JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2015 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER INC.

VOL. 66, NO. 23, 2015
ISSN 0735-1097/$36.00
http://dx.doi.org/10.1016/j.jacc.2015.10.023

THE PRESENT AND FUTURE

STATE-OF-THE-ART REVIEW

Statistical Controversies in Reporting of Clinical Trials

Part 2 of a 4-Part Series on Statistics for Clinical Trials

Stuart J. Pocock, PhD,* John J.V. McMurray, MD,† Tim J. Collier, MSc*

ABSTRACT

This paper 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 steer a wise course in their statistical reporting, thereby giving readers a balanced account of trial findings. (J Am Coll Cardiol 2015;66:2648–62)
© 2015 by the American College of Cardiology Foundation.

Last week's review paper covered the fundamentals of statistical analysis and reporting of randomized 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 investigators due to 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 an individual patient's absolute benefits and absolute risks. The merit of analysis by intention to treat (ITT) 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, which are often quite small in scale.

All of these controversies are summarized in the **Central Illustration** and are 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). So, out of the potential chaos of 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

Listen to this manuscript's audio summary by *JACC* Editor-in-Chief Dr. Valentin Fuster.

From the *Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, United Kingdom; and the †Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, United Kingdom. The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

Manuscript received September 8, 2015; revised manuscript received October 12, 2015, accepted October 19, 2015.

condensed trial report is fair in what it includes; that is, how do we resist the temptation to "play up the positive" by devoting more space in the results and conclusions sections 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) that is fully signed off before database locking 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 (CI), and p value. For time-to-event outcomes this is commonly a hazard ratio (HR) (and 95% CI) with log-rank 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; for example, as in **Table 1** for the PEGASUS-TIMI 54 (Prevention of Cardiovascular Events in Patients With Prior Heart Attack Using Ticagrelor Compared to Placebo on a Background of Aspirin–Thrombolysis In Myocardial Infarction 54) trial (2), comparing 2 doses of ticagrelor with aspirin in patients with prior myocardial infarction (MI). In this instance, the interpretation appears to be straightforward because 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 dyspnea 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 (Prospective pioglitazone clinical trial in macrovascular events) (3) trial of pioglitazone versus placebo in 5,238 diabetic patients had a primary composite endpoint of death, MI, stroke, acute coronary syndrome, endovascular surgery, or leg amputation. Over a mean 3 years of follow-up, the HR was 0.90 (95% CI: 0.80 to 1.02; p = 0.095). The main secondary endpoint, the composite of death, MI, and stroke, had an HR of 0.84 (95% CI: 0.72 to 0.98; p = 0.027). The publication's conclusions highlighted the latter and downplayed the lack of statistical significance for the primary endpoint, whereas a more cautious interpretation is usually warranted.

In contrast, the publication of the MATRIX (Minimizing Adverse Hemorrhagic Events by Transradial Access Site and Systemic Implementation of Angiox) trial (4), comparing bivalirudin or unfractionated heparin in acute coronary syndromes, had conclusions confined to the coprimary endpoints of major adverse cardiovascular (CV) events (death, MI, or stroke) and net adverse clinical events (death, MI, stroke, or major bleed), both of which "were not significantly lower with bivalirudin than with unfractionated heparin." Whereas the focus on primary endpoints is appropriate, there is a danger that it can hide important differences among secondary (component) endpoints. Although cautious interpretation is essential across a multiplicity of secondary endpoints, the 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 the potential harm of a treatment, controversy is likely to ensue. For instance, in the SAVOR-TIMI 53 (Saxagliptin Assessment of Vascular Outcomes Recorded in Patients with Diabetes Mellitus–Thrombolysis In Myocardial Infarction 53) trial (5) of saxagliptin versus placebo in diabetic patients, the composite primary endpoint (CV death, MI, and stroke) showed no treatment difference, but 1 of several secondary endpoints, heart failure hospitalization, revealed an excess on saxagliptin (HR: 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. The 2 subsequent EXAMINE (Examination of Cardiovascular Outcomes with Alogliptin versus Standard of Care) (6) and TECOS (Trial Evaluating Cardiovascular Outcomes with Sitagliptin) (7) trials of drugs in the same class, alogliptin and sitagliptin, respectively, revealed no excess of heart failure, and there is no plausible

ABBREVIATIONS AND ACRONYMS

CABG = coronary artery bypass graft

CI = confidence interval

DES = drug-eluting stent

ITT = intention to treat

MACCE = major adverse cardiac or cerebrovascular events

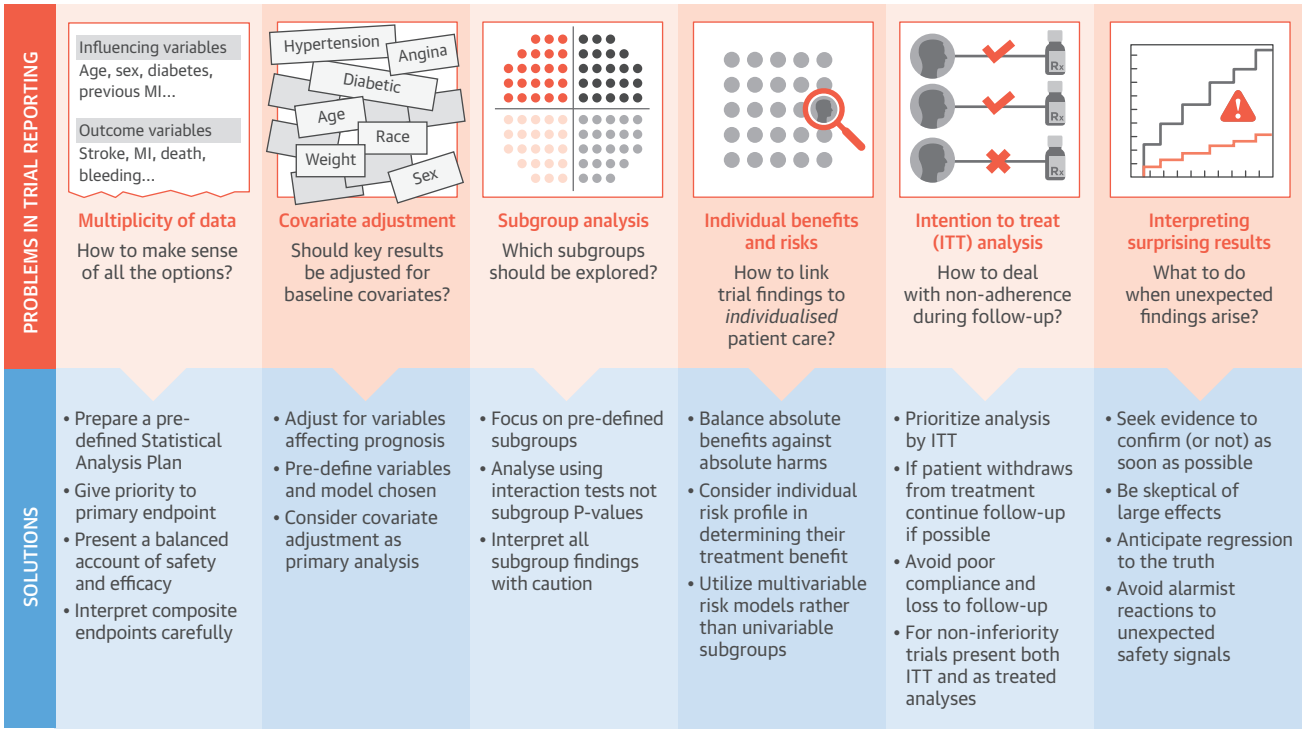
MI = myocardial infarction

PCI = percutaneous coronary intervention

RCT = randomized clinical trial

SAP = statistical analysis plan

CENTRAL ILLUSTRATION **Statistical Controversies in Reporting of Clinical Trials: Problems and Solutions**



Pocock, S.J. et al. J Am Coll Cardiol. 2015; 66(23):2648–62.

ITT = intention to treat; MI = myocardial infarction.

biological explanation as to why the drugs might differ in this respect. Furthermore, a statistical test of heterogeneity comparing the 3 trials' HRs for heart failure is not statistically significant (interaction p = 0.16), and the combined HR is 1.13 (p = 0.04) and 1.12 (p = 0.18) for fixed and random-effect meta-analyses, respectively (Figure 1). This analysis partly hinges on the concept that similar effects should be expected

TABLE 1 **Efficacy Endpoints for the PEGASUS-TIMI 54 Trial**

	Ticagrelor, 90 mg (n = 7,050)	Ticagrelor, 60 mg (n = 7,045)	Placebo (n = 7,067)	Ticagrelor, 90 mg vs. Placebo		Ticagrelor, 60 mg vs. Placebo	
				HR (95% CI)	p Value	HR (95% CI)	p Value
Cardiovascular death, MI, 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, MI, 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 MI	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 MI	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
MI	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

Values are n (%) unless otherwise indicated. Number of events, 3-year Kaplan-Meier estimates, and hazard ratios for efficacy endpoints in the PEGASUS-TIMI 54 (Prevention of Cardiovascular Events in Patients With Prior Heart Attack Using Ticagrelor Compared to Placebo on a Background of Aspirin—Thrombolysis In Myocardial Infarction 54) trial. Data from Bonaca et al. (2). *Primary endpoint.

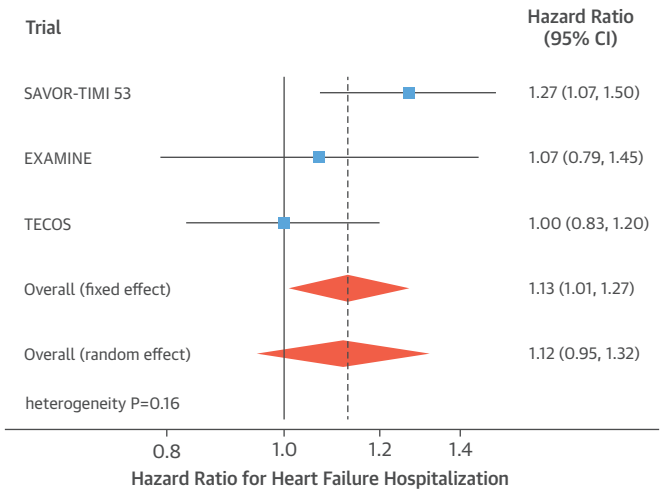
CI = confidence interval; HR = hazard ratio; MI = myocardial infarction.

for drugs in the same class. This is often the case, but there are exceptions: for example, torcetrapib versus other cholesteryl ester transfer protein inhibitors, and ximelagatran versus other direct thrombin inhibitors regarding liver abnormalities. Thus, although 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 published medical data (usually with less emphasis than authors might wish), although 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-TIMI 53 trial (5). There is an asymmetry here in that the corresponding extent of evidence in the direction of treatment benefit would receive scant attention. Although 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 on the basis of weak evidence of potential harm arising from data dredging across a multiplicity of endpoints.

COMPOSITE ENDPOINTS. These are commonly used in CV RCTs to combine evidence across 2 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 (Synergy between Percutaneous Coronary Intervention with Taxus and Cardiac Surgery) trial (8,9) of bypass surgery (CABG) versus the TAXUS drug-eluting stent (DES) in 1,800 patients with left main or triple-vessel disease had a major adverse cardiac or cerebrovascular events (MACCE) composite primary endpoint comprising death, stroke, MI, and repeat revascularization; results at 1 and 5 years of follow-up are shown in **Table 2**. At 1 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 revascularization, of which the great majority is repeat PCIs. 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 1 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 (for example, complete or culprit lesion intervention in primary PCI) is that clinically driven interventions should not be part of the primary endpoint.

FIGURE 1 Assessing the Evidence for Excess Heart Failure Across 3 Related Placebo-Controlled Trials in Type 2 Diabetes



Forest plot showing study-specific and pooled (fixed effect and random effects) hazard ratios for heart failure from the SAVOR-TIMI 53, EXAMINE, and TECOS trials. CI = confidence interval; EXAMINE = Examination of Cardiovascular Outcomes with Alogliptin versus Standard of Care; SAVOR-TIMI 53 = Saxagliptin Assessment of Vascular Outcomes Recorded in Patients with Diabetes Mellitus–Thrombolysis In Myocardial Infarction 53; TECOS = Trial Evaluating Cardiovascular Outcomes with Sitagliptin.

TABLE 2 A Summary of Key 1- and 5-Year Findings From the SYNTAX Trial

	1-Year Event Rates			5-Year Event Rates		
Endpoint	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 revascularization	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			

Values are % of patients experiencing the composite primary endpoint (MACCE) and its components at 1 and 5 years in the SYNTAX trial. *MACCE is the pre-defined primary composite of death, MI, stroke, and repeat revascularization.

CABG = coronary artery bypass graft; DES = drug-eluting stents; MACCE = major adverse cardiac or cerebrovascular events; MI = myocardial infarction; PCI = percutaneous coronary intervention; SYNTAX = Synergy between Percutaneous Coronary Intervention with Taxus and Cardiac Surgery.

TABLE 3 Efficacy Results From the EMPHASIS-HF Trial, With and Without Covariate Adjustment

	Adjusted HR (95% CI)	p Value	Unadjusted HR (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
Hospitalization 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 hospitalization for heart failure. HRs and 95% CIs 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 glomerular filtration rate, ejection fraction, body mass index, hemoglobin value, heart rate, systolic blood pressure, diabetes mellitus, history of hypertension, previous MI, atrial fibrillation, left bundle-branch block, or QRS duration >130 ms. Data from Zannad et al. (12).

CV = cardiovascular; EMPHASIS-HF = Eplerenone in Mild Patients Hospitalization and Survival Study in Heart Failure; other abbreviations as in Table 1.

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 (Evaluation of XIENCE versus coronary artery bypass surgery for effectiveness of left main revascularization) (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 CV 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)? Practice varies widely: for some RCTs only unadjusted results are presented, others have covariate adjustment as their primary analysis, and yet others use it as a secondary sensitivity analysis. This inconsistency of approach across trials is, perhaps, tolerated because in major trials, randomization ensures good balance across treatments for baseline variables, and hence, covariate adjustment usually makes little difference.

The EMPHASIS-HF (Eplerenone in Mild Patients Hospitalization and Survival Study in Heart Failure) trial (12) of eplerenone versus placebo in 2,737 chronic heart failure patients illustrates the consequences of covariate adjustment. The investigators pre-defined use of a proportional hazards model adjusting for 13 baseline covariates: age, estimated glomerular filtration rate (GFR), ejection fraction, body mass index, hemoglobin value, heart rate, systolic blood pressure, diabetes mellitus, history of hypertension, previous MI, atrial fibrillation, left bundle-branch block, and QRS duration >130 ms. Selection was sensibly on the basis of prior knowledge/suspicion of each variable's association with patient prognosis. Table 3 shows the adjusted and unadjusted HRs for eplerenone versus placebo for the primary endpoint and also for its 2 separate components. In all 3 instances, the adjusted HR was slightly further from 1, as one would expect when adjusting for factors that are related to prognosis (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 (CI width changes little); rather, point estimates, (e.g., odds ratio, HR) 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 randomization. This was specified in IMPROVE-IT (Improved Reduction of Outcomes: Vytorin Efficacy International Trial) (14) in acute coronary syndrome, where stratification factors were prior lipid-lowering therapy, type of acute coronary syndrome, and enrollment 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, although harmless, might be considered of little value.

Adjustment for geographic region is also sometimes performed. For example, PARADIGM-HF (Prospective Comparison of ARNI [Angiotensin Receptor-Neprilysin Inhibitor] with ACEI [Angiotensin-Converting-Enzyme Inhibitor] to Determine Impact on Global Mortality and Morbidity in Heart Failure) (15) adjusted HRs for 5 regions, 1 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 (Gruppo Italiano per lo Studio della Sopravvivenza nell'Insufficienza cardiaca-Heart Failure) trial (16) adjusted for variables that were unbalanced between randomized groups at $p < 0.1$. As a secondary sensitivity analysis, it can add reassurance that the primary analysis makes sense. However, if the covariates with imbalance again do not affect the prognosis, such adjustment will make a 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 (Candesartan in Heart failure: Assessment of Reduction in Mortality and morbidity) trial (17) in 7,599 heart failure patients, the unadjusted HR (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 an HR of 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 the covariate-adjusted HR for CV death was 0.87 (95% CI: 0.78 to 0.96; $p = 0.006$).

In general, we believe 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. On the basis of 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 nonsignificant predictors, or adding in new significant predictors after database lock) should be avoided in the primary analysis because 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 supported by experience of what influences prognosis, then it makes sense to have the covariate-adjusted analysis as primary (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.

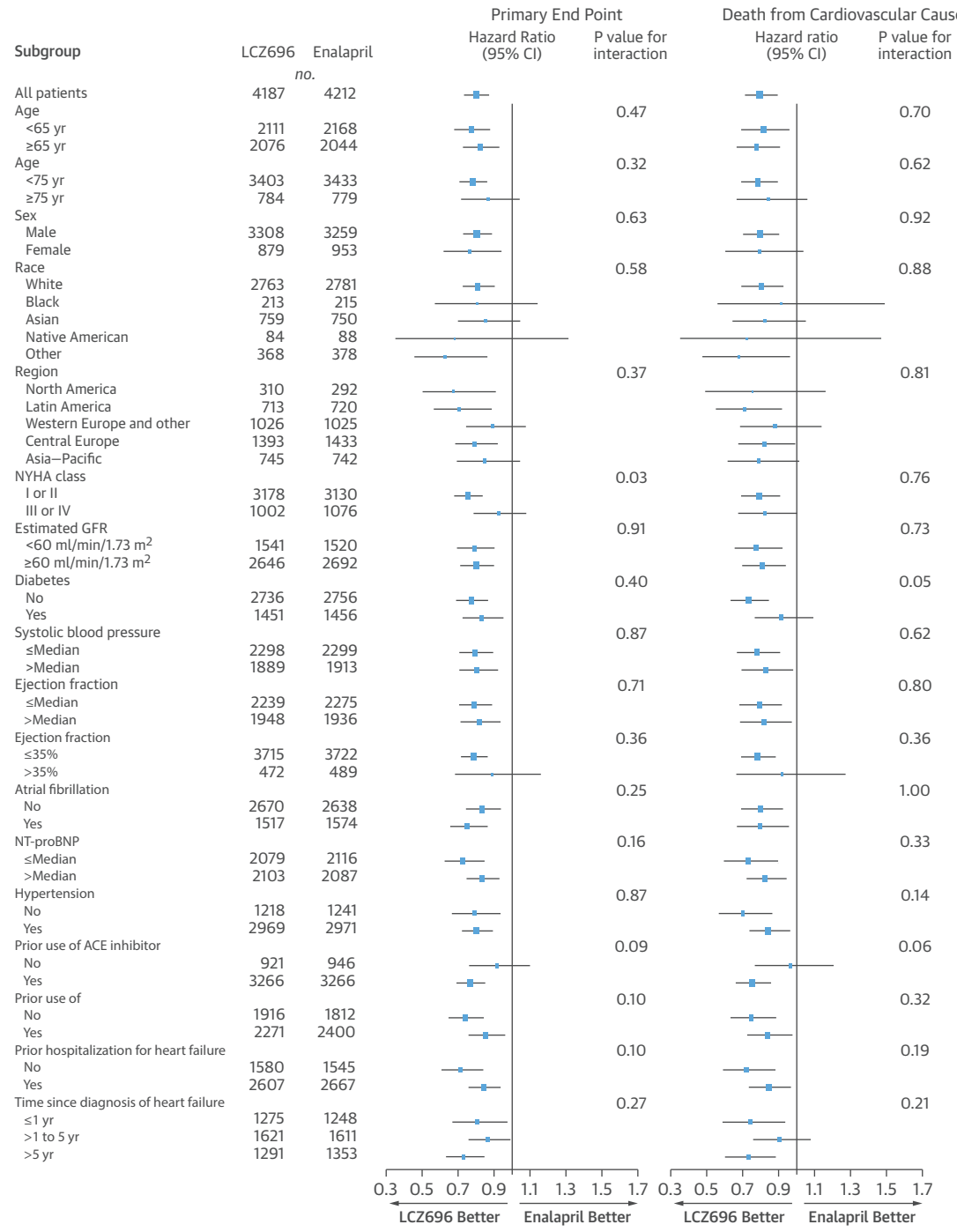
Of all multiplicity problems in reporting RCTs, interpretation of subgroup analyses presents a particular challenge (19). First, trials usually lack power to reliably detect subgroup effects. Second, there are many possible subgroups that could be explored, and one needs to guard against data dredging, 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) 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 (an HR 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 with enalapril in this heterogeneous heart failure population for both the primary endpoint and CV death. For reference, the results for all patients, with their inevitably tighter CIs, are shown at the top of **Figure 2**.

Scanning across subgroups, one can see that estimated HRs vary by chance and CIs are wider for smaller subgroups. Some CIs overlap the line of unity, indicating that the subgroup p value does not reach 5% significance; 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 HRs across subgroups may be attributed to chance variation. For the primary endpoint, just 1 interaction test is statistically significant: $p = 0.03$ for New York Heart Association class I or II versus III or IV, which suggests a possible greater benefit of LCZ696 in less symptomatic patients, although no such interaction exists for CV death (interaction $p = 0.76$). Given that 18 subgroup analyses have been performed for each of 2 outcomes, one could expect at least 1 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

FIGURE 2 Pre-Specified Subgroup Analyses in the PARADIGM-HF Trial



Hazard ratios for the primary endpoint (death from cardiovascular causes or first hospitalization for heart failure) and for death from cardiovascular causes among patients in pre-specified subgroups. The **size of the square** corresponds to the number of patients in each subgroup. Data from McMurray et al. (15). ACE = angiotensin-converting enzyme; CI = confidence interval; GFR = glomerular filtration rate; NT-proBNP = N-terminal pro-B-type natriuretic peptide; NYHA = New York Heart Association; PARADIGM-HF = Prospective Comparison of ARNI (Angiotensin Receptor-Neprilysin Inhibitor) with ACEI (Angiotensin-Converting-Enzyme Inhibitor) to Determine Impact on Global Mortality and Morbidity in Heart Failure.

there is a particular subgroup in which the treatment effect is favorable. In this context, subgroup claims require an especially cautious interpretation in a journal publication. Furthermore, it is highly unlikely that regulators, such as the Food and Drug Administration (FDA), would approve a drug on the basis of such a positive subgroup claim.

The CHARISMA (Clopidogrel For High Atherothrombotic Risk, Ischemic Stabilization, Management, And Avoidance) trial (20) is an interesting case in point. Against a background of low-dose aspirin, 15,603 patients at high risk of atherothrombotic events were randomized 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% versus 7.9% ($p = 0.046$). In contrast, the results trended in the opposition direction in asymptomatic patients: 6.6% versus 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 that "the charisma of extracting favorable subgroups should be resisted" (21). Why? Well, this was 1 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; that is, so-called qualitative interactions rarely arise across clinical medicine. The *New England Journal of Medicine* has subsequently toughened its policy regarding subgroup analyses (22), so that they are seen more as exploratory and hypothesis-generating, rather than as 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 (Clinical Evaluation of the XIENCE V Everolimus Eluting Coronary Stent System) IV (23) trial comparing everolimus-eluting stents (EES) or paclitaxel-eluting stents in 3,687 patients showed overall superiority for EES for the primary endpoint, target lesion failure at 1 year: 4.2% versus 6.8% ($p = 0.001$). But, in 1,140 diabetic patients (1 of 12 subgroup analyses), there was no evidence of a treatment difference: 6.4% versus 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 2-year outcomes across 4 RCTs of EES versus paclitaxel-eluting stents in 6,780 patients (24) revealed marked superiority of EES in nondiabetic subjects for death, MI, stent thrombosis, and ischemia-driven target lesion revascularization, whereas no such benefits of EES existed for diabetic patients. All 4 interaction tests were convincingly significant: $p = 0.02$, $p = 0.01$, $p = 0.0006$, and $p = 0.02$, respectively.

Although such confirmatory evidence of an initial subgroup finding is highly desirable, it is not always achievable. But, regulatory decisions still need to be made on whether apparent lack of efficacy in a subgroup merits a specific restriction with regard to drug approval. For instance, the European Medicines Agency restricted use of ivabradine in chronic heart failure to patients with heart rate ≥75 beats/min on the basis of a significant interaction in 1 pivotal trial (25).

Overall, there is a responsibility to perform and present subgroup findings from major RCTs. Prespecification of a limited number of intended subgroup analyses is a helpful guard against post-hoc manipulations of data, but interpretations are still restricted by a lack of statistical power and a multiplicity of hypotheses, so due caution is required to not overreact to any subgroup claims. Subgroup analysis becomes most convincing when it relates to just 1 pre-declared factor of especial interest (e.g., troponin-positive vs. 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., on the basis of HRs or odds ratios), there may well be important differences between individuals as regards absolute treatment benefits (26).

For instance, in the EMPHASIS-HF trial (12) of eplerenone versus placebo in heart failure, patients with mild symptoms, the composite primary endpoint, CV death, and heart failure hospitalization, showed a marked benefit over a median 21-month follow-up (HR: 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 groups on the basis of a multivariable risk score using 10 commonly-recognized prognostic features (27). **Table 4** shows the consequent treatment benefits by risk group, on both relative and absolute scales. As anticipated, the HR was similar in all risk groups. But, the absolute benefits varied markedly by

TABLE 4 Primary Endpoint Event Rates by Risk Group and Treatment in the EMPHASIS-HF Trial

Risk Group	Treatment Group	Number of Patients	Number of Events	Rate*	HR (95% CI)	Rate Difference (95% CI)
Low	Placebo	643	103	7.61	0.74 (0.56, 0.99)	−1.98 (−3.89 to 0.06)
	Eplerenone	648	81	5.63		
Mid	Placebo	478	164	19.00	0.64 (0.50, 0.82)	−6.80 (−10.54 to 3.06)
	Eplerenone	445	104	12.20		
High	Placebo	252	125	39.42	0.63 (0.49, 0.82)	−15.22 (−23.57 to 6.88)
	Eplerenone	271	103	24.19		

*Rate per 100 person-years; primary endpoint is the composite of CV death and heart failure hospitalization. Event rates, HRs, and rate differences for the primary endpoint (CV death and hospitalization for heart failure) by risk group in the EMPHASIS-HF trial. Data from Collier et al. (27).

Abbreviations as in Tables 1 and 3.

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.

Pooling of data for 3 trials (28) of routine invasive versus selective invasive strategies in acute coronary syndrome found a significant difference in the 5-year risk of CV death or MI (HR: 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 5-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 2 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 each case. Note that this is achieved by multivariable risk analysis, rather than univariate subgroups. This becomes particularly important if treatment efficacy is offset by a risk of side effects. For instance, TRITON-TIMI 38 (Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition with Prasugrel–Thrombolysis In Myocardial Infarction 38) (29) compared prasugrel with 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% vs. 12.1%; HR: 0.81; 95% CI: 0.73 to 0.90; $p < 0.001$). This benefit was mainly due to a reduction in nonfatal MIs: 7.3% versus 9.5%. However, there were more bleeding events on prasugrel; for example, TIMI (Thrombolysis In Myocardial Infarction) major bleed (2.4% vs. 1.8%; $p = 0.03$).

To weigh the benefits and risks of prasugrel versus clopidogrel on an individual patient basis, Salisbury et al. (30) used multivariable logistic models to separately predict any patient's risk of: 1) the primary ischemic endpoint; and 2) TIMI major or minor bleed, taking account of both randomized treatment and patient characteristics. The intent is to quantify, on an absolute scale, how the tradeoff between treatment differences in ischemic efficacy and bleeding harm is patient-specific. For instance, the bleeding risk is of greater concern in elderly women, whereas in a younger man with known CV risk factors, avoidance of future ischemic events is paramount. Clopidogrel is the drug of choice for the former, whereas prasugrel is a better choice for the latter. These multivariable models are a quantitative aid to such clinical judgment, and may be of particular use in the broader spectrum of patients for whom the efficacy/safety tradeoff 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 (Randomized Evaluation of Long-Term Anticoagulation Therapy) (31) and ENGAGE AF-TIMI 48 (Effective Anticoagulation with Factor Xa Next Generation in Atrial Fibrillation–Thrombolysis In Myocardial Infarction 48) (32) trials are to be commended for comparing 2 different anticoagulant drug doses (of dabigatran and edoxaban, respectively) against warfarin. In both 3-arm trials, the higher dose appeared to be more effective in reducing the risk of stroke and systemic embolism, whereas the lower dose had fewer bleeding events. At present, these 2 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 patients (if any) would

have a net benefit from the lower dose. Ideally, such risk models should be on the basis of 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 of 1 baseline variable at a time. Therefore, the opportunity to link trial findings to individualized patient care on the basis of "whole patient" risk profiles is largely being missed (33).

ANALYSIS BY ITT AND OTHER OPTIONS

Analysis by ITT means that a trial's results include the totality of patient follow-up for all randomized 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 or for how long. Everyone is included, with no escape! Such logic appears soundly based, but there are 2 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?

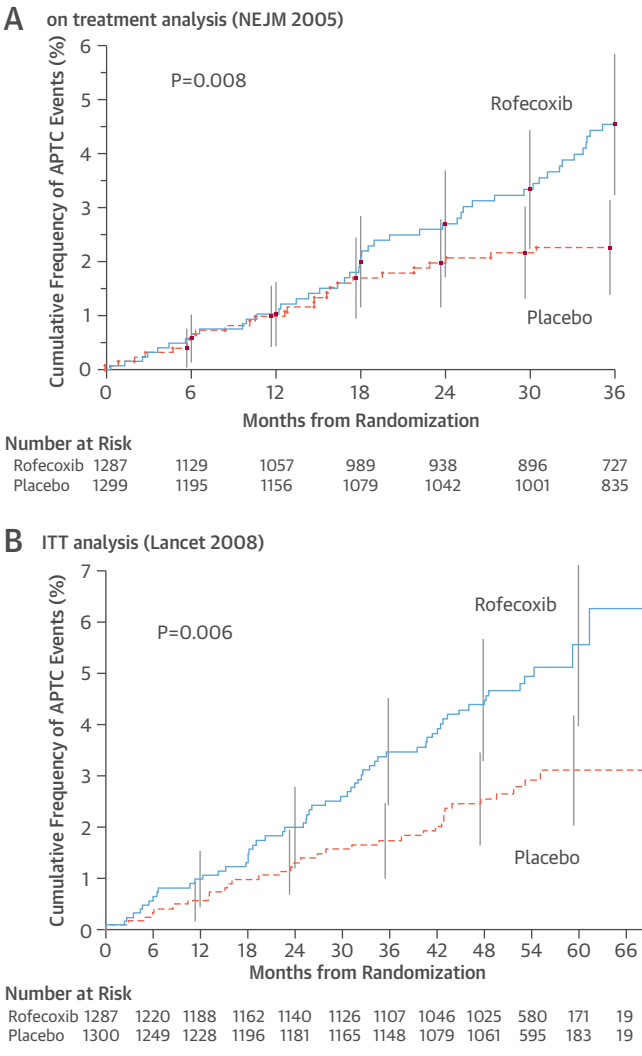
On the first point, the more patients that are lost to follow-up, the further the attempted analysis deviates from true ITT. In trial conduct, it is important to minimize loss to follow-up. High treatment compliance is a first step. Also, when patients do withdraw from treatment, their follow-up should 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 (among others) censoring is sensibly assumed to be noninformative. But, if patients are lost to follow-up at an earlier stage, this cannot be assumed to occur at random. For example, patients who drop out may be sicker and hence at higher risk of a primary event, which goes unrecorded. Thus, loss to follow-up is potentially informative censoring and could lead to a biased treatment comparison. This becomes particularly serious if the dropout rates, and their reasons, differ between treatment groups.

The ATLAS ACS 2 (Anti-Xa Therapy to Lower Cardiovascular Events in Addition to Standard Therapy in Subjects With Acute Coronary Syndrome ACS 2) trial (34) illustrates this problem. This 3-arm trial compared 2 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, although 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 the trial's end. This problem cast sufficient doubt on the robustness of the trial findings to influence the FDA panel not to recommend approval. There is no fixed guidance as to what level of dropout becomes unacceptable. ATLAS ACS 2 clearly did less well in this regard than several other recent trials in acute coronary syndrome, although, of course, none could achieve 100% follow-up (35).

Although presenting ITT analyses, ATLAS ACS 2 (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 the prime focus 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-randomization exclusions, because these could lead to bias (36). More acceptable modifications are exclusions of ineligible patients incorrectly randomized and, in double-blind trials, exclusions of any patients who never got a single dose of the study drug.

The APPROVe (Adenomatous Polyp Prevention on Vioxx) trial (37,38) illustrates how obtaining an appropriate ITT analysis is important in reaching valid conclusions. The trial found an excess risk of CV events on rofecoxib compared with placebo in 2,586 patients with a history of colorectal adenomas. The first report only included events occurring during treatment and up to 14 days after the last dose: 46 patients versus 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 (**Figure 3A**). 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 patients versus 34 patients with thrombotic events ($p = 0.006$), and the evidence appeared compatible with an early increase in risk that persists

FIGURE 3 Excess Risk of Thrombotic Events on Rofecoxib in the APPROVe Trial, First as On-Treatment Analysis and Subsequently as ITT Analysis



Kaplan-Meier cumulative frequency of thrombotic events in the APPROVe trial calculated from an **(A)** on-treatment analysis and **(B)** ITT analysis. Data from Bresalier et al. (37) and Baron et al. (38). APPROVe = Adenomatous Polyp Prevention on Vioxx; APTC = Antiplatelet Trialists' Collaboration; ITT = intention-to-treat.

1 year after stopping treatment (**Figure 3B**). Rofecoxib was withdrawn from worldwide markets due to these safety concerns, although it is worth noting that the second report (38) was published over 3 years after this withdrawal.

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

ROCKET-AF (Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation) (39), which compared rivoroxaban and warfarin in 14,264 patients with atrial fibrillation, had a similar construct, with both noninferiority and superiority hypotheses. In ITT analysis, the primary event, stroke or systematic embolism, occurred less frequently on rivaroxaban: 269 events versus 306 events (HR: 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 events versus 240 events (HR: 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 toward a claim of superiority and the published conclusion was "rivaroxaban was noninferior to warfarin for the prevention of stroke or systemic embolism." In general, per-protocol analyses introduce bias: nonadherent 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 (Placement of Aortic Transcatheter Valves) trial (40) of transcatheter aortic valve replacement versus aortic valve surgery, there were 4 (1%) and 38 (11%) patients, respectively, who did not get their intended treatments. ITT from randomization was the primary analysis. However, a supplementary as-treated analysis from time of treatment (excluding those 42 patients) helped to confirm that there was no mortality difference, but a borderline significant excess of strokes. In this situation, ITT will tend to

dilute any real treatment differences (although giving a valid comparison of strategies along with their protocol deviations), and an as-treated analysis may be biased. For example, it may be higher-risk patients who declined surgery. A consistency across the 2 analyses is helpful. For noninferiority trials, both ITT and per-protocol analyses should be presented, a point we will clarify further in the last paper 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 borne in mind. A new dramatic claim (whether benefit or harm) is often on the basis of 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 of the selection biases that have occurred, for example, across multiple analyses (endpoints, trials, and 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, for example, 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 (Simvastatin and Ezetimibe in Aortic Stenosis) trial (41) in 1,873 patients with aortic stenosis, the active treatment, simvastatin plus ezetimibe, had an unexpected excess of cancers compared with placebo: 105 incident cancers versus 70 incident cancers ($p = 0.01$), and 39 cancer deaths versus 23 cancer deaths ($p = 0.05$). Given the wealth of safety data available on statins, the potential culprit was thought to be ezetimibe. There was an urgent need to study the totality of evidence regarding ezetimibe and cancer. Hence, 2 ongoing, large trials of ezetimibe versus placebo on background statins, SHARP (Study of Heart and Renal Protection) and IMPROVE-IT, published their combined interim findings: 313 incident cancers versus 329 incident cancers ($p = 0.61$) and 97 cancer deaths versus 72 cancer deaths ($p = 0.07$) for the ezetimibe and placebo groups, respectively (42). There was also no logical pattern with respect to specific cancers. The conclusion was that "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 2 trials.

A similar pattern emerged with an apparent excess of MIs on rosiglitazone, first proposed in a meta-analysis by Nissen and Wolski (43) (odds ratio vs. control: 1.43; 95% CI: 1.03 to 1.98; $p = 0.03$). The main subsequent evidence came from the RECORD (Rosiglitazone Evaluated for Cardiovascular Outcomes in Oral Agent Combination Therapy for Type 2 Diabetes) trial (44) in 4,447 diabetic patients followed for a mean of 5.5 years: the HR for MI (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 MI. However, the overall safety profile, especially risks of heart failure and bone fractures, meant that the marketing authorization 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 randomized 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 MI attributed to rofecoxib came from the VIGOR (Vioxx Gastrointestinal Outcomes Research) trial (45) of rofecoxib versus naproxen in patients with rheumatoid arthritis: 20 MIs versus 4 MIs, relative risk 5.00 (95% CI: 1.68 to 20.13). Curiously, the original publication reported it as a benefit of naproxen, with 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 CI. Subsequent evidence, both from the APPROVe trial (37) (see previous text) and from meta-analyses, showed an effect closer to a doubling of risk, rather than a 5-fold increase. Furthermore, the meta-analysis showed no heterogeneity of this risk across the class of Cox-2 inhibitors (46). So, was rofecoxib "the unlucky one" to first focus attention on this class phenomenon?

We now turn to claims of treatment efficacy on the basis of apparently large benefits in small trials. For instance, a trial of acetylcysteine versus placebo (47) for prevention of contrast-induced nephropathy in 83 patients revealed 1 acute reduction versus 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.

TABLE 5 CONSORT Checklist of Items to Include When Reporting a Randomized Trial

Section/Topic	Item Number	Checklist Item
Title and abstract		
	1a	Identification as a randomized 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
Randomization		
Sequence generation	8a	Method used to generate the random allocation sequence
	8b	Type of randomization; 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 (e.g., 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 number of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome
	13b	For each group, losses and exclusions after randomization, 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 analyzed	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
Generalizability	21	Generalizability (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

Data from Moher et al. (51).
CONSORT = Consolidated Standards of Reporting Trials.

There is an ongoing controversy concerning the perioperative use of beta-blockers in noncardiac surgery. The small DECREASE (Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography) trial (48) of bisoprolol showed apparently marked benefits: in 112 patients there were 2 deaths versus 9 deaths ($p = 0.02$) and 0 MIs versus 9 MIs ($p < 0.001$) in the bisoprolol and control

groups, respectively. In general, it is wise to consider such dramatic findings on the basis of 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 concerning the value of beta-blockers is inconclusive. Again, a large, well-designed trial is needed.

The PRAMI (Preventive Angioplasty in Acute Myocardial Infarction) trial (49) of preventive angioplasty versus stenting of the culprit lesion only is an example of an apparently very large treatment effect on the basis of relatively small numbers of events: 53 primary endpoints versus 21 primary endpoints (refractory angina, MI, or cardiac death; HR: 0.35; 95% CI: 0.21 to 0.58; $p < 0.001$). For any intervention to reduce an event rate by more than one-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 (Complete vs Culprit-only Revascularization to Treat Multi-vessel Disease After Primary PCI for STEMI) (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 (Consolidated Standards of Reporting Trials) (51) is an established set of guidelines for reporting clinical trials to which many journals, including *JACC*, expect authors to adhere. There is a helpful checklist of items to include (**Table 5**) that covers all sections of a paper, including the methods, results, and conclusions. Specific issues covered in CONSORT extensions include noninferiority trials, pragmatic trials, reporting of harms, and what to include in a trial abstract (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., noncompliance, dropouts) and analysis (e.g., was ITT analysis achieved). Relevant to the controversies discussed in this paper is that authors should make the pre-defined analyses and any priorities among them clear: for example, the primary endpoint's overall analysis should dominate the conclusions and the abstract, whereas any important safety issues should also be 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 paper has provided a statistical insight that will help trialists to present and readers to acquire a balanced perspective.

REPRINT REQUESTS AND CORRESPONDENCE: Prof. Stuart Pocock, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom. E-mail: Stuart.Pocock@LSHTM.ac.uk.

REFERENCES

1. Cook RJ, Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. J R Stat Soc Ser A Stat Soc 1996;159:93–110.

2. Bonaca MP, Bhatt DL, Cohen M, et al., for the 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.

3. Dormandy JA, Charbonnel B, Eckland DJA, et al., for the PROactive Investigators. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. Lancet 2005;366:1279–89.

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

5. Scirica BM, Bhatt DL, Braunwald E, et al., for the 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.

6. White WB, Cannon CP, Heller SR, et al., for the EXAMINE Investigators. Alogliptin after acute coronary syndrome in patients with type 2 diabetes. N Engl J Med 2013;369:1327–35.

7. Green JB, Bethel MA, Armstrong PW, et al., for the TECOS Study Group. Effect of sitagliptin on cardiovascular outcomes in type 2 diabetes. N Engl J Med 2015;373:232–42.

8. Serruys PW, Morice MC, Kappetein AP, et al., for the SYNTAX Investigators. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med 2009;360:961–72.

9. Mohr FW, Morice MC, Kappetein AP, et al. Coronary artery bypass graft surgery versus percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. Lancet 2013;381:629–38.

10. Abbott Vascular. 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). 2015. Available at: https://clinicaltrials.gov/ct2/show/NCT01205776. Accessed October 19, 2015.

11. Pocock SJ, Assmann SE, Enos LE, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current

practice and problems. Stat Med 2002;21:2917–30.

12. Zannad F, McMurray JJV, Krum H, et al., for the EMPHASIS-HF Study Group. Eplerenone in patients with systolic heart failure and mild symptoms. N Engl J Med 2011;364:11–21.

13. Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. Stat Med 2002;21:2899–908.

14. Cannon CP, Blazing MA, Giugliano RP, et al., for the IMPROVE-IT Investigators. Ezetimibe added to statin therapy after acute coronary syndromes. N Engl J Med 2015;372:2387–97.

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

16. GISSI-HF Investigators. Effect of rosuvastatin in patients with chronic heart failure (the GISSI-HF trial): a randomised, double-blind, placebo-controlled trial. Lancet 2008;372:1231–9.

17. Yusuf S, Pfeffer MA, Swedberg K, et al., for the 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.

18. Committee for Proprietary Medicinal Products (CPMP). Committee for Proprietary Medicinal Products (CPMP) points to consider on adjustment for baseline covariates. Stat Med 2004;23:701–9.

19. Assmann SF, Pocock SJ, Enos LE, et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355:1064–9.

20. Bhatt DL, Fox KAA, Hacke W, et al., for the CHARISMA Investigators. Clopidogrel and aspirin versus aspirin alone for the prevention of atherothrombotic events. N Engl J Med 2006;354:1706–17.

21. Pfeffer MA, Jarcho JA. The charisma of subgroups and the subgroups of CHARISMA. N Engl J Med 2006;354:1744–6.

22. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine—reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357:2189–94.

23. Stone GW, Rizvi A, Newman W, et al., for the SPIRIT IV Investigators. Everolimus-eluting versus paclitaxel-eluting stents in coronary artery disease. N Engl J Med 2010;362:1663–74.

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., for the SHIFT Investigators. Ivabradine and outcomes in chronic heart failure (SHIFT): a randomised placebo-controlled study. Lancet 2010;376:875–85.

26. Pocock SJ, Lubsen J. More on subgroup analyses in clinical trials. N Engl J Med 2008;358:2076, author reply 2076–7.

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;34:2823–9.

28. Fox KAA, Clayton TC, Damman P, et al., for the FIR Collaboration. 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. J Am Coll Cardiol 2010;55:2435–45.

29. Wiviott SD, Braunwald E, McCabe CH, et al., for the TRITON-TIMI 38 Investigators. Prasugrel versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2007;357:2001–15.

30. Salisbury AC, Wang K, Cohen DJ, et al. Selecting antiplatelet therapy at the time of percutaneous intervention for an acute coronary syndrome: weighing the benefits and risks of prasugrel versus clopidogrel. Circ Cardiovasc Qual Outcomes 2013;6:27–34.

31. Connolly SJ, Ezekowitz MD, Yusuf S, et al., for the RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med 2009;361:1139–51.

32. Giugliano RP, Ruff CT, Braunwald E, et al., for the ENGAGE AF-TIMI 48 Investigators. Edoxaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2013;369:2093–104.

33. Pocock SJ, Stone GW, Mehran R, et al. Individualizing treatment choices using quantitative methods. Am Heart J 2014;168:607–10.

34. Mega JL, Braunwald E, Wiviott SD, et al., for the ATLAS ACS 2-TIMI 51 Investigators. Rivaroxaban in patients with a recent acute coronary syndrome. N Engl J Med 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). J Am Coll Cardiol 2013;62:777–81.

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., for the Adenomatous Polyp Prevention on Vioxx (APPROVe) Trial Investigators. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med 2005;352:1092–102.

38. Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. Lancet 2008;372:1756–64.

39. Patel MR, Mahaffey KW, Garg J, et al., for the ROCKET AF Steering Committee for the ROCKET AF Investigators. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med 2011;365:883–91.

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

41. Rossebø AB, Pedersen TR, Boman K, et al., for the SEAS Investigators. Intensive lipid lowering with simvastatin and ezetimibe in aortic stenosis. N Engl J Med 2008;359:1343–56.

42. Peto R, Emberson J, Landray M, et al. Analyses of cancer data from three ezetimibe trials. N Engl J Med 2008;359:1357–66.

43. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 2007;356:2457–71.

44. Home PD, Pocock SJ, Beck-Nielsen H, et al., for the RECORD Study Team. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. Lancet 2009;373:2125–35.

45. Bombardier C, Laine L, Reicin A, et al., for the VIGOR Study Group. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. N Engl J Med 2000;343:1520–8.

46. Kearney PM, Baigent C, Godwin J, et al. 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;332:1302–8.

47. Tepel M, van der Giet M, Schwarzfeld C, et al. Prevention of radiographicl-agent–induced reductions in renal function by acetylcysteine. N Engl J Med 2000;343:180–4.

48. Poldermans D, Boersma E, Bax JJ, et al., for the Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. N Engl J Med 1999;341:1789–94.

49. Wald DS, Morris JK, Wald NJ, et al., for the PRAMI Investigators. Randomized trial of preventive angioplasty in myocardial infarction. N Engl J Med 2013;369:1115–23.

50. Population Health Research Institute. Complete vs culprit-only revascularization to treat multi-vessel disease after primary PCI for STEMI (COMPLETE). 2015. Available at: https://clinicaltrials.gov/ct2/show/study/NCT01740479. Accessed October 19, 2015.

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., for the CONSORT Group. CONSORT for reporting randomised trials in journal and conference abstracts. Lancet 2008;371:281–3.

KEY WORDS intention-to-treat analysis, logistic models, proportional hazards models, randomized controlled trials as topic, risk assessment, statistics