Challenging Issues in Clinical Trial Design:
Part 4 of a 4-part Series on Statistics for Clinical Trials

Brief title: Challenges in Trial Design

Stuart J. Pocock, PhD,* Tim C. Clayton, MSc,* Gregg W. Stone, MD†

From the: *London School of Hygiene and Tropical Medicine, London, United Kingdom; †Columbia University Medical Center, New York-Presbyterian Hospital and the Cardiovascular Research Foundation, New York, New York

<COR> Reprint requests and correspondence:
Prof. Stuart J. Pocock,
Department of Medical Statistics,
London School of Hygiene and Tropical Medicine,
Keppel Street,
London, WC1E 7HT, United Kingdom
Telephone: +44 20 7927 2413
Fax: +44 20 7637 2853
E-mail: stuart.pocock@lshtm.ac.uk

Disclosures:

The authors declare no conflicts of interest for this paper.

Abstract

As a sequel to last week's article on the fundamentals of clinical trial design, this article tackles related controversial issues: noninferiority trials; the value of factorial designs; the importance and challenges of strategy trials; Data Monitoring Committees (including when to stop a trial early); and the role of adaptive designs. All topics are illustrated by relevant examples from cardiology trials.

<KW>Key words: Noninferiority trials; Factorial designs; Strategy trials, Data Monitoring Committees; Statistical stopping guidelines; Adaptive designs; Randomized Controlled Trials As Topic;

Abbreviations

ACS = acute coronary syndrome
CABG = coronary artery bypass graft
CI = confidence interval
CV = cardiovascular
DMC = Data Monitoring Committee
FDA = Food and Drug Administration
MACE = major adverse cardiovascular event
OMT = optimal medical therapy
PCI = percutaneous coronary intervention

Introduction

Randomized controlled trials are the cornerstone of clinical guidelines informing best therapeutic practices, however their design and interpretation may be complex and nuanced. This review explores challenging issues that may arise and builds on the fundamentals of trial design covered in last week’s paper. Specifically, we offer guidance on how to design and interpret noninferiority trials where the goal is to demonstrate that the efficacy of a new treatment is as good as that achieved with a standard treatment. Factorial trials, where 2 (or more) therapeutic issues are simultaneously evaluated in the same study, present an interesting opportunity that should be considered more often in cardiology research. Trials that compare substantially different alternative treatment strategies can be of great value in enhancing good patient management, and we present guidance on the topic to stimulate greater interest in overcoming the difficulties in undertaking such pragmatic studies. All major cardiology trials have both ethical and practical needs for data monitoring of the accumulating evidence over time. We provide insights into how Data Monitoring Committees (DMCs) should function, offering statistical guidelines and practical decision-making considerations as to when to stop a trial early. Finally, there is a growing interest in adaptive designs, but few instances of their implementation in cardiology trials. We focus on adaptive sample size re-estimation and enrichment strategies, with guidance on when and how they may be used. All of these issues are illustrated by experiences from actual cardiology trials, demonstrating the real-world implications of trial design decisions.

Noninferiority Trials

Increasingly, major trials are conducted to see if the efficacy of a new treatment is as good as a standard treatment (1-3). The new treatment usually has some other advantage (e.g., fewer side effects, ease of administration, lower cost), making it worthwhile to demonstrate noninferiority in respect of efficacy.

The standard approach to designing a noninferiority trial is to predefine a noninferiority margin, commonly called delta, for the primary endpoint. This is the smallest treatment difference, which, if true, would mean that the new treatment is declared inferior. This is based on the belief that any difference smaller than this would constitute clinically accepted grounds of "therapeutic interchangeability" (4). The trial's conclusions then depend on where the 95% confidence interval (CI) for the treatment difference ends up in relation to this margin. If the upper bound of the 2-sided 95% CI is less than delta, one can claim evidence that the new treatment is noninferior.

For instance, the ACUITY trial compared bivalirudin with the standard treatment of heparin plus a glycoprotein IIb/IIIa inhibitor in patients with acute coronary syndrome (ACS) for 30-day composite ischemia (death, myocardial infarction [MI] or revascularization)(5). The noninferiority margin was set at a relative risk of 1.25. The trial's findings revealed composite ischemia rates of 7.8% and 7.3% in the bivalirudin and control groups, respectively (relative risk 1.08; 95% CI: 0.93 to 1.24). Because the upper bound of the CI of 1.24 was less than the pre-declared delta of 1.25, one can conclude that there is evidence of noninferiority. The reason this matters is that bivalirudin also had a markedly lower risk of major bleeding, an important consideration when choosing between antithrombin therapies.

A common misunderstanding is that lack of a statistically significant difference between 2 therapies implies that they are equivalent. For instance, the INSIGHT trial compared nifedipine with co-amilozide in hypertension. The authors concluded that the treatments were "equally effective in preventing cardiovascular complications", on the basis of p = 0.35 for the primary composite endpoint of cardiovascular (CV) death, MI, heart failure, or stroke (6). But the observed relative risk of 1.10 had a 95% CI of 0.91 to 1.34. This includes up to a 34% excess risk on nifedipine, making it unwise to conclude that nifedipine is as good as (i.e., noninferior to) co-amilozide.

Figure 1 shows a conceptual plot of how to interpret the results of noninferiority trials. Scenario C (noninferior) indicates what happened in the ACUITY trial. If we suppose that the INSIGHT trial had the same delta, 1.25, then it would have fallen under scenario F (inconclusive). Had more patients been enrolled, the 95% CI would have narrowed, and noninferiority might then have been declared.

Sometimes, the treatment effect (and its delta) is expressed as a difference in percentages, rather than as a relative risk or hazard ratio (the argument being that absolute differences are more clinically relevant than relative risks). For instance, the OPTIMIZE trial compared a 3-month versus a 12-month duration of dual antiplatelet therapy after implantation of a zotarolimus-eluting stent (7). For the composite primary endpoint of net adverse clinical events (death, MI, stroke, or major bleed) at 1 year, a 2.7% percentage difference was set as the noninferiority margin. The observed difference was +0.2%, with a 95% CI of -1.5% to +1.9%. Because this excludes the margin of +2.7%, noninferiority of the 3-month duration of treatment was claimed.

This example raises a few issues. When the noninferiority margin is a difference in percentages, it becomes easier (perhaps too easy) to achieve noninferiority if the overall event rate is lower than expected. OPTIMIZE had an anticipated 9% event rate in the control arm, but the observed event rate was 6%. This made the 2.7% margin equivalent to a relative risk margin of 1.45, which is undesirably large. Conversely, if the overall event rate is greater than expected, it may become unreasonably difficult to achieve noninferiority. The opposite considerations of anticipated versus observed event rates apply if a relative risk is chosen for the margin.

Also, the endpoint chosen in OPTIMIZE was not of optimal relevance. The true issue in considering a shorter period of dual antiplatelet treatment concerns the balance between the increased risks of stent thrombosis and MI against the reduced risk of major bleeding. To force these diverse endpoints into a single composite would bias results toward the null. A preferable approach is to pre-specify and study separately-powered efficacy and safety endpoints, typically one for superiority and one for noninferiority. However, a very large sample size may be required to adequately power both the efficacy and safety endpoints.

A composite net adverse clinical events endpoint, consisting of combined safety and efficacy endpoints, has been used in some trials, reflecting the recognition that both types of endpoints (e.g., major bleeding and stent thrombosis) are deleterious and strongly associated with subsequent mortality. However, interpretation of such a combined safety and efficacy endpoint may be challenging, especially if the different components do not have similar impacts on patient well-being or survival. Moreover, because safety and efficacy endpoints often move in different directions (e.g., in response to more potent antithrombotic therapies), their combination in a composite endpoint may mask differences between therapies, making careful examination of each component measure essential.

A key question is the choice of noninferiority margin, which has implications for the required trial size. Power calculations for noninferiority trials (not presented here) indicate that trial size is inversely proportional to the square of the margin delta. For instance, had ACUITY chosen a 10% increase, rather than a 25% increase (i.e., relative risk 1.1, rather than 1.25), more than 6 times as many patients would have been required for the same power (i.e., over 50,000 in total). Thus, the choice of margin requires a realistic balancing of scientific goals with an achievable sample size.

The choice of margin is sometimes related to prior knowledge of the efficacy of the active control compared to placebo. A sensible goal is that the new treatment should preserve at least 50% of the effect demonstrated in prior trials of the control treatment against placebo (the so-called “putative placebo” approach). For instance, in the CONVINCE trial of verapamil versus standard antihypertensive treatment with a diuretic or beta-blocker, the noninferiority margin for the composite of stroke, MI, or CV death was set at a hazard ratio of 1.16 (8). This was because of the need for evidence that verapamil was at least half as effective as the standard treatment, relative to placebo. Regulatory agencies accept this method to establish a noninferiority margin, and provide guidance for its determination (1).

In addition to the assumed event rates, margin, and desired power, the sample size of a noninferiority trial depends on whether the delta will be tested against the upper bound of a 1- or 2-sided 95% CI (the latter being equivalent to a 1-sided 97.5% confidence limit). The latter conservative approach is the standard for regulatory approval of new pharmaceuticals (and many devices). However, some devices, such as the FilterWire EX system to prevent distal embolization during percutaneous coronary intervention (PCI) of diseased saphenous vein grafts in the FIRE trial (9), have been approved on the basis of a noninferiority design with a 1-sided alpha of 5%. Utilizing a 1-sided alpha of 5%, rather than 2.5%, reduces the sample size by approximately 20%, although this is generally frowned upon. Accepting greater alpha error may be acceptable, however, when the experimental device provides additional benefits not evident in the primary endpoint.

A noninferiority design may also be applied to exclude a safety concern in a treatment with known efficacy. Such safety trials can include comparison of the experimental agent to an active comparator. (e.g., as in the ENTRACTE trial performed to exclude excess CV risk for tocilizumab compared to etanercept in patients with rheumatoid arthritis) (10). But in type 2 diabetes, Food and Drug Administration (FDA) guidance requires assessment of the CV risk of any new drug relative to placebo (11). Many such placebo-controlled trials in high-risk patients already on appropriate antiglycemic therapy are either currently in progress or recently completed. The primary safety endpoint is typically the composite of CV death, MI, and stroke, and the noninferiority margin is set at a hazard ratio of 1.3. This requires a trial of many thousands of patients because approximately 700 primary events are needed to provide convincing evidence of noninferiority. For a new, effective antidiabetic drug, the FDA also requires preliminary evidence of CV safety for initial approval, using a hazard ratio noninferiority margin of 1.8. The larger safety trial to confirm noninferiority on the basis of the tougher margin of 1.3 then ensues.

It is sometimes argued that noninferiority trials should emphasize a per-protocol (or as-treated) analysis, rather than analysis by intention-to-treat, thereby excluding any follow-up after a patient withdraws from randomized treatment (or after a short period following withdrawal to capture rebound events). The logic is that including off-treatment follow-up (possibly with crossovers) may dilute any real treatment differences, thereby artificially enhancing any claim of noninferiority. However, per-protocol and as-treated analyses introduce other biases. We suggest that both types of analyses be presented in noninferiority trials, hopefully demonstrating a consistency of findings.

When undertaking a noninferiority trial, one can also propose a superiority hypothesis, with no statistical penalty. That is, once the trial results confirm noninferiority, one can go on to test for superiority (see scenario A in **Figure 1**). For instance, some CV safety trials of antidiabetic drugs have been made larger to accommodate this superiority hypothesis. One such trial (EMPA-REG OUTCOME) of empagliflozin versus placebo recently demonstrated some evidence of a reduction in the primary endpoint of CV death, MI, or stroke, with a hazard ratio of 0.86 (95% CI: 0.74 to 0.99; p = 0.04), while also showing a significant reduction in all-cause death with a hazard ratio of 0.68 (95% CI: 0.57 to 0.82, p <0.001)(12).

Factorial Designs

Sometimes, one can pursue 2 separate treatment comparisons within the same major trial by randomizing each patient twice: once to treatment A versus its control and, at the same time, to treatment B and its control. This is known as a 2-way factorial design (13,14).

Factorial designs have numerous practical benefits, such as adding in a second randomization within the framework of a trial funded for a different purpose, affording the opportunity to investigate an inexpensive treatment that would otherwise be difficult to fund and test in its own trial. For instance, the HOPE factorial trial studied ramipril versus placebo and then also vitamin E versus its placebo in high-risk patients (15,16). Ramipril significantly reduced CV events, whereas vitamin E did not.

In planning a factorial design, one presumes that the treatment effect in 1 randomized comparison is not likely to depend on the other randomized treatment: that is, there is no expectation of an interaction between the 2 randomized treatments. Thus, the trial is powered to examine the main effects of the 2 randomized comparisons separately. By doing so, one neatly gets “2 trials for the price of 1”; that is, in principle adding in the second randomization does not increase the trial size. In practice, it may be wise to somewhat inflate trial size when a factorial design is contemplated because: 1) if both treatments are effective, the overall event rate will be lower; and 2) one may wish to guard against a modest quantitative interaction being present.

The CURRENT OASIS 7 trial randomized 25,086 ACS patients referred for an invasive strategy to both: 1) double-dose versus standard-dose clopidogrel; and 2) higher-dose versus lower-dose aspirin (17). The primary outcome was CV death, MI, or stroke within 30 days, and the findings are shown in **Table 1**. The 2 main effect analyses showed that neither the clopidogrel dose nor the aspirin dose appeared to have any effect on the primary endpoint, p = 0.30 and p = 0.61 respectively. Exploring the potential interaction between the 2 drug doses, however, revealed a curious finding: the observed event rate was lower on double-dose than standard-dose clopidogrel (3.8% vs. 4.6%) when given with higher-dose aspirin, but this was reversed (4.5% vs. 4.2%) when given with lower-dose aspirin. This apparent qualitative interaction did reach conventional statistical significance, interaction p = 0.04. The authors believed that this unexpected finding lacks a known biological mechanism and may be due to the play of chance, which is a reasonable supposition. Conversely, if a possible biological explanation for the interaction may be posited, the validity of the conclusions drawn from both arms may be jeopardized, an inherent risk of factorial designs. Factorial designs should therefore only be contemplated when the expectation of a real interaction between the 2 therapies is low. In principle, one can still undertake a factorial trial when a plausible interaction between the 2 treatment factors is contemplated, but this would require a major increase in trial size to be adequately powered to detect such an interaction.

Another useful option is a partial (or nested) factorial design, where all recruited patients get 1 random treatment allocation, but only some patients are eligible for the second randomized treatment. For instance, the HORIZONS-AMI trial randomized 3,602 ST-segment elevation MI patients to bivalirudin versus heparin plus a glycoprotein IIb/IIIa inhibitor (in a 1:1 ratio)(18,19). Among these patients, 3,006 met additional anatomic inclusion criteria and underwent a second randomization to PCI with paclitaxel-eluting versus bare-metal stents (in a 3:1 ratio).

Occasionally the factorial design can take on more than 2 treatment factors. For instance, the ISIS 4 trial randomized 58,050 patients with MI to: 1) oral captopril versus placebo; 2) oral mononitrate versus placebo; and 3) intravenous magnesium sulfate versus open control in a 2 × 2 × 2 factorial design (20). Finally, the MATRIX trial is an example of a 3-level randomization with a nested factorial approach. In MATRIX, 8,404 patients with ACS undergoing cardiac catheterization were randomized to radial versus femoral vascular access. Among this group, 7,213 patients in whom PCI was selected for treatment were randomized again to procedural anticoagulation with heparin versus bivalirudin. Finally, the

3,610 bivalirudin-assigned patients were randomized a third time to either a post-procedural prolonged bivalirudin infusion or to no infusion (21,22).

When circumstances are right, the factorial design is a useful means of investigating 2 (or more) different treatment innovations within 1 trial. Overall, trialists need to give more attention to the imaginative use of factorial designs.

Trials of Alternative Treatment Strategies

Trials of fundamentally different treatment strategies, for example, surgery versus PCI or medical therapy, or invasive versus conservative approaches in patients with ACS, are an exciting challenge and can make a substantial impact on guidelines and clinical practice (23,24). Such “strategy” trials are, however, more difficult to undertake than studies comparing different drugs or different devices to each other.

When the randomized strategies differ substantially in their perception by both investigators and patients, particular challenges arise. Investigators (often across specialties [e.g., cardiac surgeons and interventional cardiologists]) need to accept that the patient may truly receive either strategy without being disadvantaged (i.e., a state of equipoise is indeed present). Even if solid evidence is lacking, physicians (and patients) may express strongly held beliefs in the superiority of one treatment compared to another, based either on anecdotal experiences or reports, nondefinitive evidence (e.g., uncontrolled observational comparisons or small randomized trials), or prior positive trials using surrogate endpoints. These preconceived beliefs can make enrollment more difficult, and may result in a biased cohort being recruited. Obtaining informed patient consent is also less routine in strategy trials than in standard randomized drug or device studies. Strategy trials also typically require multidisciplinary cooperation, greater resources, and a longer period for full recruitment, and are thus more expensive. Strategy trials often lack a single funding source from industry, and therefore often require pure governmental and/or institutional support, collaboration between multiple companies, or a private-public partnership. Thus, major challenges in strategy trials include randomizing a high enough proportion of eligible patients in a reasonable timeframe, and raising appropriate funds.

For instance, the ISCHEMIA trial is a major multinational trial of routine invasive versus conservative strategies in patients with stable coronary disease and at least moderate ischemia (25). A strong evidence-based case can be made for either approach in such patients (26). A prior survey of interested cardiologists asked if they would enroll their eligible patients in a randomized trial with a 50% chance of being conservatively managed without cardiac catheterization; 80% responded positively (27). ISCHEMIA initially planned to recruit 8,000 patients, but after more than 2 years, only ~2,000 patients have been randomized, which may require a protocol amendment to reduce the sample size. Such lower than desired recruitment is a common problem with strategy trials.

Strategy trials are particularly important when evaluating a new therapeutic approach. For instance, transcatheter aortic valve replacement has emerged as an alternative to surgical aortic valve replacement in patients at high and prohibitive operative risk (28,29). Ongoing trials are now being performed in patients at lower surgical risk. Key aspects here are to decide when in the learning curve of such a new technology one should undertake such a trial; to define the risk profile of patients that should initially be recruited; and to create the right collaborative atmosphere for general cardiologists, interventionalists, and surgeons to participate.

The results of strategy trials require careful interpretation, especially when crossovers occur. For instance, the COURAGE trial studied optimal medical therapy (OMT) with and without initial PCI in 2,287 patients with stable coronary disease (30). The primary endpoint, the composite rate of death or nonfatal MI, showed no significant difference between the PCI and medical therapy groups after a median 4.6 years of follow-up. A naive interpretation is that PCI is no better than medical therapy (and thus PCI should never be performed), but this ignores the strategic concept of the trial. In COURAGE, 32.5% of patients assigned to OMT went on to receive revascularization (mostly PCI) during follow-up, primarily for progressive or unstable symptoms. Thus, the trial really compared “PCI (plus OMT) now” with “OMT now, with the option of later PCI (or coronary artery bypass graft (CABG)), as needed.” The pure question “does PCI improve prognosis?” is not directly answerable because the investigators could not continue with medical therapy alone.

An additional concern of particular relevance to strategy trials is that given their inherently protracted nature (slow recruitment with long follow-up), the standard of care frequently evolves prior to their finish. For instance, in the SYNTAX trial, CABG was shown to be superior to PCI using a first-generation paclitaxel-eluting stent (31). However, by the time SYNTAX was completed, second-generation drug-eluting stents had been developed, which have been associated with reduced rates of death, MI, and repeat revascularization compared with paclitaxel-eluting stents (32). Studies have suggested that this advance alone might have eliminated the difference between the 2 strategies (33). Confirming such a hypothesis requires performance of another time-consuming and costly randomized trial, which, in turn, risks further advances in technology before its completion.

Despite the practical difficulties in undertaking randomized trials of alternative strategies, they are of key importance in evaluating radically different approaches to patient care. Otherwise, we are forced to rely on nonrandomized comparisons on the basis of patient registries. They, too, provide a wealth of interesting data, but always with the caveat that substantial selection bias is typically present, resulting in unmeasured confounders that cannot be accounted for in statistical analysis (34,35).

One exciting development is the growth of pragmatic trials that are embedded within routine care delivery (i.e., trials with patient registries, such as the TASTE trial of thrombus aspiration for MI (36)). Such trials greatly enhance patient representativeness, recruitment, and follow-up, with associated reduced trial costs. However, they are best suited to assess endpoints reliably tracked in administrative databases, such as all-cause mortality.

Data Monitoring for Efficacy, Safety and Futility

Most major randomized trials require interim analyses of the accumulating outcome data by treatment group. Such unblinded interim analyses are produced by an independent statistician and are evaluated by an independent DMC, comprising several clinicians plus a statistician, all of whom have no other involvement in the trial and operate under strict confidentiality (37,38).

The main DMC responsibility is to protect patient safety, that is, to identify and react to any evidence of harm occurring to patients, especially on the new treatment. Adverse events may relate to predefined safety issues (e.g., bleeding on antiplatelet drugs), unexpected event types, or inferiority with regard to primary or secondary event outcomes. The DMC should meet regularly so that any ethical concerns as regards potential harm can be dealt with in a timely fashion. If safety issues become evident, the DMC may request more data analyses, and schedule follow-up meetings more frequently. The DMC can recommend to the study leadership that the trial be stopped or altered. However, given the likelihood of chance variations in repeated looks at accumulating data, major alterations should only be recommended if truly convincing evidence of harm is present, with a lower threshold to modify or stop the trial for concerns relating to increased mortality, as opposed to other endpoints.

A second DMC responsibility may be to evaluate whether there is overwhelming evidence for superiority of the new treatment, which is sufficiently convincing to merit stopping the trial early. However, trials that are stopped early tend to overestimate true treatment effects. Thus, early trial stoppage should only be recommended for situations in which continuing would truly place the control group patients at harm (e.g., increased mortality, resulting in an ethical imperative to unblind and expedite approval of the experimental treatment).

Sometimes there is a third futility issue for the DMC to consider. That is, does the accumulating evidence indicate that the new treatment lacks efficacy? If there is little chance of the trial achieving a clinically-relevant positive outcome, the trial may be stopped early for futility. Such a decision needs careful consideration, as even if the primary endpoint lacks efficacy, secondary endpoints with real clinical value may emerge as positive (even if only hypothesis-generating).

A further DMC responsibility is to look at trial quality issues. For instance, if problems with noncompliance, missing visits/data, or slowness in event adjudication are evident, the DMC should provide feedback to the study leadership to facilitate improvements.

After every interim report and meeting, the DMC needs to promptly communicate its recommendations to the trial's principal investigator (e.g., the chair of the Executive Committee) or, sometimes, directly to the trial sponsor in writing (or sooner by phone, if major issues of patient safety are apparent).

All DMC-related activities should be documented in a DMC Charter (39). This should include any statistical stopping guidelines (40), recognizing that these are not formal rules: the recommendation to stop rests on the wise judgment of the DMC, on the basis of the totality of evidence at their disposal, both within the trial and externally. Note that the DMC only makes recommendations: any decisions on stopping or modifying the trial are the responsibility of the trial Executive Committee or sponsor. So what makes for sensible statistical stopping guidelines?

First, **stopping for superiority** of a new treatment requires proof beyond a reasonable doubt. For example, $p < 0.001$ is often used, or even $p < 0.0001$ at a relatively early interim analysis. Furthermore, it is wise not to look too early or too often for superiority: 2 or 3 interim looks should suffice. For instance, the PARADIGM-HF trial of LCZ696 versus enalapril in chronic heart failure required p <0.001 for both the composite primary endpoint (CV death or hospitalization for heart failure) and CV death alone at its second interim analysis, when two-thirds of primary events had occurred (41). Both boundaries were crossed, and the DMC duly recommended stopping.

Of note, achieving a statistical guideline does not automatically mean the trial is stopped. For instance, in the SHIFT trial of ivabradine versus placebo, superiority was present at the second planned interim analysis for both the composite primary endpoint (CV death and hospitalization for heart failure) and all-cause death: p <0.0001 and p = 0.0014, respectively (42). The predefined stopping boundary was p <0.001 for the primary endpoint. However, the DMC recommended continuation: there were only a few months to go to complete enrollment; important subgroup issues needing resolving; event adjudication was incomplete; and a previous related trial (BEAUTIFUL) had been neutral (43). Upon trial completion, the primary endpoint finding was confirmed, but all-cause mortality was no longer significant (p = 0.09). Such “regression to the truth” may often arise. That is, interim findings that cross a stopping boundary may be “on a random high,” so that subsequent results (if the trial continues) may end up less impressive (44).

Secondly, **stopping for futility** has 2 types of statistical guidelines (40,45). One approach is to see if the 95% CI for the primary endpoint effect estimate excludes a predeclared minimum benefit and then stop the trial early. For instance, in the PERFORM trial of terutroban versus aspirin in patients with cerebral ischemic events, the primary endpoint was the composite of CV death, MI, or ischemic stroke (46). At the 20th safety report, the hazard ratio was 1.04 (95% CI: 0.95 to 1.14). This excluded the predefined 7%

benefit (i.e., a hazard ratio of 0.93), and so the DMC recommended that the trial be stopped for futility.

An alternative approach uses conditional power: that is, if the interim data indicate only a slim chance of achieving statistical significance upon trial completion, then stopping early for futility may be reasonable. This method was applied in the RED-HF trial of darbepoetin alfa versus placebo in heart failure patients with anemia (47). Futility was considered at each interim analysis: if the conditional power under the protocol-specified hazard ratio of 0.8 for the composite primary endpoint (death or heart failure hospitalization) was <30%, then the DMC could recommend the trial be stopped. This boundary was eventually crossed, but the DMC decided to allow the trial to continue: there were no safety concerns and there were significant quality of life improvements (a secondary endpoint).

Thirdly, **stopping for safety** usually requires more frequent looks at interim data because there is an ethical obligation to stop promptly if a new treatment is causing harm (48). Also, the stopping boundary needs to be less stringent; for example, p <0.01 going the wrong way for the primary endpoint or all-cause mortality is a useful simple guideline. For instance, in the ILLUMINATE trial of torcetrapib versus placebo in high-risk patients, the DMC observed 82 deaths in the treatment arm versus 51 deaths with control (p = 0.007), which was the prime reason for stopping the trial for harm (49). As a consequence, the sponsor withdrew the drug immediately from any further investigation worldwide.

Similarly, the PALLAS trial of dronedarone versus placebo in permanent atrial fibrillation was stopped early when both coprimary endpoints of: 1) stroke, MI, systemic embolism, or CV death; and 2) unplanned hospitalization for a CV cause or death, demonstrated an excess on dronedarone, both p <0.01 (50). This was particularly surprising, given that the earlier ATHENA trial of dronedarone in nonpermanent/paroxysmal atrial fibrillation had shown a highly significant benefit (51). This illustrates the importance of the safety role of a DMC, no matter how promising the prior evidence from other sources.

Stopping early for harm may not relate to the efficacy endpoints, but to specific safety problems instead. For instance, at an early interim report, the APPRAISE 2 trial of apixaban versus placebo in ACS patients showed significant increases in major bleeding events on apixaban (52). Numbers of events were small, but given that the primary efficacy endpoint of CV death, MI, or ischemic stroke had thus far showed no benefit, this safety signal was deemed sufficient to halt the trial. In such scenarios of potential harm, it is difficult to have a statistical stopping guideline that adequately captures the ethical concern, which needs balancing against potential benefit regarding efficacy endpoints. Such matters depend on an experienced DMC acting wisely, being fully aware of the ethical and practical consequences of its actions.

Let us conclude this section with potential stopping guidelines for a planned placebo-controlled trial of a new drug for patients at high CV risk. The trial is to recruit 13,000 patients and completion is planned when 1,600 primary MACE (major adverse CV events) have occurred, anticipated to be over 5 years duration in total. This gives 90% power to detect a 15% risk reduction (i.e., hazard ratio: 0.85). The trial plans to have 2 interim analyses, after 50% and 75% of primary events have occurred, and the proposed stopping boundaries for superiority and for futility are shown in **Table 2**.

First, the timing of these boundaries recognizes that stopping for either superiority or futility should not be contemplated before at least half the trial's evidence has accumulated. The superiority guideline of p <0.0002 reflects the spirit of only stopping when there is overwhelming evidence. It is interesting to note that to stop early, the hazard ratio for the MACE primary endpoint at the 2 interim looks would need to be <0.768 and <0.806, respectively, considerably more beneficial than the 0.85 hazard ratio used in the power calculation. Given the tough stopping boundary, the final $p < 0.05$ for a positive outcome is not compromised, and with 1,600 primary events, an observed hazard ratio <0.906 would reach statistical significance.

The stopping guidelines for futility in **Table 2** are on the basis of conditional power calculations. With 50% of the event data in (800 primary endpoint events), if the hazard ratio is only very slightly in a positive direction (hazard ratio >0.979) or in the opposite direction, then the trial may stop for futility. Adding a further 25% of events at the second interim analysis (1,200 events), one needs a somewhat stronger indication of treatment benefit to continue: hazard ratio >0.931 is considered sufficient to stop for futility. Note these are not intended as absolute rules. There may be other issues (secondary endpoints, safety concerns, subgroup findings, external evidence) that could sway the totality of evidence in a positive or negative direction.

Lastly, note the lack of any formal stopping boundaries for safety. Experience dictates that it is impractical to capture all the scenarios and nuances of potential harms in statistical guidelines. Rather, the trial DMC will receive frequent safety reports every few months, and collectively make judgments on the strength of evidence and the absolute magnitude and seriousness of any safety signals.

Adaptive Designs

The conventional wisdom in clinical trial design is that once the study protocol is finalized, the trial should proceed with no further changes to its intent. Protocol amendments are permitted under certain circumstances, but should be made without knowledge of interim results by treatment groups: that is, the DMC should have no involvement in such changes. Such amendments may be of a practical nature; for example clarifications of patient eligibility, endpoint definitions, or drug dose modifications. Amendments in response to knowledge of ongoing blinded results for all treatments combined are also permitted. For instance, if the incidence of the primary endpoint pooled across randomized groups is substantially lower than anticipated, the target sample size might be increased, the eligibility criteria might be changed to recruit higher-risk patients, the duration of follow-up might be prolonged, or the primary endpoint might even be altered (e.g., by expanding a composite to include additional types of outcomes). In principle, such adaptations are acceptable and carry no statistical penalties, although they may prompt concerns that someone involved had an awareness of unblinded results. In particular, changing the primary endpoint often evokes suspicion, even if unwarranted.

An emerging and more controversial type of adaptive design is where protocol changes are made on the basis of the unblinded interim results (53,54). Both European and U.S. regulators have issued guidance on the use (and possible misuse) of such adaptations (55,56). It is key that any such potential changes should be predefined in an Adaptive Charter; that they should not affect the trial's overall integrity; and that they should preserve statistical rigor: that is, an unbiased verdict is still reached on the treatments' relative merits.

The most common adaptation using unblinded data concerns **sample size re-estimation** (57). Other types of proposed adaptive designs (54) include seamless phase II/III trial designs, whereby from multiple new treatments (e.g., different drug doses), one drops some arms at interim analysis on the basis of a surrogate outcome, thereafter examining clinical outcomes (58); enrichment designs, in which after the adaptation, selected subgroups of patients are preferentially enrolled in whom the event rates were observed to be high or evidence of treatment effect appeared particularly robust (59); and "play the winner," whereby the randomization ratio is adjusted to put a higher proportion of future patients on the treatment with better interim results (60). All have a methodological appeal, but introduce logistical and interpretive challenges.

Hence, we now concentrate on adaptive sample size re-estimation. The logic is that if the observed treatment difference for the primary endpoint at a preplanned interim analysis is somewhat smaller than that assumed in the original power calculation, trial size may be increased to provide adequate power to detect such a more modest treatment effect. For this approach to be valid, the interim results need to be in a “promising zone”: that is, 1) the observed interim treatment difference, although smaller than hoped for, is still trending in the right direction and is big enough to be of clinical relevance, and 2) the expansion in sample size takes the conditional power from a current 50%+ to a desired 80%, or higher. Then the type I error may be preserved without any statistical adjustments. A sample size increase could also be considered if the effect size is preserved, but the endpoint rates at the interim analysis are lower than anticipated.

Figure 2 gives a conceptual outline of how adaptive sample size re-estimation could work. Suppose an interim analysis is performed after half the original trial’s results are known, and that you are prepared to increase the size (if necessary) up to double that originally planned. Then, whether to make any size increase depends on how the observed treatment difference compares to the preplanned treatment difference used in the original power calculation. If these rates are at least similar, then the trial is “on track” and there is no need to increase trial size. We call this the Favorable Zone: in **Figure 2,** this extends to point B, where the observed difference is approximately 90% of pre-planned difference A.

The Promising Zone refers to the scenario where the observed difference is less than hoped for, but conditional power can still be upped to the desirable 90% by increasing the sample size. This works fine if the observed difference is at least 66% of the pre-planned difference (point C in **Figure 2**) for which a doubling of size is needed.

One can then extend the Promising Zone into less optimistic territory, where a doubling is still to be done, even though the conditional power cannot make it to the desired

90%. For instance, point D in **Figure 2** occurs when the interim difference is only slightly more than half of the preplanned difference. Doubling the trial size can raise the conditional power up to more than 50%; not hopeless, but a gamble as to whether the trial will end up positive. If the interim results are worse than that, then one is in the Unfavorable Zone. The trial then continues to its original size (unless findings are very unfavorable, in which case stopping for futility may be considered). Note that **Figure 2** is just conceptual: precise statistical details would need to be calculated (57) and specified in an Adaptive Charter.

Preplanned adaptive sample size re-estimation has been used in 2 trials of cangrelor versus clopidogrel in PCI patients. In the CHAMPION-PCI trial, after 70% of patients were enrolled, an interim analysis of the 48-h primary endpoint was performed to determine whether the intended sample size (9,000 patients) needed expanding up to a maximum of 15,000 (61,62). The Adaptive Charter also considered potential enrichment with more diabetic, troponin-positive, or clopidogrel-naive patients if it would enhance statistical power. Unfortunately, there was no interim evidence that cangrelor was superior to clopidogrel, and the trial was stopped early for futility.

The more recent CHAMPION-PHOENIX trial also planned for adaptive sample size re-estimation, but, in this instance, because the interim analysis showed clear evidence of cangrelor's superiority, there was no need to expand beyond the original sample size target of 10,900 patients (63).

These 2 examples provide a reality check to the burgeoning enthusiasm some trialists express about adaptive designs. If a trial is well planned, with a realistic size and alternative hypothesis, then the "promising zone" needing actual expansion of trial size is a relatively narrow window of opportunity. We favor incorporating preplanned adaptive sample size re-estimation into clinical trial designs, but investigators should realize that the likelihood of actively changing the study size or patient eligibility composition (enrichment) is modest.

Thus, organizational simplicity, rather than complex statistical algorithms, is recommended. Also, these calculations can be nuanced, and a statistician experienced in adaptive design methodology should be involved.

Despite these caveats, small biotechnology or medical device companies that do not have the initial resources to plan an appropriately large trial upfront often consider an adaptive approach. Thus, they start with a smaller trial with a potentially unrealistic treatment-effect size, and then use “positive” interim data to persuade funders to expand the trial. This raises an important concern about adaptive designs: the implicit leaking of interim findings beyond the strict confidentiality of the DMC. Only the adaptive decision-makers should be privy to interim results. If the rationale for a trial’s adaptive increase in size is known, people will infer the nature of the interim findings. It is a matter of debate as to whether such wider leakage compromises the trial’s integrity (e.g., by altering patient recruitment patterns).

Conclusions

The **Central Illustration** summarizes the key issues in the diverse collection of design topics we have tackled. In this series of 4 consecutive articles on clinical trials (2 on analysis and reporting, 2 on design) the aim has been to cover those statistical and scientific issues of importance, with a focus on practical insights of relevance to cardiologists.

There is a substantial literature of a more technical nature that statisticians need to master, but such issues tend to be secondary in importance compared with grasping the essential nontechnical factors we have discussed, many of which represent the application of common sense to trial design and statistics. Some topics we chose not to tackle. For example, Bayesian methods are absent, partly because it is hard to do them justice in a few pages, but also reflecting our view that they have a limited role: there is a paucity of examples where their use in cardiology trials achieved insights not reachable by conventional methods.

It is our hope that this series may help clinical trialists and sponsors to more effectively design studies, statisticians interfacing with study leadership to bring forward the most relevant issues to jointly address, and cardiologists to critically interpret and appraise published studies so as to effectively translate clinical trial evidence to patient care.

REFERENCES

1. Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for Industry: Non-Inferiority Clinical Trials. Draft Guidance. 2010. Available at: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm202140.pdf. Accessed November 2, 2015.
2. Piaggio G, Elbourne DR, Pocock SJ, et al.; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012;308:2594-604.
3. Mulla SM, Scott IA, Jackevicius CA, et al. How to use a noninferiority trial: users' guides to the medical literature. JAMA 2012;308:2605-11.
4. Ware JH, Antman EM. Equivalence Trials. N Engl J Med 1997;337:1159-61.
5. Stone GW, McLaurin BT, Cox DA, et al.; ACUITY Investigators. Bivalirudin for patients with acute coronary syndromes. N Engl J Med 2006;355:2203-16.
6. Brown MJ, Palmer CR, Castaigne A, et al. Morbidity and mortality in patients randomised to double-blind treatment with a long-acting calcium-channel blocker or diuretic in the International Nifedipine GITS study: Intervention as a Goal in Hypertension Treatment (INSIGHT). Lancet 2000;356:366-72.
7. Feres F, Costa RA, Abizaid A, et al.; OPTIMIZE Trial Investigator. Three vs twelve months of dual antiplatelet therapy after zotarolimus-eluting stents: the OPTIMIZE randomized trial. JAMA 2013;310:2510-22.
8. Black HR, Elliott WJ, Grandits G, et al.; CONVINCE Research Group. Principal results of the Controlled Onset Verapamil Investigation of Cardiovascular End Points (CONVINCE) trial. JAMA 2003;289:2073-82.

9. Stone GW, Rogers C, Hermiller J, et al.; FilterWire EX Randomized Evaluation (FIRE) Investigators. Randomized comparison of distal protection with a filter-based catheter and a balloon occlusion and aspiration system during percutaneous intervention of diseased saphenous vein aorto-coronary bypass grafts. Circulation 2003;108:548-53.

10. Hoffman-LaRoche. A Study of RoActemra/Actemra (Tocilizumab) in Comparison to Etanercept in Patients With Rheumatoid Arthritis and Cardiovascular Disease Risk Factors. 2015. Available at: https://clinicaltrials.gov/ct2/show/NCT01331837. Accessed November 2, 2015.

11. Food and Drug Administration, Center for Drug Evaluation and Research (CDER). Guidance for Industry: Diabetes Mellitus — Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes. 2008. Available at: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071627.pdf.. Accessed November 2, 2015.

12. Zinman B, Wanner C, Lachin JM, et al. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. N Engl J Med 2015 Sep 17 [E-pub ahead of print], http://dx.doi.org/0.1056/NEJMoa1504720. Accessed November 2, 2015.

13. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol 2003;3:26.

14. Lubsen J, Pocock S. Factorial trials in cardiology: pros and cons. Eur Heart J 1994;15:585-8.

15. 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.

16. Heart Outcomes Prevention Evaluation Study Investigators. Vitamin E supplementation and cardiovascular events in high-risk patients. N Engl J Med 2000;342:154-60.

17. CURRENT–OASIS 7 Investigators. Dose comparisons of clopidogrel and aspirin in acute coronary syndromes. N Engl J Med 2010;363:930-42.

18. Stone GW, Witzenbichler B, Guagliumi G, et al.; HORIZONS-AMI Trial Investigators. Bivalirudin during primary PCI in acute myocardial infarction. N Engl J Med 2008;358:2218-30.

19. Stone GW, Lansky AJ, Pocock SJ, et al.; HORIZONS-AMI Trial Investigators. Paclitaxel-eluting stents versus bare-metal stents in acute myocardial infarction. N Engl J Med 2009;360:1946-59.

20. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. Lancet 1995;345:669-85.

21. Valgimigli M, Frigoli E, Leonardi S, et al.; MATRIX Investigators. Bivalirudin or unfractionated heparin in acute coronary syndromes. N Engl J Med 2015;373:997-1009.

22. Valgimigli M, Gagnor A, Calabró P, et al. MATRIX Investigators. Radial versus femoral access in patients with acute coronary syndromes undergoing invasive management: a randomised multicentre trial. Lancet;385:2465-76.

23. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003;290:1624-32.

24. Pocock SJ, Gersh BJ. Do current clinical trials meet society's needs? : a critical review of recent evidence. J Am Coll Cardiol 2014;64:1615-28.

25. New York University School of Medicine. International Study of Comparative Health Effectiveness With Medical and Invasive Approaches (ISCHEMIA). 2015. Available at: https://clinicaltrials.gov/ct2/show/NCT01471522. Accessed November 2, 2015.

26. Stone GW, Hochman JS, Williams DO, et al. Medical therapy with versus without routine revascularization in patients with stable ischemic heart disease and moderate or severe ischemia: the case for community equipoise. J Am Coll Cardiol 2015; In press.

27. Maron DJ, Stone GW, Berman DS, et al. Is cardiac catheterization necessary before initial management of patients with stable ischemic heart disease? Results from a Web-based survey of cardiologists. Am Heart J 2011;162:1034-1043.e13.

28. Smith CR, Leon MB, Mack MJ, et al.; PARTNER Trial Investigators. Transcatheter versus surgical aortic-valve replacement in high-risk patients. N Engl J Med 2011;364:2187-98.

29. Adams DH, Popma JJ, Reardon MJ, et al.; U.S. CoreValve Clinical Investigators. Transcatheter aortic-valve replacement with a self-expanding prosthesis. N Engl J Med 2014;370:1790-8.

30. Boden WE, O'Rourke RA, Teo KK, et al.; COURAGE Trial Research Group. Optimal medical therapy with or without PCI for stable coronary disease. N Engl J Med 2007;356:1503-16.

31. Serruys PW, Morice MC, Kappetein AP, et al.; SYNTAX Investigators. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease [Erratum appears in N Engl J Med 2013;368:584]. N Engl J Med 2009;360:961-72.

32. Palmerini T, Benedetto U, Biondi-Zoccai G, et al. Long-term safety of drug-eluting and bare-metal stents: evidence from a comprehensive network meta-analysis. J Am Coll Cardiol 2015;65:2496-507.

33. Windecker S, Stortecky S, Stefanini GG, et al. Revascularisation versus medical treatment in patients with stable coronary artery disease: network meta-analysis. BMJ 2014;348:g3859.

34. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? N Engl J Med 2000;342:1907-9.

35. Brown ML, Gersh BJ, Holmes DR, et al. From randomized trials to registry studies: translating data into clinical information. Nat Clin Pract Cardiovasc Med 2008;5:613-20.

36. Lagerqvist B, Fröbert O, Olivecrona GK, et al. Outcomes 1 year after thrombus aspiration for myocardial infarction. N Engl J Med 2014;371:1111-20.

37. Ellenberg SS, Fleming TR, DeMets DL. Data Monitoring Committees in Clinical Trials: A Practical Perspective. Chichester, England: John Wiley & Sons, 2003.

38. Zannad F, Gattis Stough W, McMurray JJV, et al. When to stop a clinical trial early for benefit: lessons learned and future approaches. Circ Heart Fail 2012;5:294-302.

39. DAMOCLES Study Group. A proposed charter for clinical trial data monitoring committees: helping them to do their job well. Lancet 2005;365:711-22.

40. Pocock SJ. Current controversies in data monitoring for clinical trials. Clin Trials 2006;3:513-21.

41. McMurray JJV, Packer M, Desai AS, et al.; PARADIGM-HF Investigators and Committees. Angiotensin-neprilysin inhibition versus enalapril in heart failure. N Engl J Med 2014;371:993-1004.

42. Swedberg K, Komajda M, Böhm M, et al.; SHIFT Investigators. Ivabradine and outcomes in chronic heart failure (SHIFT): a randomised placebo-controlled study. Lancet 2010;376:875-85.

43. Fox K, Ford I, Steg PG, et al.; BEAUTIFUL Investigators. 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.

44. Montori VM, Devereaux PJ, Adhikari NJ, et al. Randomized trials stopped early for benefit: a systematic review. JAMA 2005;294:2203-9.

45. Lachin JM. Futility interim monitoring with control of type I and II error probabilities using the interim Z-value or confidence limit. Clin Trials 2009;6:565-73.

46. Bousser M-G, Amarenco P, Chamorro A, et al.; PERFORM Study Investigators. Terutroban versus aspirin in patients with cerebral ischaemic events (PERFORM): a randomised, double-blind, parallel-group trial. Lancet 2011;377:2013-22.

47. Swedberg K, Young JB, Anand IS, et al.; RED-HF Investigators. Treatment of anemia with darbepoetin alfa in systolic heart failure. N Engl J Med 2013;368:1210-9.

48. DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. Lancet 1999;354:1983-8.

49. Barter PJ, Caulfield M, Eriksson M, et al.; ILLUMINATE Investigators. Effects of torcetrapib in patients at high risk for coronary events. N Engl J Med 2007;357:2109-22.

50. Connolly SJ, Camm AJ, Halperin JL, et al.; PALLAS Investigators. Dronedarone in high-risk permanent atrial fibrillation. N Engl J Med 2011;365:2268-76.

51. Hohnloser SH, Crijns HJGM, van Eickels M, et al.;ATHENA Investigators. Effect of dronedarone on cardiovascular events in atrial fibrillation. N Engl J Med 2009;360:668-78.

52. Alexander JH, Lopes RD, James S, et al.; APPRAISE-2 Investigators. Apixaban with antiplatelet therapy after acute coronary syndrome. N Engl J Med 2011;365:699-708.

53. Gallo P, Chuang-Stein C, Dragalin V, et al.; PhRMA Working Group. Adaptive Designs in Clinical Drug Development—An Executive Summary of the PhRMA Working Group. J Biopharm Stat 2006;16:275-83; discussion 285-91.

54. Chow SC, Chang M. Adaptive design methods in clinical trials-a review. Orphanet J Rare Dis 2008;3:11.

55. Committee for Medicinal Products for Human Use (CHMP). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency. 2007. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf. Accessed November 2, 2015.

56. Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics: Draft Guidance. 2010. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf. Accessed November 2, 2015.

57. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. Stat Med 2011;30:3267-84.

58. Kunz CU, Friede T, Parsons N, et al. Data-driven treatment selection for seamless phase II/III trials incorporating early-outcome data. Pharm Stat 2014;13:238-46.

59. Simon N, Simon R. Adaptive enrichment designs for clinical trials. Biostatistics 2013;14:613-25.

60. Rosenberger WF. Randomized play-the-winner clinical trials: review and recommendations. Control Clin Trials 1999;20:328-42.

61. Harrington RA, Stone GW, McNulty S, et al. Platelet inhibition with cangrelor in patients undergoing PCI. N Engl J Med 2009;361:2318-29.

62. Mehta C, Gao P, Bhatt DL, et al. Optimizing trial design: sequential, adaptive, and enrichment strategies. Circulation 2009;119:597-605.

63. Bhatt DL, Stone GW, Mahaffey KW, et al.; CHAMPION PHOENIX Investigators. Effect of platelet inhibition with cangrelor during PCI on ischemic events. N Engl J Med 2013;368:1303-13.

FIGURE LEGENDS

Central Illustration. Key Challenges in Trial Design

Figure 1: Possible Outcomes in a Noninferiority Trial (observed difference and 95% CI)

Caption: Conceptual figure for interpreting non-inferiority trials based on the estimated absolute difference, 95% CI and a non-inferiority margin of delta,. The vertical line at 0 represents no treatment difference. CIs to the left of the delta line indicate non-inferiority of the new treatment. Note delta can sometimes be specified on a relative risk scale.

Figure 2: Conceptual Outline of Sample Size Re-estimation When in the Promising Zone

Caption: Possible outcomes from an interim analysis for adaptive sample size re-estimation. The three scenarios are: (1) the Favorable Zone when the observed treatment difference is similar to the pre-planned treatment difference, (2) the Promising Zone when the observed difference is less than that hoped for but where reasonable conditional power can be achieved by increasing the sample size, and (3) the Unfavorable Zone when interim results show poor conditional power and the trial continues to its original size.

Table 1: Displaying the Results of a Factorial Design: CURRENT OASIS 7

First main effect	**Double-dose clopidogrel** **(N = 12,520)**	**Standard-dose clopidogrel** **(N = 12,566)**
Primary event rates	4.17%	4.43%
	HR: 0.94; 95% CI: 0.83 to 1.06; p = 0.30	
Second main effect	**Higher-dose aspirin** **(N = 12,507)**	**Lower-dose aspirin** **(N = 12,579)**
Primary event rates	4.24%	4.36%
	HR: 0.97; 95% CI: 0.86 to 1.09; p = 0.61	
Potential interaction	**Primary events by both treatments simultaneously**	
	Double-dose clopidogrel	Standard-dose clopidogrel
Higher-dose aspirin	3.8%	4.6%
Lower-dose aspirin	4.5%	4.2%
Interaction test	p = 0.04	

CI = confidence interval; HR = hazard ratio.

Table 2: Planned Stopping Boundaries for an Event-Driven, 13,000-Patient,Placebo-Controlled Trial of Patients at High CV Risk

Interim Analysis	Number of Primary Events	Stopping Boundaries*	
		Superiority	**Futility**
1	800 (50%)	p <0.0002	p >0.758
		HR <0.768	HR >0.979
2	1200 (75%)	p <0.0002	p >0.216
		HR <0.806	HR >0.931
Final	1,600 (100%)	p <0.05	
		HR <0.906	

*There are no formal stopping boundaries for safety.

CV = cardiovascular; HR = hazard ratio.