the primary outcome of a trial fails to achieve statistical significance, we propose that researchers ask a series of searching questions that will help them clarify whether the new treatment may still have value. The options are to claim "success" anyway on the basis of the total evidence (an option that is rarely used), to plan a future trial with design improvements (a costly option), or to accept that the new treatment is likely to be ineffective (a frustrating option). However, the best option is to avoid this scenario altogether through rigorous upfront planning. By making sure that there is evidence of strong pathophysiological and mechanistic underpinnings that are common to both the new therapy and the disease, by selecting appropriate patients and end points, by calculating an adequate sample size, by paying meticulous attention to dosing, definitions of disease and outcomes, and all procedural processes, and by anticipating the ways in which the trial might fail and give rise to criticisms, one can enhance the likelihood of reaching a decisive conclusion.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Tim Collier for his help in producing earlier versions of the figures.

REFERENCES

1. Sterne JA, Davey Smith G. Sifting the evidence — what's wrong with significance tests? BMJ 2001;322:226-31.
2. Stone GW, Pocock SJ. Randomized trials, statistics, and clinical inference. J Am Coll Cardiol 2010;55:428-31.
3. Pocock SJ, McMurray JJV, Collier TJ. Making sense of statistics in clinical trial reports: part 1 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol 2015;66:2536-49.
4. Bousser MG, Amarenco P, Chamorro A, et al. Terutroban versus aspirin in patients with cerebral ischaemic events (PERFORM): a randomised, double-blind, parallel-group trial. Lancet 2011;377:2013-22.
5. Calverley PMA, Anderson JA, Celli B, et al. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. N Engl J Med 2007;356:775-89.
6. CIBIS Investigators and Committees. A randomized trial of β-blockade in heart failure: the Cardiac Insufficiency Bisoprolol Study (CIBIS). Circulation 1994;90:1765-73.
7. CIBIS-II Investigators and Committees. The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. Lancet 1999;353:9-13.
8. Dormandy JA, Charbonnel B, Eckland DJA, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. Lancet 2005;366:1279-89.
9. Bhatt DL, Lincoff AM, Gibson CM, et al. Intravenous platelet blockade with cangrelor during PCI. N Engl J Med 2009;361:2330-41.
10. Leonardi S, Truffa AA, Neely ML, et al. A novel approach to systematically implement the universal definition of myocardial infarction: insights from the CHAMPION PLATFORM trial. Heart 2013;99:1282-7.
11. Bhatt DL, Stone GW, Mahaffey KW, et al. Effect of platelet inhibition with cangrelor during PCI on ischemic events. N Engl J Med 2013;368:1303-13.
12. Fox KAA, Ford I, Steg PG, Tendera M, Ferrari R. Ivabradine for patients with stable coronary artery disease and left-ventricular systolic dysfunction (BEAUTIFUL): a randomised, double-blind, placebo-controlled trial. Lancet 2008;372:807-16.
13. Fox K, Ford I, Steg PG, Tardif J-C, Tendera M, Ferrari M. Ivabradine in stable coronary artery disease without clinical heart failure. N Engl J Med 2014;371:1091-9.
14. Swedberg K, Komajda M, Böhm M, et al. Ivabradine and outcomes in chronic heart failure (SHIFT): a randomised pla-

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

cebo-controlled study. Lancet 2010;376:875-85.
15. Topol EJ, Moliterno DJ, Herrmann HC, et al. Comparison of two platelet glycoprotein IIb/IIIa inhibitors, tirofiban and abciximab, for the prevention of ischemic events with percutaneous coronary revascularization. N Engl J Med 2001;344:1888-94.
16. Cohn JN, Pfeffer MA, Rouleau J, et al. Adverse mortality effect of central sympathetic inhibition with sustained-release moxonidine in patients with heart failure (MOXCON). Eur J Heart Fail 2003;5:659-67.
17. Bonaca MP, Bhatt DL, Cohen M, et al. Long-term use of ticagrelor in patients with prior myocardial infarction. N Engl J Med 2015;372:1791-800.
18. Pitt B, Pfeffer MA, Assmann SF, et al. Spironolactone for heart failure with preserved ejection fraction. N Engl J Med 2014;370:1383-92.
19. Pfeffer MA, Claggett B, Assmann SF, et al. Regional variation in patients and outcomes in the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) trial. Circulation 2015;131:34-42.
20. Pfeffer MA, McMurray JJV, Velazquez EJ, et al. Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both. N Engl J Med 2003;349:1893-906.
21. 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.
22. Pocock SJ, McMurray JJ, Collier TJ. Statistical controversies in reporting of clinical trials: part 2 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol 2015;66:2648-62.
23. Mohr FW, Morice MC, Kappetein AP, et al. Coronary artery bypass graft surgery versus percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. Lancet 2013;381:629-38.
24. Kappetein AP,Serruys PWSabik JF, et al. Design and rationale for a randomized comparison of everolimus-eluting stents and coronary artery bypass graft surgery in selected patients with left main coronary artery disease: the EXCEL trial. EuroIntervention 2016 (in press).
25. Dahlöf B, Sever PS, Poulter NR, et al. Prevention of cardiovascular events with an antihypertensive regimen of amlodipine adding perindopril as required versus atenolol adding bendroflumethiazide as required, in the Anglo-Scandinavian Cardiac Outcomes Trial-Blood Pressure Lowering Arm (ASCOT-BPLA): a multicentre randomised controlled trial. Lancet 2005;366:895-906.
26. Valgimigli M, Frigoli E, Leonardi S, et al. Bivalirudin or unfractionated heparin in acute coronary syndromes. N Engl J Med 2015;373:997-1009.
27. Stone GW, Witzenbichler B, Guagliumi G, et al. Bivalirudin during primary PCI in acute myocardial infarction. N Engl J Med 2008;358:2218-30.
28. The Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) Investigators. High-dose atorvastatin after stroke or transient ischemic attack. N Engl J Med 2006;355:549-59.
29. Gupta SK. Intention-to-treat concept: a review. Perspect Clin Res 2011;2:109-12.
30. Velazquez EJ, Lee KL, Deja MA, et al. Coronary-artery bypass surgery in patients with left ventricular dysfunction. N Engl J Med 2011;364:1607-16.
31. Velazquez EJ, Lee KL, Jones RH, et al. Coronary-artery bypass surgery in patients with ischemic cardiomyopathy. N Engl J Med 2016;374:1511-20.
32. The BARI 2D Study Group. A randomized trial of therapies for type 2 diabetes and coronary artery disease. N Engl J Med 2009;360:2503-15.
33. Rogers JK, Pocock SJ, McMurray JJ, et al. Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. Eur J Heart Fail 2014;16:33-40.
34. Knopp RH, d'Emden M, Smilde JG, Pocock SJ. Efficacy and safety of atorvastatin in the prevention of cardiovascular end points in subjects with type 2 diabetes: the Atorvastatin Study for Prevention of Coronary Heart Disease Endpoints in non-insulin-dependent diabetes mellitus (ASPEN). Diabetes Care 2006;29:1478-85.
35. Colhoun HM, Betteridge DJ, Durrington PN, et al. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. Lancet 2004;364:685-96.
36. Rader DJ, deGoma EM. Future of cholesteryl ester transfer protein inhibitors. Annu Rev Med 2014;65:385-403.
37. The CAPRICORN Investigators. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. Lancet 2001;357:1385-90.
38. Bhatt DL, Kandzari DE, O'Neill WW, et al. A controlled trial of renal denervation for resistant hypertension. N Engl J Med 2014;370:1393-401.
39. Fröbert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med 2013;369:1587-97.
40. Jolly SS, Cairns JA, Yusuf S, et al. Outcomes after thrombus aspiration for ST elevation myocardial infarction: 1-year follow-up of the prospective randomised TOTAL trial. Lancet 2016;387:127-35.
Copyright © 2016 Massachusetts Medical Society.

MY NEJM IN THE *JOURNAL* ONLINE
Individual subscribers can store articles and searches using a feature on the *Journal*'s website (NEJM.org) called "My NEJM." Each article and search result links to this feature. Users can create personal folders and move articles into them for convenient retrieval later.

The New England Journal of Medicine
Downloaded from nejm.org at LONDON SCH HYGIENE & TROPICAL MED on October 13, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.