Design of Major Randomized Trials:
Part 3 of a 4-part Series on Statistics for Clinical Trials

Brief title: Clinical Trial Design

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

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, UK
Tel: +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

This paper provides practical guidance on the fundamentals of design for major randomized controlled trials. Topics covered include the choice of patients, choice of treatment and control groups, choice of primary and secondary endpoints, methods of randomization, appropriate use of blinding, and determination of trial size. Insights are made with reference to contemporary major trials in cardiology.

Abbreviations
ACS = acute coronary syndrome
CV = cardiovascular
MACE = major adverse cardiovascular event
MI = myocardial infarction
NSTEMI = non-ST-segment elevation myocardial infarction
PCI = percutaneous coronary intervention
STEMI = ST-segment elevation myocardial infarction

Introduction

The properly designed randomized controlled trial is recognized as providing the highest level of evidence in determining guidelines for therapeutic practice (1). However, numerous pitfalls in trial design may occur, which, if succumbed to, may seriously impair the reliability of or confidence in the results. In this paper, we review the key features that determine whether a trial's design is "up to scratch." These fundamentals include: selection of patients, treatments and outcomes; randomization methods; appropriate blinding; and determination of trial size. Of course, one first has to propose a relevant clinical question for which true equipoise exists.

Our focus is on major, pivotal (phase III) trials that are intended to directly influence clinical practice, and throughout we bring the issues to life with topical examples from recent cardiology trials. Our underlying intent is to elucidate matters of trial design that should be considered and clearly documented in a study protocol (2,3). Once the study protocol is finalized, it is good practice to register the trial on a website (prior to patient enrollment) and to make the protocol readily available. The core goals are to plan each trial so that it will achieve unbiased conclusions for treatment comparisons of direct public health relevance that are both clinically meaningful and statistically precise. It is fundamentally important to differentiate a statistically significant difference from a clinically meaningful difference. Very large studies may sometimes result in small, statistically significant differences in outcomes that have little or no clinical relevance. It is, of course, also essential that the trial can be successfully completed. That is, from ethical, scientific, organizational, and funding standpoints, the protocol is deliverable. This pragmatism underpins all that follows here.

Choice of Patients and Centers

The delineation of precise eligibility criteria is important in determining the population of patients to which the trial findings can be extrapolated (3). Setting the criteria too specifically can inhibit successful patient recruitment and restrict generalizability, whereas unduly broad entry criteria may dilute the opportunity to identify a treatment effect in a specific population.

In acute coronary syndromes (ACS), the PLATO trial of ticagrelor versus clopidogrel studied a broad population: patients with ACS with or without ST-segment elevation and treated with or without invasive procedures (4). In contrast, the HORIZONS-AMI trial compared anticoagulation regimens only in patients undergoing primary percutaneous coronary intervention (PCI) in acute myocardial infarction (MI) (5). This was appropriate, as the ACUITY trial had already studied bivalirudin in ACS patients with high-risk unstable angina or non-ST-segment elevation MI (NSTEMI) (6). As the prognosis and treatment pathways for patients with ACS with and without ST-segment elevation are quite different, it is advisable to avoid pooling these 2 groups in a single trial design, unless each cohort is separately powered for safety and efficacy.

In chronic heart failure, the CHARM trial of candesartan had broad entry criteria, with no entry ejection fraction restriction, although recognizing that patients with preserved left ventricular systolic function should be interpreted as a separate stratum, CHARM PRESERVED (7,8). In contrast, the RALES trial of spironolactone focused on high-risk patients (history of New York Heart Association class IV and ejection fraction ≤35%). This facilitated using all-cause mortality as a realistic primary endpoint, but restricted the generalizability of the trial findings (9). Alternatively, some trials recruit only high-risk patients in order to achieve a sufficient number of outcome events; for example, SAVOR and other cardiovascular (CV) safety trials in diabetes (10).

Nearly all major trials in cardiology are multicenter, and many have a global recruitment strategy. The occasional exception happens (e.g., the single-center TAPAS and HEAT-PPCI trials in ST-segment elevation MI (STEMI) patients)(11,12), but then their generalizability is uncertain. The results of single-center trials may be unique to the specific practices and techniques employed at that institution; adequate masking and absence of independent monitoring and core laboratories may be issues; resource commitment is often less than in multicenter trials; and treatment effect sizes can sometimes be at odds with multicenter investigations.

However, global trials raise other issues: problems of ensuring trial quality in all countries; issues with center to center variability and training; easier recruitment in certain regions (e.g., Eastern Europe), which affects the distribution of geographic representation; and concerns over geographic heterogeneity in treatment effect, as occurred in the TOPCAT trial of spironolactone in preserved cardiac function heart failure (**Figure 1**) (13). Nonetheless, large-scale multicenter trials are preferable to single-center studies, with the understanding that the results apply to the geographies and practices of the participants. Secondary analysis should be undertaken to ensure there are no major geographic disparities in the results (14).

Choice of Treatments and Control Groups

In drug trials, usually the new treatment has 1 specific dose regimen selected on the basis of earlier research investigating the agent's safety profile, and often using surrogate endpoints to evaluate efficacy. Occasionally, a major trial proceeds with 2 different doses of the new drug; for example, RE-LY in atrial fibrillation and PEGASUS in the post-MI setting (15,16). If the increase in trial size necessitated by 2 experimental drug regimens is achievable (with appropriate adjustments to preserve alpha), examining different drug doses (or, occasionally,

even different experimental treatments) is, in principle, a valuable design option, although controversies can ensue in interpretation of findings.

The appropriate choice of control group is crucial in trial design (17). Many opt for a placebo control, whereby the new drug or placebo is assigned on top of current therapeutic practice; for example, CHARM and SHIFT (assessing ivabradine) in chronic heart failure (7,18). This is common for regulatory trials; evidence for FDA approval can be straightforward, but in an advancing field, such as heart failure, it does mean that patients face a growing polypharmacy, with little evidence that any drugs get withdrawn from routine use (19).

Opting for an active comparator poses additional challenges. For example, in the PARADIGM-HF trial in heart failure, comparing LCZ696 with enalapril (20), the consequent superiority of LCZ696 is liable to have a far greater impact than if a placebo control had been chosen. Others will choose an active comparator to demonstrate noninferiority in efficacy, anticipating other benefits (e.g., safety). An example is the ACUITY trial in ACS, which compared bivalirudin with heparin plus a glycoprotein IIb/IIIa inhibitor (noninferior rates of composite adverse ischemic events with reduced major bleeding) (6). A placebo control would be unethical in many such scenarios. We review noninferiority trials in next week's article.

Another option is to have a third combination treatment arm, new drug + active comparator, as was done in the VALIANT trial of valsartan versus enalapril versus both in MI with complications (21). In this instance, the combination demonstrated an excess of adverse events without improving survival, a warning that combination arms need careful preparatory study to see if their safety profiles are acceptable.

Another consideration for chronic disease trials is whether to have a run-in period prior to randomization on either the new drug, the comparator, or both in sequence (as in PARADIGM-

HF) (20). In principle, this helps to confine enrollment to patients who can tolerate both treatments, resulting in greater adherence, but care is needed to ensure that a fair and relevant treatment comparison is not compromised. Trials with a run-in period tend to result in better drug adherence and tolerability than would otherwise occur. For instance, if a trial of ticagrelor had a run-in period, patients experiencing dyspnea in the short-term would likely be excluded, thereby under-reporting dyspnea in the consequent randomized comparison.

Trials comparing radically different strategies, such as PCI versus coronary artery bypass graft, face unique challenges, which are discussed in the next article in this series.

Choice of Primary and Secondary Outcomes

Most major trials in cardiology focus attention on disease event outcomes. Treatment effects on potential surrogate endpoints (22) and biomarkers in earlier-phase trials can be a useful motivation and help determine drug dose, but pivotal trials typically require clear evidence of patient benefit using clinical endpoints if they are to influence treatment practice. Nonetheless, many drugs (and some devices) are approved on the basis of surrogate endpoints (e.g., blood pressure or glucose control). Given the cost and complexity of pivotal randomized trials, there is increasing interest in considering surrogates as pivotal trial endpoints, but the criteria for true surrogacy are difficult, though not impossible, to achieve (23). For example, angiographic late loss after coronary stent implantations has been demonstrated to be a surrogate for ischemia-driven target lesion revascularization (24).

Although selection of which clinical events to collect follow-up data on may be straightforward, precise definitions are needed in the study protocol. Furthermore, many trials have an independent Events Adjudication Committee (25) blinded to treatment assignment, a

potentially important quality control measure to ensure a greater consistency of event arbitration than can be achieved by relying on individual investigators.

An important challenge is to define the primary endpoint for the trial (26). In trials of high-risk patients, such as the PARTNER trials of transaortic valve replacement (27), all-cause death is the primary outcome. However, for most major trials, a composite primary endpoint (26) comprising several nonfatal events along with mortality is chosen: no one type of event adequately captures the treatment effect, and such a composite has the potential to enhance statistical power by the greater frequency of patients experiencing at least 1 component of the composite during follow-up.

But what events should contribute to a composite primary endpoint? This depends on the disease and treatments being studied. In ACS and diabetes CV safety trials, the usual composite is CV death, MI, and stroke. Some are tempted to add in extra components; for example, including unstable angina or ischemia-driven revascularization into a broader major adverse CV events (MACE) composite. This boosts the numbers of events, but dilutes the impact and meaning of the composite. For instance, the most frequent (and often least clinically relevant) component tends to be the driver of event rates (e.g., enzymatic MIs or revascularization in ACS trials).

In chronic heart failure, the standard composite primary endpoint is CV death and hospitalization for heart failure. The weakness here is that it emphasizes the first hospitalization and ignores any subsequent hospitalizations or death thereafter. Greater insight and enhanced statistical power may be achieved by a primary endpoint analysis that incorporates both repeat hospitalizations for heart failure and CV death (28), as in the ongoing PARAGON trial of LCZ696 versus valsartan in preserved ejection fraction heart failure (29).

Several options exist for the analysis of repeat events (e.g., hospitalizations) in chronic disease (e.g., heart failure) trials and are illustrated in reanalyses of the CHARM and CORONA trials (28,30). It is important to recognize that repeat events within a patient are not independent. For example, a few patients will have many hospitalizations, whereas many will have none at all. The negative binomial approach allows for this, while producing event rates per treatment and hence a rate ratio (and its confidence interval). One complexity is the competing risk of mortality: patients often have high event rates prior to death, but their follow-up is then informatively censored. Joint frailty models can simultaneously estimate both hospitalization and mortality risks: the challenge is how to make such statistical complexity comprehensible to readers of trial reports with repeat events data.

Novel methods of handling composite endpoints are of value when the components vary in their clinical severity and importance. The least serious component may tend to occur earlier (and more frequently), and thus is afforded undue priority in a conventional analysis. These novel methods; for example, the Finkelstein-Schoenfeld method or the win ratio, invert the priorities to match with clinical severity (31-33). The principle behind the win ratio method is that one compares every possible pair of patients on the new treatment and the standard treatment. First, the "new" patient is declared the winner or loser if they did better or worse on the most serious outcome (e.g., who lived longer). If neither died, then you decide the winner or loser on the basis of who had the next most serious outcome (e.g., hospitalization) first and so on, if there are more (less-serious) events to help determine who did worse. Then, across all pairs, one notes the number of winners and number of losers for the new treatment (ignoring any "tied" pairs who had no relevant events): the win ratio is the former number divided by the latter.

This concept can be extended to include nonevent outcomes (e.g., ejection fraction), as a "tie-breaker" in patients who are otherwise event-free.(34)

Another challenge is how to incorporate predefined safety concerns; for example, major bleeding in antiplatelet and anticoagulant drug trials, into the overall outcome priorities. Some trials (e.g., ACUITY and OPTIMIZE (6,35)), have extended MACE to net adverse clinical events by adding major bleeding as an extra component of the primary composite outcome (26). If the included safety and efficacy components are weighted similarly in terms of their clinical importance, such an endpoint may reflect the overall net clinical benefit to the patient of the therapies being evaluated. However, if adequate power is present, it may be more desirable to evaluate efficacy and safety separately, as coprimary endpoints. For example, major bleeding was predefined as the primary safety outcome in the HORIZONS-AMI trial (5).

In general, efforts to collect quality data on potential safety concerns are important. Although MedDRA coding of adverse events and serious adverse events is well established (36), the process has its limitations. Thus, if prior insight into which safety concerns may plausibly arise is available, specific extra reporting procedures may be undertaken for such events.

Having defined the primary endpoint, it is important to then provide a list of secondary endpoints in the protocol, along with any predefined priorities among them. Specifically, the separate components of the primary endpoint should be included here. For regulatory trials, interest focuses on how to interpret secondary endpoint data, should the primary be "positive"; that is, can one extend the label to include any positive secondary endpoints? Across a multiplicity of outcomes there are 2 options in this regard. First, one can list a set of secondary endpoints in order of priority, so that one can proceed down that list when results are known, counting all in sequence as positive until one fails to reach statistical significance (hierarchical

testing). Alternatively, one can pre-declare, on equal terms, a specific set of secondary outcomes and use a statistical correction for multiplicity (e.g., the Hochberg procedure (37)), in determining which endpoints meet criteria as "positive." If a method to control for multiplicity was not pre-specified, the data with confidence intervals should just be presented, and clearly identified as secondary outcomes. Although the implications from such secondary endpoints should usually be considered exploratory and hypothesis generating, they still may provide a persuasive level of evidence if highly positive (e.g., p <0.001).

Because patient-centered outcomes and economic evaluations are of increasing importance, secondary outcomes may also include quality of life, patient preferences, and resource utilization data.

Methods of Randomization

The justification for randomization, and the problems in interpreting nonrandomized treatment comparisons are well known (38,39). Herein, we discuss the actual methods of setting up and delivering random allocation of patients to treatments (40).There are 3 key elements: 1) the statistical methods used to set up the sequence of randomized treatment assignments; 2) the practical means by which the investigator assigns each randomized treatment; and 3) ensuring that patient eligibility, investigator agreements, and informed patient consent are all in place before randomization. Informed consent is a vital component of ethical trial conduct, but its proper delivery lies beyond the scope of this more statistical article. We do note, however, that there are special challenges to obtaining consent in an emergency setting; for example, primary PCI in STEMI patients.

A central tenet is that the investigator cannot predict the assignment in advance because that could bias the choice of the next patient to be randomized. The CONSORT guidelines state

that a trial publication should document these randomization methods (41). For example, the MATRIX trial stated: "before start of angiography, patients were centrally assigned (1:1) to radial or femoral access... using a web-based system to ensure adequate concealment of allocation. The randomized sequence was computer generated, blocked and stratified by... use of ticagrelor or prasugrel, type of ACS (STEMI or NSTEMI) and anticipated use of immediate PCI" (42). Thus, MATRIX used one of the most common approaches: a web-based allocation system and random permuted blocks within strata, with $2 \times 2 \times 2 = 8$ strata in this instance. What is not stated here is the number of patients in each randomized block: commonly this is set at 4 (two per treatment) or 6 (three per treatment) in a random order. Because MATRIX is an "open" study, blocks of varying size would be recommended to avoid prediction of some assignments in advance, whereas this consideration is less important for blinded trials.

For the CURRENT-OASIS 7 trial, a 24-h computerized, central automated voice-response system, with permuted block randomization stratified by center, was used (43). Note the need for continuous day and night randomization in an acute setting.

These 2 examples, 1 stratifying by 3 patient factors (but not center) and the other stratifying by center only, reflect the diversity of approaches to stratifying randomization. Indeed, some trials (e.g., CHARM in chronic heart failure (7)) have just a simple randomization list with permuted blocks. So why stratify at all? Stratification offers insurance that for a key patient feature, a serious imbalance between treatment groups will not arise by chance. However, for large, pivotal trials, this becomes very unlikely and, if it does occur, it may be accounted for in a covariate-adjusted analysis. In addition, trialists often have difficulty agreeing on the choice of stratification variables, and the gain in statistical efficiency by stratification is negligible.

However, showing near-perfect balance in treatment groups for key baseline variables is a convenient bonus of stratification.

An alternative method of achieving balance across 2 or more baseline features is called minimization (44). Minimization is a dynamic approach (randomization lists are not prepared in advance) whereby, as each patient enters the trial, a computerized algorithm determines which treatment assignment would achieve the better overall balance across multiple patient characteristics. That treatment is then assigned, or, to preserve a chance element, it is assigned with a high probability (e.g., 0.75). For example, the TITRe2 trial of liberal versus restrictive transfusion after cardiac surgery "used cohort minimization to balance assignments according to center and type of surgery" (45). Whatever method is chosen, more important than stratification is ensuring that randomization is carried out in a rigorous manner.

Sometimes an unequal randomization is used, for example, with a 2:1 or 3:2 ratio in favor of the new treatment (46). This has 2 possible advantages: it increases information on the new treatment (e.g., to allay safety concerns) and may also stimulate greater enthusiasm by investigators and patients knowing they have more than a 50-50 chance of getting the new treatment. For instance, the SYMPLICITY-HTN 3 trial of renal denervation versus a sham procedure in resistant hypertension had a 2:1 randomization ratio, whereas the HORIZONS-AMI trial in primary PCI had a 3:1 ratio of drug-eluting versus bare-metal stents (5,47). But there is a loss of statistical power with unequal randomization. For example, for 3:2, 2:1, and 3:1 ratios, the total number of patients needs to increase by 4.2%, 12.5%, and 33.3%, respectively, to preserve the same power as a trial with equal randomization.

Blinding

It is standard practice for drug trials to be double-blind: that is, the patients and those responsible for their treatment and follow-up evaluation do not know which randomized treatment they are assigned (48,49). This is the optimal way to ensure that any potential influence that awareness of treatment might have on patient perceptions, patient management, and evaluation of endpoints is avoided.

So what are we to make of trials that are not double-blinded? For instance, the 3-arm RE-LY trial in atrial fibrillation compared 2 fixed doses of dabigatran in a blinded manner with open-label use of warfarin (15). Does this bias the comparison with warfarin? Having all primary and secondary outcome events adjudicated by an independent group who were centrally blinded helps reduce bias, but we are still reliant on unblinded patients and investigators to fairly report relevant events. Other similar trials; for example, the ARISTOTLE trial of apixaban vs. warfarin, were double-blinded, with the consequent complexity of carrying out repeat international normalized ratio testing on all patients, including those not on warfarin (50). Which of these approaches is better is open to debate.

Of course, many trials of devices and alternative treatment strategies cannot possibly be double-blinded. For instance, the MATRIX trial of radial versus femoral access for ACS patients undergoing invasive management inevitably had patients and treating physicians aware of their randomized assignment, but outcome assessors were masked (blinded)(42). Such a PROBE (Prospective Randomized Open Blinded Endpoint) design is commonly used (51), but many researchers believe that the data obtained is less reliable than in a double-blind study.

Blinding becomes particularly important when the primary endpoint is not on the basis of "hard" clinical events (with objective definitions), but on softer endpoints, such as quality of life indicators, or a measurement such as blood pressure. The SYMPLICITY HTN-3 trial in resistant

hypertension is an important illustration of the value of comparing a device (renal denervation) with a sham procedure. Both the patients and the blood-pressure assessors were unaware of the study-group assignments (47). In prior unblinded and uncontrolled studies, marked blood pressure lowering was attributed to renal denervation. In contrast, in the sham-controlled randomized trial, comparable reductions in blood pressure were noted in the treatment and control groups, suggesting a substantial placebo effect with some regression to the mean (52). The neutral finding of this trial, the first objective, unbiased evaluation of renal denervation, has importantly affected future trial design considerations for medical devices in general.

Trial Size

A key role of statistical reasoning in trial design is the provision of power calculations to determine the required trial size (53). Focusing on the trial's primary endpoint, one needs to specify the expected event rate in the control group, the magnitude of treatment effect that the trial is required to detect (the alternative hypothesis), and the degree of certainty that such detection should occur (the statistical power: 1 minus the risk of a type II error [β] and the significance level that qualifies as "detected" [the risk of a type I error (α)]).

For instance, in the CHAMPION-PHOENIX trial in patients undergoing PCI, the composite primary endpoint was death, MI, ischemia-driven revascularization, and stent thrombosis within 48 h. The investigators planned for a difference of 5.1% in patients treated with clopidogrel versus 3.9% in patients treated with cangrelor (a 24.5% relative risk reduction) to be detected with 85% power and type I error of 0.05. This required a total of 10,900 randomized patients (5,450 per group). They actually recruited 11,145 patients and the observed difference was in fact 5.9% versus 4.7%, p $=$ 0.005 (54).

Note the large sample size that was required. Had the trial planners hypothesized an unrealistic 50% relative risk reduction with 80% power, the required trial size would have been reduced to 1,780 patients. With observed rates of 5.9% versus 4.7% the p value would have been 0.25, well short of statistical significance. Fortunately, such overoptimistic planning is becoming less common nowadays in major CV trials, although this requires commitment to relatively large trials, with modestly realistic, but still clinically important, expectations of treatment benefit.

For chronic disease trials, sample size determination involves considering length of follow-up as well as patient numbers. For instance, the PARADIGM trial in chronic heart failure was planned to recruit 8,000 patients with a mean 34 months follow-up. For the composite primary endpoint of heart failure hospitalization and CV death, they anticipated an annual event rate of 14.5% on enalapril, and had 97% power to detect a 15% relative risk reduction on LCZ696, with type I error of 0.05. Why set the power so high? They also wanted good statistical power (80%) to also detect the same relative risk reduction in the secondary endpoint of CV death alone. A wise choice, in fact, because the published trial results showed marked treatment benefits for both outcomes (20).

A cynical view of statistical power calculations is that people "juggle with the numbers" until they get the sample size they always wanted (or could afford). There is an element of truth to this: there are practical limits on what is achievable patient recruitment. Hence, some compromise is needed in not making the target so difficult as to put the consequent number of patients out of range for a real world study.

So how can we "play with the numbers" while preserving scientific integrity? Take CHAMPION-PHOENIX as an example, and consider varying the parameters that feed into the calculations:

1. **Control group event-rate.** If one were to double the anticipated 5.1% rate to 10.2%, while keeping the same relative risk reduction and power, one could halve the required number of patients. But this would be an artificial deception, unless recruitment was restricted to very high-risk patients, an unrealistic option in this case.

2. **Alternative hypothesis.** The required number of patients is inversely proportional to the square of the declared absolute treatment difference. So if one increased this difference from 5.1% (control) − 3.9% (treatment) = 1.2% to 5.1% − 2.7% = 2.4%, one would only need less than a quarter of the patients, that is, 2,400 rather than 10,900. In contrast, halving the difference to a mere 0.6% takes the sample size to a staggering 46,300 patients. The declared 1.2% difference was a realistic compromise.

3. **Power.** There is nothing magical about the chosen power. The greater the power, the larger the required patient numbers (but the smaller the likelihood of failing to detect a difference that is truly present). A common choice is 90% power: it carries a reasonable sense of guarantee that one would be unlucky (10% chance) to miss out on detecting the specified treatment difference, if it were true. If one reduces the power to 85% or 80%, then trial size can be reduced by 14% and 25%, respectively. Reducing power down to 50% reduces sample size by 63% compared with 90% power, but only a gambler would take such a risk. Few trials are performed with <80% power.

4. **Type I error.** One could choose options other than the conventional 0.05. For instance, changing to 0.1 or 0.01 would make trial size around 19% smaller and 42% bigger, respectively, but the former falls short of conventional significance and the latter is not commonly chosen.

An important fact when determining trial size is to recognize that statistical power depends mainly on the total number of patients in the trial experiencing the primary event during follow-up. Thus, some trials are event-driven, meaning that all patients are followed until that calendar date when the required number of events has occurred. Applying this logic, how many events should a trial aim for? A useful approximate formula is the following: one declares the hazard ratio (or risk ratio) for the alternative hypothesis and its associated statistical power $(1- \beta)$ (often 90% or 80%) and the 2-sided type I error (α) (usually 0.05). Then, the total events required:

$$D = \frac{(1+R)^2}{(1-R)^2} \times f(\alpha, \beta)$$

where $f(\alpha, \beta) = \left(z_\beta + z_{\alpha/2}\right)^2$, which equates to 10.51 and 7.85 for 90% and 80% power, respectively, with $\alpha = 0.05$. z is a standardized normal deviate.

Thus, if one wishes to detect a hazard ratio of 0.7 (i.e., a 30% relative risk reduction) with 80% power, then 252 total primary events are required. A much tougher challenge is to detect a hazard ratio of 0.85 (i.e., a 15% relative risk reduction) with 90% power because this requires a total of 1,194 primary events.

Figure 2 displays the power curves to assist the reader in determining their own trial size (required number of events) for any choice of hazard ratio, with options for 90% and 80% power, as desired.

It has been argued that all power calculations enforce an arbitrary logic by focusing on a particular alternative hypothesis. In practice, the trial statistician could produce extensive tabulations by varying all 4 of the above parameters, thus guiding the investigators to a sensible trial size. Overall, a concise logic, specified in both the trial protocol and the publications' Methods section, helps to instill confidence that the trialists chose wisely (and achieved) their

target sample size. Of course, "the bigger the better" is a crude maxim whereby the larger the

trial becomes, the greater the precision of any treatment effect estimate (and the more likely are

the results to approximate "the truth"). This scientific optimization always needs to be placed in

perspective alongside the need to keep the trial's scale, duration, and costs finite.

Concluding Remarks

A sound knowledge of the essential issues in the design of major randomized trials, as

outlined in this paper, is of importance to cardiologists and others involved in trials research and

their interpretation. Although there are many details of relevance, the essential elements of

designing a worthwhile investigation include:

1. Identifying a valuable therapeutic concept worthy of a major trial;

2. Focusing on the essentials of the trial protocol: defining exactly which patients, which
   treatment comparison(s) and which primary (and secondary) endpoints should be studied;

3. Ensuring the trial provides a reliable (unbiased) treatment comparison by appropriate
   randomization, blinding and quality delivery of the protocol's intent;

4. Making the trial large enough so that it is adequately powered to detect (or refute) any
   treatment differences of clinical importance.

REFERENCES

1.   Jacobs AK, Anderson JL, Halperin JL. The evolution and future of ACC/AHA clinical practice guidelines: a 30-year journey: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2014;64:1373-84.

2.   Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Int Med 2013;158:200-7.

3.   Chan AW, Tetzlaff JM, Gøzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ 2013;346:e7586.

4.   Wallentin L, Becker RC, Budaj A, et al. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2009;361:1045-57.

5.   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.

6.   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.

7.   Pfeffer MA, Swedberg K, Granger CB, et al.; CHARM Investigators and Committees. Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM-Overall programme [Erratum appears in Lancet 2009;374:1744]. Lancet 2003;362:759-66.

8.   Yusuf S, Pfeffer MA, Swedberg K, et al.; CHARM Investigators and Committees. Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial. Lancet 2003;362:777-81.

9.   Pitt B, Zannad F, Remme WJ, et al.; Randomized Aldactone Evaluation Study

       Investigators. The effect of spironolactone on morbidity and mortality in patients with

       severe heart failure. N Engl J Med 1999;341:709-17.

10.   Scirica BM, Bhatt DL, Braunwald E, et al.; SAVOR-TIMI 53 Steering Committee and

       Investigators. Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes

       mellitus. N Engl J Med 2013;369:1317-26.

11.   Svilaas T, Vlaar PJ, van der Horst IC, et al. Thrombus aspiration during primary

       percutaneous coronary intervention. N Engl J Med 2008;358:557-67.

12.   Shahzad A, Kemp I, Mars C, et al.; HEAT-PPCI trial investigators. Unfractionated

       heparin versus bivalirudin in primary percutaneous coronary intervention (HEAT-PPCI):

       an open-label, single centre, randomised controlled trial [Erratum appears in Lancet

       2014;384:1848]. Lancet 2014;384:1849-58.

13.   Pfeffer MA, Claggett BP, 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.

14.   Pocock S, Calvo G, Marrugat J, et al. International differences in treatment effect: do

       they really exist and why? Eur Heart J 2013;34:1846-52.

15.   Connolly SJ, Ezekowitz MD, Yusuf S, et al.; RE-LY Steering Committee and

       Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med

       2009;361:1139-51.

16.   Bonaca MP, Bhatt DL, Cohen M, et al.; PEGASUS-TIMI 54 Steering Committee And

       Investigators. Long-term use of ticagrelor in patients with prior myocardial infarction. N

       Engl J Med 2015;372:1791-800.

17.	ICH Harmonized Tripartite Guideline: choice of control group and related issues in clinical trials E10. International Conference On Harmonisation Of Technical Requirements For Registration Of Pharmaceuticals For Human Use 2000:1-35.

18.	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.

19.	Rossello X, Pocock SJ, Julian DG. Long-term use of cardiovascular drugs: challenges for research and for patient care. J Am Coll Cardiol 2015;66:1273-85.

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

21.	Pfeffer MA, McMurray JJV, Velazquez EJ, et al.; Valsartan in Acute Myocardial Infarction Trial Investigators. Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both [Erratum appears in N Engl J Med 2004;350:203]. N Engl J Med 2003;349:1893-906.

22.	Psaty BM, Weiss NS, Furberg CD, et al. Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. JAMA 1999;282:786-90.

23.	Weintraub WS, Lüscher TF, Pocock S. The perils of surrogate endpoints. Eur Heart J 2015;36:2212-8.

24.	Pocock SJ, Lansky AJ, Mehran R, et al. Angiographic surrogate end points in drug-eluting stent trials: a systematic evaluation based on individual patient data from 11 randomized, controlled trials. J Am Coll Cardiol 2008;51:23-32.

25.   Granger CB, Vogel V, Cummings SR, et al. Do we need to adjudicate major clinical events? Clinical Trials 2008;5:56-60.

26.   Gómez G, Gómez-Mateu M, Dafni U. Informed choice of composite end points in cardiovascular trials. Circ Cardiovasc Qual Outcomes 2014;7:170-8.

27.   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.

28.   Rogers JK, Pocock SJ, McMurray JJV, 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.

29.   Novartis Pharmaceuticals. Efficacy and Safety of LCZ696 Compared to Valsartan, on Morbidity and Mortality in Heart Failure Patients With Preserved Ejection Fraction (PARAGON-HF). 2015. Available at: https://clinicaltrials.gov/ct2/show/NCT01920711. Accessed October 26, 2015.

30.   Rogers JK, Jhund PS, Perez AC, et al. Effect of rosuvastatin on repeat heart failure hospitalizations: the CORONA Trial (Controlled Rosuvastatin Multinational Trial in Heart Failure). J Am Coll Cardiol HF 2014;2:289-97.

31.   Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med 2010;29:3245-57.

32.   Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Stat Med 1999;18:1341-54.

33.   Pocock SJ, Ariti CA, Collier TJ, et al. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J 2012;33:176-82.

34.   Packer M. Proposal for a new clinical end point to evaluate the efficacy of drugs and devices in the treatment of chronic heart failure. J Card Fail 2001;7:176-82.

35.   Feres F, Costa RA, Abizaid A, et al.; OPTIMIZE Trial Investigators. Three vs twelve months of dual antiplatelet therapy after zotarolimus-eluting stents: the OPTIMIZE randomized trial. JAMA 2013;310:2510-22.

36.   Schroll JB, Maund E, Gøtzsche PC. Challenges in coding adverse events in clinical trials: a systematic review. PLoS One 2012;7:e41174.

37.   Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800-2.

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

39.   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.

40.   Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002;359:515-9.

41.   Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

42.   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 2015;385:2465-76.

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

44.   Treasure T, MacRae KD. Minimisation: The platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does. BMJ 1998;317:362-3.

45.   Murphy GJ, Pike K, Rogers CA, et al.; TITRe2 Investigators. Liberal or restrictive transfusion after cardiac surgery. N Engl J Med 2015;372:997-1008.

46.   Dumville JC, Hahn S, Miles JNV, Torgerson DJ. The use of unequal randomisation ratios in clinical trials: a review. Contemp Clin Trials 2006;27:1-12.

47.   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.

48.   Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002;359:696-700.

49.   Psaty BM, Prentice RL. Minimizing bias in randomized trials: the importance of blinding. JAMA 2010;304:793-4.

50.   Granger CB, Alexander JH, McMurray JJV, et al.; ARISTOTLE Committees and Investigators. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2011;365:981-92.

51.   Antman EM. Evidence and education. Circulation 2011;123:681-5.

52.   Gulati R, Raphael CE, Negoita M, et al. The rise and fall and possible resurrection of renal denervation. Nat Rev Cardiol 2015; In press.

53.   Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002;24:39-53.

54.   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: **Fundamentals of Design of Major Randomized Controlled Trials**

Figure 1: **Geographic Disparities in the TOPCAT Trial's Results for the Primary Endpoint**
(cardiovascular death and heart failure hospitalization).

Figure 2: **Total Number of Primary Events Needed to Detect a Specific True Hazard Ratio
as Statistically Significant** ($\alpha = 0.05$).

Table 1: Key Issues Regarding Composite Endpoints

Considerations	Discussion
Statistical power	For adequate power, trials with a single endpoint may be too large to be practical. Including several endpoints increases event frequency and may increase power.
Selecting components of the composite	A single endpoint may not fully document treatment effect. Composite endpoint captures several clinical aspects. Choose components so overall effect is meaningful.

Evaluating treatment effect	Interpretation can be difficult:
	
	    - Components may differ in clinical importance.
	
	    - Impact of treatment may vary across components.
	
	It is best to have separate safety and efficacy endpoints.

Analysis considerations	Emphasis usually on time to first event, which is often less serious (e.g., hospitalization) than subsequent events (e.g., death). Always show results for each separate component. Methods exist for repeat events and for competing risks, but complexity is a challenge. Win Ratio method prioritizes events of greater clinical importance.

Table 2: Key Issues Regarding Randomization Methods

	Discussion
Allocation concealment	Ensure treatment assignment cannot be predicted <u>in advance</u> of patient entry.
Double-blind	Patients, investigators, and those evaluating outcome remain unaware of assigned treatment <u>after</u> randomization.
Single-blinding	Investigators (and possibly patients) are aware of the assigned treatment, whereas those evaluating outcomes remain blinded. Superior to an unblinded design, but introduces greater opportunities for bias than a double-blind design.
Methods of randomization	
Simple randomization	Like coin tossing, with no connection between allocations. May lead to treatment imbalance in numbers or key patient factors.
Random permuted blocks	Treatment numbers are equal after each block of patients. Order of treatments within each block is random. Block sizes may vary to avoid predictability if trial not double-blinded.
Stratification	Aims to ensure balance for key patient factors across treatment groups. Each combination of factors (e.g., center and sex) has its own random permuted blocks. Must avoid over-stratification (e.g., 4 binary factors = 16 strata), which may introduce imbalances.
Minimization	A dynamic approach. Each treatment allocation done to achieve the best balance across several patient factors.
Unequal randomization	Can allocate more patients on new treatment (e.g. 2:1 ratio). Increases knowledge of new treatment and may enhance investigator/patient enthusiasm Requires more patients.