**Statistical Controversies in Reporting of Clinical Trials**

Stuart J. Pocock, PhD\*, John JV McMurray, MD† and Tim J. Collier, MSc\*

\* Department of Medical Statistics, London School of Hygiene & Tropical Medicine

† Institute of Cardiovascular and Medical Sciences, University of Glasgow

**Corresponding Author:** Stuart Pocock, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK.

email: Stuart.Pocock@LSHTM.ac.uk.

**Declarations:** There are no disclosures for any authors - this is an educational article.

**Abstract**

This article tackles several statistical controversies that are commonly faced when reporting a major clinical trial. Topics covered include: multiplicity of data, interpreting secondary endpoints and composite endpoints, the value of covariate adjustment, the traumas of subgroup analysis, assessing individual benefits and risks, alternatives to analysis by intention to treat, interpreting surprise findings (good and bad), and the overall quality of clinical trial reports. All is put in the context of topical cardiology trial examples, and is geared to help trialists to steer a wise course in their statistical reporting, thereby giving readers a balanced account of trial findings.

**Word count: 6830**

**Abbreviations**

RCT: Randomised Clinical Trial; SAP: Statistical Analysis Plan; CI: Confidence Interval; MACCE: Major Adverse Cardiac or Cerebrovascular Events; ITT: Intention to Treat; CABG: Coronary Artery Bypass Grafting; DES: Drug Eluting Stent; PCI: Percutaneous Coronary Intervention; TAVI: Trans-Aortic Valve Implant; MI: Myocardial Infarction.

**Introduction**

Last week’s review article covered the fundamentals of statistical analysis and reporting of randomised clinical trials (RCTs). We now extend those ideas to discuss several controversial statistical issues that are commonly faced in the presentation and interpretation of trial findings.

We explore the problems faced by the multiplicity of data available from any RCT, especially regarding multiple endpoints and subgroup analyses. Interpreting composite endpoints is a particular challenge. There is an inconsistency regarding the use of covariate-adjusted analyses. There is a need for more trials to assess how their overall findings can be translated into assessment of individual patient absolute benefits and absolute risks. The merit of analysis by intention to treat is considered alongside other options such as on-treatment analysis. One rarely discussed topic is how to interpret surprisingly large treatment effects (both good and bad) in new trials that are often quite small in scale.

All these controversies are summarised in the Central Illustration and illustrated by topical examples from cardiology trials. The overall aim in clarifying these issues is to enhance the quality of clinical trial reports in medical journals. The same principles apply to conference presentations and sponsor press releases which are even more prone to distortive reporting.

**Multiplicity of Data**

The key challenge in any report of a major RCT is how to provide a balanced account of the trial’s findings given the large number of variables collected at baseline and during follow-up, commonly called a “multiplicity of data” ([1](#_ENREF_1)). So out of the potential chaos of all the innumerable Tables and Figures that could be produced for purposes of treatment comparison, how do we validly select what to include in the finite confines of a trial publication in a major journal? Especially, how do we ensure that such a condensed trial report is fair in what it includes, i.e.we resist temptation to “play up the positive” by devoting more space in Results and Conclusions to those findings that put the new treatment in a good light?

A first step to overcome this is to have a pre-defined Statistical Analysis Plan (SAP) which is fully signed off before database lock and study unblinding. This SAP is prepared by trial statisticians and approved by the Trial Executive all of whom must be blind to any interim results by treatment group. A good SAP will not only document exactly which analyses are to be done, but will also elucidate relevant priorities in their interpretation, especially regarding the primary hypothesis, secondary hypotheses, any pre-defined safety concerns and a potential plethora of exploratory analyses (e.g. subgroup analyses) which are more hypothesis generating in spirit.

A particular focus is on the pre-defined primary endpoint with clear definition of the endpoint itself, the time of follow-up included (either a fixed period, e.g. 90 days, or a fixed calendar date for follow-up of all patients), and the precise statistical method for determining its point estimate, confidence interval and P-value. For time-to-event outcomes this is commonly a hazard ratio (and 95% CI) with logrank P-value. But sometimes a covariate-adjusted analysis is primary (see later discussion on this).

It is good practice to have a pre-defined and limited set of secondary endpoints for treatment efficacy. Their results are shown alongside those of the primary endpoint, e.g. as in Table 1 for the PEGASUS trial ([2](#_ENREF_2)) comparing two doses of ticagrelor with aspirin in patients with prior myocardial infarction. In this instance, interpretation appears straightforward since the primary endpoint achieved statistical significance for each ticagrelor dose versus placebo and all secondary efficacy endpoints showed trends in the same direction, except for no difference in all cause death for the higher ticagrelor dose. However, excesses of major bleeding and dyspnoea on ticagrelor, mean that such efficacy is offset by safety concerns.

But when the primary endpoint findings are inconclusive, claims of efficacy for any secondary endpoints are more of a challenge. For instance, the PROactive ([3](#_ENREF_3)) trial of pioglitazone versus placebo in 5238 diabetic patients had a primary composite endpoint of death, myocardial infarction, stroke, acute coronary syndrome, endovascular surgery or leg amputation. Over a mean 3 years follow-up the hazard ratio was 0.90 (95% CI 0.80 to 1.02) P=0.095. The main secondary endpoint, the composite of death, myocardial infarction and stroke had hazard ratio 0.84 (95% CI 0.72 to 0.98) P=0.027. The publication’s conclusions highlighted the latter downplaying the lack of statistical significance for the primary endpoint, whereas a more cautious interpretation is usually warranted.

In contrast, publication of the MATRIX trial ([4](#_ENREF_4)) comparing bivalirudin or unfractionated heparin in acute coronary syndromes had conclusions confined to the co-primary endpoints MACE (death, MI or stroke) and NACE (death, MI, stroke or major bleed) both of which “were not significantly lower with bivalirudin than with unfractionated heparin”. While the focus on primary endpoints is appropriate, there is a danger that it hides important differences amongst secondary (component) endpoints. While cautious interpretation is essential across a multiplicity of secondary endpoints, conclusions would have benefited from mentioning that bivalirudin had more stent thromboses (P=0.048), fewer major bleeds (P<0.001) and fewer deaths (P=0.04). Such intriguing secondary findings need clarification from other related trials.

When a secondary endpoint reveals potential harm of a treatment, controversy is likely to ensue. For instance, in the SAVOR trial ([5](#_ENREF_5)) of saxagliptin versus placebo in diabetic patients the composite primary endpoint (CV death, MI and stroke) showed no treatment difference but one of several secondary endpoints, heart failure hospitalisation revealed an excess on saxagliptin hazard ratio 1.27 (95% CI 1.07 to 1.51) P=0.007. The risk of a type I error (false positive) runs high when looking at multiple endpoints (1 primary and 10 secondary in this instance), so the play of chance cannot be ruled out. Two subsequent trials EXAMINE ([6](#_ENREF_6)) and TECOS ([7](#_ENREF_7)) of drugs in the same class, alogliptin and sitagliptin respectively, reveal no excess of heart failure and there is no plausible biological explanation as to why the drugs might differ in this respect. Furthermore, a statistical test of heterogeneity comparing the three trials’ hazard ratios for heart failure is not statistically significant (interaction P=0.16) and the combined hazard ratio is 1.13 (P=0.04) and 1.12 (P=0.18) for fixed and random effect meta-analyses respectively (see Figure 1). This analysis partly hinges on the concept that similar effects should be expected for drugs in the same class. This is often the case but there are exceptions e.g. torcetrapib versus other CETP inhibitors, and ximelagatran versus other direct thrombin inhibitors regarding liver abnormalities. This, while one cannot rule out the possibility of a real problem here unique to saxagliptin, the evidence of harm lacks conviction and should be interpreted with caution.

Regulatory authorities and trial publications in medical journals have somewhat different perspectives when it comes to interpreting secondary endpoints. If the primary endpoint is neutral, the efficacy claims for secondary endpoints may be cautiously expressed in the medical literature (usually with less emphasis than authors might wish), while it is highly unlikely that regulators will approve a drug on this basis. Regulators face a dilemma when secondary endpoint suggestions of potential harm arise, as in the SAVOR trial ([5](#_ENREF_5)). There is an asymmetry here in that the corresponding extent of evidence in the direction of treatment benefit would receive scant attention. While there is an obvious need to protect patients from any harm, regulators need to recognize the statistical uncertainties whereby effective treatments might be unjustly removed based on weak evidence of potential harm arising from data dredging across a multiplicity of endpoints.

**Composite endpoints** are commonly used in cardiovascular RCTs to combine evidence across two or more outcomes into a single primary endpoint. But there is a danger of oversimplifying the evidence by putting too much emphasis on the composite without adequate inspection of the contribution from each separate component. For instance, the SYNTAX trial ([8](#_ENREF_8),[9](#_ENREF_9)) of bypass surgery (CABG) versus the TAXUS drug-eluting stent (DES) in 1800 patients with left main or triple vessel disease had a Major Adverse Cardiac or Cerebrovascular Events (MACCE) composite primary endpoint comprising of death, stroke, myocardial infarction and repeat revascularisation: results at 1 year and 5 years follow-up are shown in Table 2. At one year there was highly significant excess of MACCE events after DES which at face value indicates that DES is inferior to CABG. But here is a more complex picture not well captured by this choice of primary endpoint. The main difference is in repeat revascularisation, the great majority of which is repeat PCI. One could argue that 10% of patients having a second PCI is less traumatic than the CABG received by 100% of the CABG group, so this component of the primary endpoint is not well representing the comparison of overall patient well-being. At one year, there is a significant excess of strokes after CABG, and no overall difference in the composite of death, MI and stroke. A general principle that often occurs in other interventional trials, e.g. complete or culprit lesion intervention in primary PCI, is that clinically driven interventions should not be part of the primary endpoint.

A second important point raised by SYNTAX is that in such strategy trials the key treatment differences may well be revealed with longer-term follow-up. At 5 years, there is a highly significant excess of MIs in the DES group and this drives the composite of death, MI and stroke also to be in favor of CABG.

This example illustrates how for composite endpoints “the devil lies in the details”. In the ongoing EXCEL ([10](#_ENREF_10)) trial of CABG versus everolimus-eluting stent in left main disease, the primary endpoint is death, MI and stroke after 3 years providing an appropriate longer-term perspective on the key major cardiovascular events.

**Covariate Adjustment**

Should the key results of an RCT be adjusted for baseline covariates, which covariates should be chosen (and how), and which results should be emphasized? ([11](#_ENREF_11)) Practice varies widely: for some RCTs only unadjusted results are presented, others have covariate adjustment as their primary analysis while yet others use it as a secondary sensitivity analysis. This inconsistency of approach across trials is perhaps tolerated because in major trials randomisation ensures good balance across treatments for baseline variables and hence covariate adjustment usually makes little difference.

The EMPHASIS-HF trial ([12](#_ENREF_12)) of eplerenone vs placebo in 2737 chronic heart failure illustrates the consequences of covariate adjustment. They pre-defined use of a proportional hazards model adjusting for 13 baseline covariates: age, estimated GFR, ejection fraction, body mass index, haemoglobin value, heart rate, systolic blood pressure, diabetes mellitus, history of hypertension, previous myocardial infarction, atrial fibrillation, and left bundle-branch block or QRS duration greater than 130 msec. Selection was sensibly based on prior knowledge/suspicion of each variable’s association with patient prognosis. Table 3 shows the adjusted and unadjusted hazard ratios for eplerenone versus placebo for the primary endpoint and also for its two separate components. In all three instances, the adjusted hazard ratio was slightly further from 1, as one would expect when adjusting for factors that are related to prognosis ([13](#_ENREF_13)). Unlike Normal regression models, covariate-adjustment for binary or time-to-event outcomes using logistic or proportional hazard models does not increase the precision of estimates (confidence interval width changes little): rather point estimates, (e.g. odds ratio, hazard ratio) tend to move further away from the null. Thus, there is a slight gain in statistical power in adjusting for covariates, but only if the chosen covariates are related to patient prognosis. If misguidedly one chooses covariates not linked to prognosis then covariate-adjustment will make no difference.

One misperception is that covariate-adjustment should be done for the stratification factors used in randomisation. This was specified in IMPROVE-IT ([14](#_ENREF_14)) in acute coronary syndrome where stratification factors were prior lipid-lowering therapy, type of acute coronary syndrome, and enrolment in another trial (yes/no). Clearly, these are not the most important issues affecting prognosis in ACS (age is the strongest risk factor) and such adjustment while harmless, might be considered of little value.

Adjustment for geographic region is also sometimes performed, e.g. PARADIGM-HF ([15](#_ENREF_15)) adjusted hazard ratios for 5 regions, one of which curiously was Western Europe plus South Africa and Israel. Again, this will do no harm but is a cosmetic exercise missing out on the real purpose of covariate-adjustment.

Some argue that one should adjust for baseline variables that show an imbalance between treatment groups. For instance, the GISSI-HF trial ([16](#_ENREF_16)) adjusted for variables that were unbalanced between randomised groups at P<0.1. As a secondary sensitivity analysis it can add reassurance that the primary analysis makes sense, but again if the covariates with imbalance do not affect prognosis such adjustment will make negligible difference.

Occasionally, when an unadjusted analysis achieves borderline significance the use of an appropriately pre-defined covariate adjustment can add weight to the conclusions. For instance, in the CHARM trial ([17](#_ENREF_17)) in 7599 heart failure patients the unadjusted hazard ratio (candesartan vs placebo) for all-cause death over a median 3.2 years was 0.90 (95% CI 0.83 to 1.00) P=0.055. A pre-specified secondary analysis, adjusting for covariates anticipated to affect prognosis, gave hazard ratio 0.90 (95% CI 0.82 to 0.99) P=0.032. This added credibility to the idea of a survival benefit for candesartan, especially given that for cardiovascular death the covariate-adjusted hazard ratio was 0.87 (95% CI 0.78 to 0.96), P=0.006.

In general, we feel that a well-defined appropriate covariate-adjusted analysis is well worth doing in major RCTs. After all, it offers a slight gain in statistical power at no extra cost and with minimal statistical effort, so why miss out on such an opportunity? The following principles should be followed:

1) Based on prior knowledge, one should specify clearly a limited number of covariates known (or thought) to have a substantial bearing on patient prognosis. Make sure such covariates are accurately recorded at baseline on all patients.

2) Document in a pre-specified SAP the precise covariate-adjusted model to be fitted. For instance, a quantitative covariate such as age can be either fitted as a linear covariate or in several categories (age-groups). Such a choice needs to be made in advance.

3) Post hoc variable selection, e.g. adding covariates unbalanced at baseline, dropping non-significant predictors or adding in new significant predictors after database lock, should be avoided in the primary analysis since suspicions may arise that such choices might have been made to enhance the treatment effect.

4) Both unadjusted and covariate-adjusted analyses should be presented, with pre-specification as to which is the primary analysis. If the choice of covariates is confidently based on experience of what influences prognosis, then it makes sense to have the covariate-adjusted analysis as primary ([18](#_ENREF_18)).

**Subgroup Analysis**

Patients recruited in a major trial are not a homogeneous bunch: their medical history, demographics and other baseline features will vary. Hence, it is legitimate to undertake subgroup analyses to see whether the overall result of the trial appears to apply to all eligible patients or whether there is evidence that real treatment effects depend on certain baseline characteristics.

Of all the multiplicity problems in reporting RCTs, interpretation of subgroup analyses presents a particular challenge ([19](#_ENREF_19)). First trials usually lack power to reliably detect subgroup effects. Second there are many possible subgroups that could be explored and one needs to guard against data dredging eliciting false subgroup claims. Third, statistical significance (or not) in a specific subgroup is not a sound basis for making (or ruling out) any subgroup claims: instead one needs statistical tests of interaction to directly infer whether the treatment effect appears to differ across subgroups.

We explore these ideas in a few examples. First, subgroup analyses for the PARADIGM-HF trial ([15](#_ENREF_15)) are shown in Figure 2. This kind of Figure, called a Forest plot, is the usual way of documenting the estimated treatment effect within each subgroup (a hazard ratio in this case) together with its 95% CI. The 18 subgroups displayed were pre-specified and show a consistency of treatment effect, all being in the direction of superiority for LCZ696 compared to enalapril in this heterogeneous heart failure population, for both the primary endpoint and cardiovascular death. For reference, the results for all patients, with their inevitably tighter confidence intervals, are shown at the top of Figure 2.

Scanning across subgroups one can see that estimated hazard ratios vary by chance and confidence intervals (CIs) are wider for smaller subgroups. Some CIs overlap the line of unity indicating that the subgroup P-value does not reach 5% significance, but this will inevitably happen, especially in smaller subgroups, and is not helpful in interpreting subgroup findings. Instead, a statistical test of interaction should accompany each subgroup display (as shown in Figure 2). This interaction test examines the extent to which the observed difference in hazard ratios across subgroups may be attributed to chance variation. For the primary endpoint just one interaction test is statistically significant: P=0.03 for NYHA class I or II versus III or IV suggests a possible greater benefit of LCZ696 in less symptomatic patients, though no such interaction exists for cardiovascular death (interaction P=0.76). Given 18 subgroups analyses have been performed for each of two outcomes, one could expect at least one interaction P<0.05 purely by chance, so these data are overall supportive of a consistency of treatment effect across a broad spectrum of patients with heart failure.

When the overall result of a major RCT is neutral, it is tempting to search across subgroups to see if there is a particular subgroup in which the treatment effect is favourable. In this context subgroup claims require an especially cautious interpretation in a journal publication. Furthermore, it is highly unlikely that regulators such as the FDA would approve a drug based on such a positive subgroup claim.

The CHARISMA trial ([20](#_ENREF_20)) is an interesting case in point. Against a background of low dose aspirin 15,603 patients at high risk of atherothrombotic events were randomised to clopidogrel or placebo. Over a median 28 months, incidence of the primary endpoint (CV death, MI or stroke) was 6.8% versus 7.3%, P=0.22. But in symptomatic patients (78% of all patients) the findings for clopidogrel looked better: 6.9% vs 7.9%, P=0.046. In contrast, in asymptomatic patients results trended in the opposition direction: 6.6% vs 5.5%, P=0.02. The interaction test had P=0.045 and the authors’ conclusions included a claim of benefit for clopidogrel in symptomatic patients.

The accompanying editorial was critical commenting “the charisma of extracting favourable subgroups should be resisted” ([21](#_ENREF_21)). Why? Well, this was one of 12 pre-specified subgroup analyses and the strength of evidence, P for interaction, was borderline. Also, it is usually biologically implausible that a true treatment effect will be in opposite directions across subgroups, i.e.so-called qualitative interactions rarely arise across clinical medicine. The New England Journal has subsequently toughened its policy regarding subgroup analyses ([22](#_ENREF_22)), so that they be seen more as exploratory and hypothesis-generating rather than becoming part of a trial’s key conclusions.

A different challenge arises when the overall results of a trial are positive, but there appears to be a lack of superiority in a particular subgroup. For instance, the SPIRIT IV ([23](#_ENREF_23)) trial comparing everolimus-eluting stent (EES) or paclitaxel-eluting stent (PES) in 3687 patients showed overall superiority for ESS for the primary endpoint target-lesion failure at one year: 4.2% vs 6.8%, P=0.001. But in 1140 diabetic patients (one of 12 subgroup analyses) there was no evidence of a treatment difference: 6.4% vs 6.9%, interaction P=0.02. This finding in itself is not definitive evidence, and hence further evidence was sought to confirm (or refute) this finding. A pooled analysis of two-year outcomes across four RCTs of EES versus PES in 6780 patients ([24](#_ENREF_24)) revealed marked superiority of EES in non-diabetics for death, myocardial infarction, stent thrombosis and ischaemia-driven target lesion revascularisation, while no such benefits of EES existed for diabetic patients. All four interaction tests were convincingly significant: P=0.02, P=0.01, P=0.0006, P=0.02 respectively.

While such confirmatory evidence of an initial subgroup finding is highly desirable it is not always achievable. But still regulatory decisions need to be made on whether apparent lack of efficacy in a subgroup merits a specific restriction re drug approval. For instance, the European Medicines Agency restricted use of ivabradine in chronic heart failure to patients with heart rate ≥75 bpm on the basis of a significant interaction in one pivotal trial ([25](#_ENREF_25)).

Overall, there is a responsibility to perform and present subgroup findings from major RCTs. Pre-specification of a limited number of intended subgroup analyses is a helpful guard against post hoc manipulations of data, but still interpretations are restricted by a lack of statistical power and a multiplicity of hypotheses, so due caution is required in not over-reacting to any subgroup claims. Subgroup analysis becomes most convincing when it relates to just one pre-declared factor of especial interest (e.g. troponin positive versus negative in GP IIb/IIIa trials) where an interaction is anticipated to exist.

**Assessing Individual Benefits and Risks**

In most RCT reports the focus is on the overall relative efficacy and relative safety of the treatments being compared. But even in the absence of any subgroup differences on a relative scale (e.g. based on hazard ratios or odds ratios) there may well be important differences between individuals as regards absolute treatment benefits ([26](#_ENREF_26)).

For instance, in the EMPHASIS-HF trial ([12](#_ENREF_12)) of eplerenone versus placebo in heart failure patients with mild symptoms the composite primary endpoint, CV death and heart failure hospitalisation showed a marked benefit over a median 21 months follow-up, hazard ratio 0.63 (95% CI 0.54 to 0.74) P<0.0001. There were no apparent subgroup effects on a hazard ratio scale. In a subsequent analysis each patient was then classified into low, medium and high risk on the basis of a multivariable risk score using 10 commonly recognised prognostic features ([27](#_ENREF_27)). Table 4 shows the consequent treatment benefits by risk group, on both relative and absolute scales. As anticipated, the hazard ratio was similar in all risk groups. But the absolute benefits varied markedly by risk: the estimated reduction due to eplerenone in the primary event rate per 100 patient years was 2.0, 6.8 and 15.2 in low, medium and high risk patients respectively. Inevitably, when there is no interaction on a relative scale, there will often be a marked interaction on an absolute scale across subgroups of differing risk status.

A pooling of data for three trials ([28](#_ENREF_28)) of routine invasive versus selective invasive strategies in acute coronary syndrome found a significant difference in 5 year risk of CV death or myocardial infarction: hazard ratio 0.81, 95% CI 0.71 to 0.93, P=0.002. This was consistent across risk groups, but when expressed on an absolute scale the benefit of a routine invasive strategy is more marked in higher risk patients: for low, intermediate and high risk patients the reductions in five year risk of CV death and MI were 2.0%, 3.8% or 11.1% respectively. In contrast, it could happen that higher risk patients have a lower relative benefit, but because of their higher risk their absolute benefit was similar to those at lower risk e.g. the elderly (age>75 years) and fibrinolysis.

These two examples illustrate how one needs to consider the individual’s risk status in determining whether the absolute benefit of an intervention is sufficient to merit its use in their case. Note this is achieved by multivariable risk analysis rather that univariate subgroups. This becomes particularly important if treatment efficacy is offset by a risk of side effects. For instance, the TRITON trial ([29](#_ENREF_29)) compared prasugrel against clopidogrel in 13, 608 patients with acute coronary syndrome. Over a median 14.5 months, incidence of the primary endpoint (CV death, MI and stroke) was less on prasugrel 9.9% versus 12.1%, hazard ratio 0.81 95% CI 0.73 to 0.90, P<0.001. This benefit was mainly due to a reduction in non-fatal MIs: 7.3% versus 9.5%. But there were more bleeding events on prasugrel, e.g. TIMI major bleed 2.4% versus 1.8%, P=0.03.

In order to weigh the benefits and risks of prasugrel versus clopidogrel on an individual patient basis, Salisbury et al ([30](#_ENREF_30)) used multivariable logistic models to separately predict any patient’s risk of 1) the primary ischaemic endpoint and 2) TIMI major or minor bleed, taking account of both randomised treatment and patient characteristics. The intent is to quantify on an absolute scale how the trade-off between treatment differences in ischaemic efficacy and bleeding harm is patient specific. For instance, for elderly women the bleeding risk is of greater concern while in a younger man with known CV risk factors avoidance of future ischaemic events is paramount: clopidogrel is the drug of choice for the former while prasugrel is a better choice for the latter. These multivariable models are a quantitative aid to such clinical judgement and may be of particular use in the broader spectrum of patients for whom the efficacy/safety trade-off is less clinically obvious.

Similar principles can apply when deciding what dose of a drug is appropriate for the individual patient. In stroke prevention for patients with atrial fibrillation, the RE-LY ([31](#_ENREF_31)) and ENGAGE AF ([32](#_ENREF_32)) trials are to be commended for comparing two different anticoagulant drug doses (of dabigatran and edoxaban respectively) against warfarin. In both three-arm trials, the higher dose appeared more effective in reducing risk of stroke and systemic embolism, while the lower dose had less bleeding events. At present, these two trials have confined attention to the overall findings and conventional subgroup analyses, whereas we would encourage more model-based approaches to better identify which types of patient (if any) would have a net benefit for the lower dose. Ideally, such risk models should be based on external data, but realistically this is often not possible.

In general, reports of clinical trials mainly focus on an overall treatment comparison with cautious reference to subgroup analyses based on one baseline variable at a time. Therefore, the opportunity to link trial findings to individualised patient care based on “whole patient” risk profiles is largely being missed ([33](#_ENREF_33)).

**Analysis by Intention to Treat and Other Options**

Analysis by intention to treat (ITT) means that a trial’s results include the totality of patient follow-up for all randomised patients. For major RCTs with a superiority hypothesis it is generally regarded as the main approach to reporting of trial findings for treatment efficacy in both medical journals and regulatory submissions. The advantage of ITT is that it provides an unbiased comparison of treatment strategies as delivered in practice: there is no scope for post hoc selection of who to include for how long. Everyone is included, with no escape! Such logic appears soundly based but there are two complications to consider: 1) do we truly have full follow-up data available for everyone? and 2) are we really happy to include all protocol deviations in a pure ITT or is a modified ITT appropriate?

On the first point, the more patients that are lost to follow-up the further the attempted analysis deviates from true ITT. In trial conduct, it is important to minimize the loss to follow-up. High treatment compliance is a first step. Also, when patients do withdraw from treatment their follow-up should continue if at all possible. In most time-to-event analyses there is variation in observed patient follow-up: commonly recruitment takes 1 to 2 years and all patients are followed to a fixed calendar date. If all patients not experiencing a primary endpoint reach that date then a true ITT analysis is done, and in producing Kaplan Meier plots etc. censoring is sensibly assumed to be non-informative. But if patients are lost to follow-up at an earlier stage this cannot be assumed to occur at random, e.g. patients who drop out may be sicker and hence at higher risk of a primary event, which goes unrecorded. Thus loss to follow is potentially informative censoring and could lead to a biased treatment comparison. This becomes particularly serious if the rate of drop-out, and its reasons, differ between treatment groups.

The ATLAS ACS 2 trial ([34](#_ENREF_34)) illustrates this problem. This three-arm trial compared two doses of rivaroxaban and placebo in 15,526 patients followed for a mean of 13 months. Published findings looked particularly good for the lower rivaroxaban dose with superiority for the primary endpoint (CV death, MI and stroke) and for all-cause death, though with some increase in bleeding events compared with placebo. However, under the scrutiny of an FDA Advisory Panel the extent of incomplete follow-up became evident: with 15.5% of patients prematurely discontinued from the study. Specifically 8.3% withdrew consent for whom the great majority had unknown vital status at trial end. This problem cast sufficient doubt on the robustness of trial findings to influence the FDA Panel not to recommend approval. There is no fixed guidance as to what level of dropout becomes unacceptable, though ATLAS ACS 2 clearly did less well in this regard than several other recent trials in acute coronary syndrome, though of course none could achieve 100% follow-up ([35](#_ENREF_35)).

While presenting ITT analyses, ATLAS ACS 2 ([33](#_ENREF_33)) put greater emphasis on what they called a modified ITT approach: that is, any events occurring more than 30 days after study drug discontinuation were excluded from analysis. This is perhaps more commonly called a “per protocol” or “on treatment” analysis and is usually downgraded to a secondary analysis with prime focus being on ITT. Use of the term “modified ITT” is quite common in clinical trial reports but there is a lack of consistency in what it means. Less desirable features are any post-randomisation exclusions, since these could lead to bias ([36](#_ENREF_36)). More acceptable modifications are exclusions of ineligible patients incorrectly randomised and, in double blind trials, exclusions of any patients who never got a single dose of study drug.

The APPROVe trial ([37](#_ENREF_37),[38](#_ENREF_38)) illustrates how obtaining an appropriate ITT analysis is important in reaching valid conclusions. The trial found an excess risk of cardiovascular events on rofecoxib compared to placebo in 2586 patients with a history of colerectal adenomas. The first report only included events occurring during treatment and up to 14 days after last dose: 46 vs 26 patients with thrombotic events, P=0.008. It was claimed that event rates were similar in the first 18 months and the excess only emerged thereafter (see Figure 3, top graph). This “on treatment” analysis did not give the whole story, and a subsequent ITT analysis, still with some missing follow-up, revealed a somewhat different pattern. There were now 59 vs 34 patients with thrombotic events, P=0.006 and the evidence appeared compatible with an early increase in risk that persists one year after stopping treatment (see Figure 3, bottom graph). Rofecoxib was withdrawn from worldwide markets due to these safety concerns, though it is worth noting that the second report ([38](#_ENREF_38)) was published over 3 years after this withdrawal.

For a trial in which both non-inferiority and superiority hypotheses are of interest then both per-protocol analysis and ITT analysis are relevant. For non-inferiority testing the per-protocol analysis is often deemed primary, on the basis that ITT includes time when patients are off study drug which may dilute any real treatment effects making it artificially too easy to claim non-inferiority. But for tests of superiority ITT gets priority. The TECOS trial ([7](#_ENREF_7)) of sitagliptin versus placebo in type 2 diabetes illustrates this approach. The primary event (CV death, MI, stroke, unstable angina) occurred in 839 vs 851 patients in ITT analysis, P=0.65. Over a median 3 years follow-up a sizeable minority of patients stopped taking their study drug, so per protocol analysis had 695 vs 695 primary events: hazard ratio 0.98 (95% CI 0.88 to 1.09). Non-inferiority of sitagliptin was clearly established and there is no evidence that it reduces risk of CV events. In the both treatment groups the event rates in ITT analysis are higher than in per-protocol analysis. That is, becoming non-adherent is associated with a high risk, which is a common feature across most RCTs.

The ROCKET-AF trial ([39](#_ENREF_39)) comparing rivoraxaban and warfarin in 14,264 patients with atrial fibrillation had a similar construct with both non-inferiority and superiority hypotheses. In ITT analysis the primary event, stroke or systematic embolism, occurred less frequently on rivaroxaban: 269 vs 306 events, hazard ratio 0.88 (95% CI 0.75 to 1.03) P=0.12. Given this lack of significance, interest turned to the events occurring on study drug (plus within 2 days of stopping), the per-protocol population: 188 vs 240 events, hazard ratio 0.79 (95% CI 0.66 to 0.96) P=0.02. However this secondary analysis for superiority, with its potential bias as always, did not sway the evidence towards a claim of superiority and the published conclusion was “rivaroxaban was non-inferior to warfarin for the prevention of stroke or systemic embolism”. In general, per-protocol analyses introduce a bias: non-adherent patients are a select group, often but not automatically at high risk of outcome events, which makes interpretation unreliable.

In unblinded trials of alternative treatment strategies, there is a risk that some patients do not pursue their randomly allocated strategy. For instance, in the PARTNER trial ([40](#_ENREF_40)) of trans-aortic valve implant (TAVI) versus aortic valve surgery there were 4 (1%) and 38 (11%) respectively who did not get their intended treatment. ITT from randomisation was the primary analysis, but a supplementary as-treated analysis from time of treatment (excluding those 42 patients) helped to confirm no mortality difference but a borderline significant excess of strokes. In this situation, ITT will tend to dilute any real treatment differences (while giving a valid comparison of strategies along with their protocol deviations) and an as-treated analysis may be biased, e.g. it may be poorer risk patients who declined surgery. A consistency across the two analyses is helpful. For non-inferiority trials both ITT and per-protocol analyses should be presented, a point we clarify further in the last article of this series.

**Interpreting Surprises, both Good and Bad**

From time to time an unexpected finding arises from a clinical trial. The surprise may relate to an endpoint for which there was no prior hypothesis, a subgroup that appears inconsistent with the overall treatment effect or an unduly large treatment effect that exceeds prior expectations. It may relate to treatment benefit or harm.

The cycle of progress in medical research needs to be born in mind. A new dramatic claim (whether benefit or harm) is often based on a small study. Accordingly, it is prone to exaggerate the issue, even if the study has no design flaw. The issue becomes high profile without adequate recognition of all the selection biases that have taken place, e.g. across multiple analyses (endpoints, trials, subgroups) if one focuses on the most extreme result it will look more impressive than is justified.

One then needs to collect more substantial evidence on the issue, e.g. continued follow-up, a larger trial or a meta-analysis of related trials. Some “regression to the truth” is liable to occur whereby the consequent effect turns out to be more modest, and sometimes not present at all. We illustrate this pattern with some examples, starting with potential safety concerns arising from RCTs.

In the SEAS trial ([41](#_ENREF_41)) in 1873 patients with aortic stenosis the active treatment simvastatim plus ezetemibe had an unexpected excess of cancers compared with placebo: 105 vs 70 incident cancers P=0.01, and 39 vs 23 cancer deaths P=0.05. Given the wealth of safety data available on statins, the potential culprit was thought to be ezetemibe. There was an urgent need to study the totality of evidence regarding ezetemibe and cancer. Hence, two ongoing large trials of ezetimibe versus placebo on background statin SHARP and IMPROVE-IT published their combined interim findings: 313 vs 329 incident cancers P=0.61 and 97 vs 72 cancer deaths P=0.07 for ezetimibe and placebo groups respectively ([42](#_ENREF_42)). There were also no logical pattern re specific cancers. The conclusion was the “the available evidence do not provide any credible evidence of any adverse effect of ezetimibe on rates of cancer”, which was confirmed by the larger numbers of events in the final results of these two trials.

A similar pattern emerged with an apparent excess of myocardial infarction on rosiglitazone, first proposed in a meta-analysis by Nissen and Wolski: odds ratio versus control 1.43 (95% CI 1.03 to 1.98) P=0.03 ([43](#_ENREF_43)). The main subsequent evidence came from the RECORD trials ([44](#_ENREF_44)) in 4447 diabetic patients followed for a mean 5.5 years: hazard ratio for myocardial infarction (rosiglitazone vs active control) was 1.14 (95% CI 0.80 to 1.63), P=0.47. After much FDA scrutiny over several years it was finally concluded that rosiglitazone does not increase the risk of myocardial infarction. However, the overall safety profile, especially risks of heart failure and bone fractures, meant that the marketing authorisation for the drug was suspended in Europe. One general lesson here is that meta-analyses of small trials require a very cautious interpretation pending more solid evidence from large prospective randomised trials.

Another possibility is that a drug’s initial signal of harm is exaggerated but further evidence does substantiate that a real problem exists. For instance, the first evidence of risk of myocardial infarction attributed to rofecoxib came from the VIGOR trial ([45](#_ENREF_45)) of rofecoxib vs naproxen in patients with rheumatoid arthritis: 20 vs 4 myocardial infarctions, relative risk 5.00 (95% CI 1.68 to 20.13). Curiously, the original publication reported it as a benefit of naproxen relative risk 0.2 (95% CI 0.1 to 0.7). Here it is important to note the small numbers of events and hence the wide confidence interval. Subsequent evidence both from the APPROVe trial ([37](#_ENREF_37)) (see above) and from meta-analyses showed an effect closer to a doubling of risk, rather than a five-fold increase. Furthermore, the meta-analysis showed no heterogeneity of this risk across the class of Cox 2 inhibitors ([46](#_ENREF_46)). So was rofecoxib “the unlucky one” to first focus attention on this class phenomenon?

We now turn to claims of treatment efficacy based on apparently large benefits in small trials. For instance, a trial of acetylcysteine vs placebo ([47](#_ENREF_47)) for prevention of contrast induced nephropathy in 83 patients revealed 1 vs 9 acute reductions in renal function, P=0.01. This topic has yielded a number of other small trials and meta-analyses of the collective evidence show that findings are too inconsistent to warrant a conclusion of efficacy. A large well designed trial is needed to resolve this issue.

There is an ongoing controversy re the peri-operative use of beta-blockers in non-cardiac surgery. The small DECREASE trial ([48](#_ENREF_48)) of bisoprolol showed apparently marked benefits: in 112 patients there were 2 vs 9 deaths (P=0.02) and 0 vs 9 myocardial infarction (P<0.001) in bisoprolol and control groups respectively. In general it is wise to consider such dramatic findings based on small numbers as being too good to be true. Sadly, in this case scientific misconduct became evident making the study untrustworthy, and current guidelines indicate that the collective valid evidence re the value of beta-blockers is inconclusive. Again, a large well designed trial is needed.

The PRAMI trial ([49](#_ENREF_49)) of preventive angioplasty versus stenting of the culprit lesion only is an example of an apparently very large treatment effect based on relatively small numbers of events: 53 vs 21 primary endpoints (refractory angina, myocardial infarction or cardiac death) hazard ratio 0.35 (95% CI 0.21 to 0.58) P<0.001. For any intervention to reduce an event rate by more than half strikes one as implausible. Here, the trial stopped early, recruitment had been slow (and hence perhaps not representative) and the trial was unblinded: all of these may have contributed to an exaggeration of effect. Findings from a larger sequel trial, COMPLETE ([50](#_ENREF_50)), are awaited with interest.

**Enhancing the Quality of Clinical Trial Reports**

We conclude with some general remarks about the overall quality of clinical trial publications. CONSORT ([51](#_ENREF_51)) is an established set of guidelines for reporting clinical trials which many journals, including JACC, expect authors to adhere to. There is a helpful checklist of items to include (see Table 5) which covers all sections of an article including Methods, Results and Conclusions. Specific issues covered in CONSORT extensions include non-inferiority trials, pragmatic trials, reporting of harms, and what to include in a trial Abstract ([52](#_ENREF_52)).

Such guidelines do help, but the overall responsibility of trialists (and journal editors and reviewers) is to ensure that an honest, balanced account of a trial’s findings is provided. In particular, the Discussion should document any limitations in a trial’s design (e.g. what potential biases exist?), conduct (e.g. non-compliance, dropouts) and analysis (e.g. was ITT analysis achieved?). Relevant to the controversies discussed in this article is that authors make clear what are the pre-defined analyses and any priorities amongst them, e.g. the primary endpoint’s overall analysis should dominate the Conclusions and the Abstract, while any important safety issues are also adequately represented in both. Any other data explorations, (e.g. secondary endpoints, subgroup analyses) are relevant background but it is the authors’ and journal’s responsibility to ensure that a cautious interpretation is maintained. Nevertheless, controversies will continue to arise and we hope this article has provided a statistical insight which will help both trialists to present, and readers to acquire, a balanced perspective.

1. Cook RJ, Farewell VT. Multiplicity Considerations in the Design and Analysis of Clinical Trials. Journal of the Royal Statistical Society. Series A (Statistics in Society) 1996;159:93-110.

2. Bonaca MP, Bhatt DL, Cohen M, et al. Long-Term Use of Ticagrelor in Patients with Prior Myocardial Infarction. New England Journal of Medicine 2015;372:1791-1800.

3. Dormandy JA, Charbonnel B, Eckland DJA, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. The Lancet 2005;366:1279-1289.

4. Valgimigli M, Frigoli E, Leonardi S, et al. Bivalirudin or Unfractionated Heparin in Acute Coronary Syndromes. New England Journal of Medicine 2015;373:997-1009.

5. Scirica BM, Bhatt DL, Braunwald E, et al. Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus. New England Journal of Medicine 2013;369:1317-1326.

6. White WB, Cannon CP, Heller SR, et al. Alogliptin after Acute Coronary Syndrome in Patients with Type 2 Diabetes. New England Journal of Medicine 2013;369:1327-1335.

7. Green JB, Bethel MA, Armstrong PW, et al. Effect of Sitagliptin on Cardiovascular Outcomes in Type 2 Diabetes. New England Journal of Medicine 2015;373:232-242.

8. Serruys PW, Morice M-C, Kappetein AP, et al. Percutaneous Coronary Intervention versus Coronary-Artery Bypass Grafting for Severe Coronary Artery Disease. New England Journal of Medicine 2009;360:961-972.

9. Mohr FW, Morice M-C, Kappetein AP, et al. Coronary artery bypass graft surgery versus percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. The Lancet 2013;381:629-638.

10. Abbott Vascular I, U.S.A. Evaluation of Xience Prime everolimus-eluting stent system (EECSS) or Xience V EECSS versus coronary artery bypass surgery for effectiveness of left main revascularization (EXCEL) clinical trial. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2010- [cited 2015 August 07]. Available from: https://clinicaltrials.gov/ct2/show/NCT01205776 Identifier: NCT01205776 2010 2010.

11. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. Statistics in Medicine 2002;21:2917-2930.

12. Zannad F, McMurray JJV, Krum H, et al. Eplerenone in Patients with Systolic Heart Failure and Mild Symptoms. New England Journal of Medicine 2011;364:11-21.

13. Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. Statistics in Medicine 2002;21:2899-2908.

14. Cannon CP, Blazing MA, Giugliano RP, et al. Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. New England Journal of Medicine 2015;372:2387-2397.

15. McMurray JJV, Packer M, Desai AS, et al. Angiotensin–Neprilysin Inhibition versus Enalapril in Heart Failure. New England Journal of Medicine 2014;371:993-1004.

16. investigators G-H. Effect of rosuvastatin in patients with chronic heart failure (the GISSI-HF trial): a randomised, double-blind, placebo-controlled trial. The Lancet;372:1231-1239.

17. Yusuf S, Pfeffer MA, Swedberg K, et al. Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial. The Lancet 2003;362:777-781.

18. Committee for proprietary medicinal products (CPMP) points to consider on adjustment for baseline covariates. Statistics in Medicine 2004;23:701-709.

19. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. The Lancet 2000;355:1064-1069.

20. Bhatt DL, Fox KAA, Hacke W, et al. Clopidogrel and Aspirin versus Aspirin Alone for the Prevention of Atherothrombotic Events. New England Journal of Medicine 2006;354:1706-1717.

21. Pfeffer MA, Jarcho JA. The Charisma of Subgroups and the Subgroups of CHARISMA. New England Journal of Medicine 2006;354:1744-1746.

22. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. New England Journal of Medicine 2007;357:2189-2194.

23. Stone GW, Rizvi A, Newman W, et al. Everolimus-Eluting versus Paclitaxel-Eluting Stents in Coronary Artery Disease. New England Journal of Medicine 2010;362:1663-1674.

24. Stone GW, Kedhi E, Kereiakes DJ, et al. Differential Clinical Responses to Everolimus-Eluting and Paclitaxel-Eluting Coronary Stents in Patients With and Without Diabetes Mellitus. Circulation 2011;124:893-900.

25. Swedberg K, Komajda M, Böhm M, et al. Ivabradine and outcomes in chronic heart failure (SHIFT): a randomised placebo-controlled study. The Lancet 2010;376:875-885.

26. Pocock SJ, Lubsen J. More on Subgroup Analyses in Clinical Trials. New England Journal of Medicine 2008;358:2076-2077.

27. Collier TJ, Pocock SJ, McMurray JJV, et al. The impact of eplerenone at different levels of risk in patients with systolic heart failure and mild symptoms: insight from a novel risk score for prognosis derived from the EMPHASIS-HF trial. Eur Heart J 2013;34:2823-2829.

28. Fox KAA, Clayton TC, Damman P, et al. Long-Term Outcome of a Routine Versus Selective Invasive Strategy in Patients With Non–ST-Segment Elevation Acute Coronary Syndrome: A Meta-Analysis of Individual Patient Data. Journal of the American College of Cardiology 2010;55:2435-2445.

29. Wiviott SD, Braunwald E, McCabe CH, et al. Prasugrel versus Clopidogrel in Patients with Acute Coronary Syndromes. New England Journal of Medicine 2007;357:2001-2015.

30. Salisbury AC, Wang K, Cohen DJ, Li Y, Jones PG, Spertus JA. Selecting Antiplatelet Therapy at the Time of Percutaneous Intervention for an Acute Coronary Syndrome: Weighing the Benefits and Risks of Prasugrel Versus Clopidogrel. Circulation: Cardiovascular Quality and Outcomes 2013;6:27-34.

31. Connolly SJ, Ezekowitz MD, Yusuf S, et al. Dabigatran versus Warfarin in Patients with Atrial Fibrillation. New England Journal of Medicine 2009;361:1139-1151.

32. Giugliano RP, Ruff CT, Braunwald E, et al. Edoxaban versus Warfarin in Patients with Atrial Fibrillation. New England Journal of Medicine 2013;369:2093-2104.

33. Pocock SJ, Stone GW, Mehran R, Clayton TC. Individualizing treatment choices using quantitative methods. American Heart Journal 2014;168:607-610.

34. Mega JL, Braunwald E, Wiviott SD, et al. Rivaroxaban in Patients with a Recent Acute Coronary Syndrome. New England Journal of Medicine 2012;366:9-19.

35. Krantz MJ, Kaul S. The ATLAS ACS 2–TIMI 51 Trial and the Burden of Missing Data: (Anti-Xa Therapy to Lower Cardiovascular Events in Addition to Standard Therapy in Subjects With Acute Coronary Syndrome ACS 2–Thrombolysis In Myocardial Infarction 51). Journal of the American College of Cardiology 2013;62:777-781.

36. Montedori A, Bonacini MI, Casazza G, et al. Modified versus standard intention-to-treat reporting: are there differences in methodological quality, sponsorship, and findings in randomized trials? A cross-sectional study. Trials 2011;12:58.

37. Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular Events Associated with Rofecoxib in a Colorectal Adenoma Chemoprevention Trial. New England Journal of Medicine 2005;352:1092-1102.

38. Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. The Lancet 2008;372:1756-1764.

39. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus Warfarin in Nonvalvular Atrial Fibrillation. New England Journal of Medicine 2011;365:883-891.

40. Smith CR, Leon MB, Mack MJ, et al. Transcatheter versus Surgical Aortic-Valve Replacement in High-Risk Patients. New England Journal of Medicine 2011;364:2187-2198.

41. Rossebø AB, Pedersen TR, Boman K, et al. Intensive Lipid Lowering with Simvastatin and Ezetimibe in Aortic Stenosis. New England Journal of Medicine 2008;359:1343-1356.

42. Peto R, Emberson J, Landray M, et al. Analyses of Cancer Data from Three Ezetimibe Trials. New England Journal of Medicine 2008;359:1357-1366.

43. Nissen SE, Wolski K. Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes. New England Journal of Medicine 2007;356:2457-2471.

44. Home PD, Pocock SJ, Beck-Nielsen H, et al. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. The Lancet 2009;373:2125-2135.

45. Bombardier C, Laine L, Reicin A, et al. Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis. New England Journal of Medicine 2000;343:1520-1528.

46. Kearney PM, Baigent C, Godwin J, Halls H, Emberson JR, Patrono C. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? Meta-analysis of randomised trials. BMJ 2006;332:1302-1308.

47. Tepel M, van der Giet M, Schwarzfeld C, Laufer U, Liermann D, Zidek W. Prevention of Radiographic-Contrast-Agent–Induced Reductions in Renal Function by Acetylcysteine. New England Journal of Medicine 2000;343:180-184.

48. Poldermans D, Boersma E, Bax JJ, et al. The Effect of Bisoprolol on Perioperative Mortality and Myocardial Infarction in High-Risk Patients Undergoing Vascular Surgery. New England Journal of Medicine 1999;341:1789-1794.

49. Wald DS, Morris JK, Wald NJ, et al. Randomized Trial of Preventive Angioplasty in Myocardial Infarction. New England Journal of Medicine 2013;369:1115-1123.

50. Population Health Research Institute; Hamilton C. Complete vs Culprit-only Revascularization to Treat Multi-vessel Disease After Primary PCI for STEMI (COMPLETE) In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2012- [cited 2015 July 16]. Available from: https://clinicaltrials.gov/ct2/show/study/NCT01740479 Identifier: NCT01740479 2012.

51. 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.

52. Hopewell S, Clarke M, Moher D, et al. CONSORT for reporting randomised trials in journal and conference abstracts. The Lancet 2008;371:281-283.

Table 1 Efficacy Endpoints for the PEGASUS trial

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Endpoint** | **Ticagrelor,** **90 mg****(N=7050)** | **Ticagrelor,****60 mg****(N=7045)** | **Placebo****(N=7067)** | **Ticagrelor, 90 mg****vs. Placebo** | **Ticagrelor, 60 mg****vs. Placebo** |
| Number (%\*) | Hazard Ratio(95% CI) | P Value | Hazard Ratio(95% CI) | P Value |
| Cardiovascular death, myocardialinfarction, or stroke† | 493 | (7.85) | 487 | (7.77) | 578 | (9.04) | 0.85 | (0.75–0.96) | 0.008 | 0.84 | (0.74–0.95) | 0.004 |
| Death from coronary heart disease, myocardial infarction, or stroke | 438 | (6.99) | 445 | (7.09) | 535 | (8.33) | 0.82 | (0.72–0.93) | 0.002 | 0.83 | (0.73–0.94) | 0.003 |
| Cardiovascular death or myocardialinfarction | 424 | (6.79) | 422 | (6.77) | 497 | (7.81) | 0.85 | (0.75–0.97) | 0.01 | 0.85 | (0.74–0.96) | 0.01 |
| Death from coronary heart disease or myocardial infarction | 350 | (5.59) | 360 | (5.75) | 429 | (6.68) | 0.81 | (0.71–0.94) | 0.004 | 0.84 | (0.73–0.96) | 0.01 |
| Cardiovascular death | 182 | (2.94) | 174 | (2.86) | 210 | (3.39) | 0.87 | (0.71–1.06) | 0.15 | 0.83 | (0.68–1.01) | 0.07 |
| Death from coronary heart disease | 97 | (1.53) | 106 | (1.72) | 132 | (2.08) | 0.73 | (0.56–0.95) | 0.02 | 0.80 | (0.62–1.04) | 0.09 |
| Myocardial infarction | 275 | (4.40) | 285 | (4.53) | 338 | (5.25) | 0.81 | (0.69–0.95) | 0.01 | 0.84 | (0.72–0.98) | 0.03 |
| Stroke |  |  |  |  |  |  |  |  |  |  |  |  |
| Any | 100 | (1.61) | 91 | (1.47) | 122 | (1.94) | 0.82 | (0.63–1.07) | 0.14 | 0.75 | (0.57–0.98) | 0.03 |
| Ischemic | 88 | (1.41) | 78 | (1.28) | 103 | (1.65) | 0.85 | (0.64–1.14) | 0.28 | 0.76 | (0.56–1.02) | 0.06 |
| Death from any cause | 326 | (5.15) | 289 | (4.69) | 326 | (5.16) | 1.00 | (0.86–1.16) | 0.99 | 0.89 | (0.76–1.04) | 0.14 |
| \* percentages are 3-year Kaplan Meier estimates † primary endpointSource: Bonaca MP, Bhatt DL, Cohen M, et al. Long-Term Use of Ticagrelor in Patients with Prior Myocardial Infarction. New England Journal of Medicine 2015;372:1791-1800. |

Table 1 Caption: Number of events, 3-year Kaplan-Meier estimates and hazard ratios for efficacy endpoints in the PEGASUS trial.

Table 2 A summary of key 1 and 5 year findings from the SYNTAX trial

|  |  |
| --- | --- |
|  | Event Rates |
| Endpoint | 1 year | 5 years |
|  | CABG(N=897) | DES(N=903) | P-value | CABG(N=897) | DES(N=903) | P-value |
| MACCE composite\* | 12.1% | 17.8% | 0.002 | 26.9% | 37.3% | <0.0001 |
| Death | 3.5% | 4.4% | 0.37 | 11.4% | 13.9% | 0.10 |
| MI | 3.3% | 4.8% | 0.11 | 3.8% | 9.7% | <0.0001 |
| Stroke | 2.2% | 0.6% | 0.003 | 2.4% | 3.7% | 0.09 |
| Death/MI/stroke | 7.6% | 7.5% | 0.98 | 16.7% | 20.8% | 0.03 |
| Repeat revascularisation | 5.9% | 13.7% | <0.001 | 13.7% | 25.9% | <0.0001 |
| PCI | 4.7% | 11.4% | <0.001 |  |  |  |
| CABG | 1.3% | 2.8% | 0.03 |  |  |  |
| \* MACCE is the pre-defined primary composite of death, MI, stroke and repeat revascularisation |

Table 2 Caption: Percentage of patients experiencing the composite primary endpoint (MACCE) and its components at 1 year and 5 years in the SYNTAX trial

Table 3 Efficacy results from the EMPHASIS-HF trial, with and without covariate adjustment

|  |
| --- |
|  |
| Outcome | Adjusted | Unadjusted |
| Hazard Ratio (95% CI) | P-value | Hazard Ratio (95% CI) | P value |
| Primary endpoint\* | 0.63 | (0.54-0.74) | <0.001 | 0.66 | (0.56-0.78) | <0.001 |
| CV death | 0.76 | (0.61-0.94) | 0.01 | 0.77 | (0.62-0.96) | 0.02 |
| Hospitalisation for heart failure | 0.58 | (0.47-0.70) | <0.001 | 0.61 | (0.50-0.75) | <0.001 |
| \*Primary endpoint is the composite of CV death and hospitalisation for heart failureSource: Zannad F, McMurray JJV, Krum H, et al. Eplerenone in Patients with Systolic Heart Failure and Mild Symptoms. New England Journal of Medicine 2011;364:11-21. |

Table 3 Caption: Hazard ratios and 95% confidence intervals with and without baseline covariate adjustment for efficacy endpoints in the EMPHASIS-HF trial. Adjustment was made using a proportional hazards model adjusting for 13 pre-defined baseline covariates: age, estimated GFR, ejection fraction, bodymass index, hemoglobin value, heart rate, systolic blood pressure, diabetes mellitus, history of hypertension, previous myocardial infarction, atrial fibrillation, and left bundle-branch block or QRS duration greater than 130 msec.

Table 4: Primary endpoint event rates by risk group and treatment in the EMPHASIS-HF trial

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Risk Group** | **Treatment****Group** | **Nos.****Patients** | **Nos.****Events** | **Rate\*** | **Hazard Ratio****(95% CI)** | **Rate Difference (95% CI)** |
| Low | *Placebo* | 643 | 103 | 7.61 | 0.74(0.56, 0.99) | -1.98(-3.89, -0.06) |
| *Eplerenone* | 648 | 81 | 5.63 |
| Mid | *Placebo* | 478 | 164 | 19.00 | 0.64 (0.50, 0.82) | -6.80(-10.54, -3.06) |
| *Eplerenone* | 445 | 104 | 12.20 |
| High | *Placebo* | 252 | 125 | 39.42 | 0.63(0.49, 0.82) | -15.22(-23.57, -6.88) |
| *Eplerenone* | 271 | 103 | 24.19 |
| \* Rate per 100 person years; Primary endpoint is composite of CV death and heart failure hospitalisation. Source: Collier TJ, Pocock SJ, McMurray JJV, et al. The impact of eplerenone at different levels of risk in patients with systolic heart failure and mild symptoms: insight from a novel risk score for prognosis derived from the EMPHASIS-HF trial. Eur Heart J 2013;34:2823-2829. |

Table 4 Caption: Event rates, hazard ratios and rate differences for the primary endpoint (CV death and hospitalisation for heart failure) by risk group in the EHPHASIS-HF trial.

Table 5 CONSORT checklist of items to include when reporting a randomised trial

|  |  |  |
| --- | --- | --- |
| Section/Topic | Item Number | Checklist item |
| Title and abstract |  |  |
|  | 1a | Identification as a randomised trial in the title |
| 1b | Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts) |
| **Introduction** |  |  |
| Background and objectives | 2a | Scientific background and explanation of rationale |
| 2b | Specific objectives or hypotheses |
| **Methods** |  |  |
| Trial design | 3a | Description of trial design (such as parallel, factorial) including allocation ratio |
| 3b | Important changes to methods after trial commencement (such as eligibility criteria), with reasons |
| Participants | 4a | Eligibility criteria for participants |
| 4b | Settings and locations where the data were collected |
| Interventions | 5 | The interventions for each group with sufficient details to allow replication, including how and when they were actually administered |
| Outcomes | 6a | Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed |
| 6b | Any changes to trial outcomes after the trial commenced, with reasons |
| Sample size | 7a | How sample size was determined |
| 7b | When applicable, explanation of any interim analyses and stopping guidelines |
| **Randomisation** |  |  |
| Sequence generation | 8a | Method used to generate the random allocation sequence |
| 8b | Type of randomisation; details of any restriction (such as blocking and block size) |
| Allocation concealment mechanism | 9 | Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned |
| Implementation | 10 | Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions |
| Blinding | 11a | If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how |
| 11b | If relevant, description of the similarity of interventions |
| Statistical methods | 12a | Statistical methods used to compare groups for primary and secondary outcomes |
| 12b | Methods for additional analyses, such as subgroup analyses and adjusted analyses |
| **Results** |  |  |
| Participant flow (a diagram is strongly recommended) | 13a | For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome |
| 13b | For each group, losses and exclusions after randomisation, together with reasons |
| Recruitment | 14a | Dates defining the periods of recruitment and follow-up |
| 14b | Why the trial ended or was stopped |
| Baseline data | 15 | A table showing baseline demographic and clinical characteristics for each group |
| Numbers analysed | 16 | For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups |
| Outcomes and estimation | 17a | For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval) |
| 17b | For binary outcomes, presentation of both absolute and relative effect sizes is recommended |
| Ancillary analyses | 18 | Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory |
| Harms | 19 | All important harms or unintended effects in each group (for specific guidance see CONSORT for harms) |
| **Discussion** |  |  |
| Limitations | 20 | Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses |
| Generalisability | 21 | Generalisability (external validity, applicability) of the trial findings |
| Interpretation | 22 | Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence |
| Other information |
| Registration | 23 | Registration number and name of trial registry |
| Protocol | 24 | Where the full trial protocol can be accessed, if available |
| Funding | 25 | Sources of funding and other support (such as supply of drugs), role of funders |

Source: 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.

Table 5 Caption: CONSORT checklist of items to include when reporting a parallel group randomised trial.

Figure 1 Assessing the evidence for excess heart failure across 3 related placebo-controlled trials in Type 2 diabetes.



Figure 1 Caption: Forest plot showing study specific and pooled (fixed effect and random effects) hazard ratio for heart failure from the SAVOR, EXAMINE and TECOS trials.

Figure 2 Pre-specified subgroup analyses in the PARADIGM-HF trial.



Source: McMurray JJV, Packer M, Desai AS, et al. Angiotensin–Neprilysin Inhibition versus Enalapril in Heart Failure. New England Journal of Medicine 2014;371:993-1004.

Figure 2 Caption: Hazard ratios for the primary end point (death from cardiovascular causes or first hospitalization for heart failure) and for death from cardiovascular causes among patients in prespecified subgroups. The size of the square corresponds to the number of patients in each subgroup.

Figure 3 Excess risk of thrombotic events on rofecoxib in the APPROVe trial first as on-treatment analysis and subsequently as ITT analysis



Source: Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular Events Associated with Rofecoxib in a Colorectal Adenoma Chemoprevention Trial. New England Journal of Medicine 2005;352:1092-1102 and Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. The Lancet 2008;372:1756-1764.

Figure 3 Caption: Kaplan-Meier cumulative frequency of thrombotic events in the APPROVe trial calculated from an (a) on-treatment analysis and (b) ITT analysis.

Central Illustration of Topics Covered

|  |  |
| --- | --- |
| Multiplicity of Data | * need a pre-defined Statistical Analysis Plan
* give priority to primary endpoint
* balanced account of efficacy and safety
* composite endpoints: interpret wisely
 |
| Covariate Adjustment  | * adjust for variables affecting prognosis
* pre-define variables and model chosen
* slight gain in statistical power
* worth considering as primary analysis
 |
| Subgroup Analysis | * focus on pre-selected subgroups
* analyse with interaction tests
* interpret as exploratory, secondary findings
* avoid over-reacting to subgroup claims
 |
| Individual Benefits and Risks | * assess individual efficacy on absolute scale
* classify patients into risk groups
* absolute efficacy versus absolute harm
* statistical models for individual treatment choice
 |
| ITT Analysis and Alternatives | * prioritize analysis by intention to treat (ITT)
* on-treatment analyses usually secondary
* report both for non-inferiority trials
* avoid poor compliance and loss to follow-up
 |
| Interpreting Surprises | * beware of small trials reporting big effects
* anticipate regression to the truth
* seek evidence to confirm (or not) asap
* alarmist safety signals require careful handling
 |