Strategies for Reporting of Primary and Secondary Outcomes in Randomized Clinical Trials

Brief title: Reporting Outcomes in Clinical Trials

Stuart J. Pocock PhD [1,2], Xavier Rossello PhD[1,2], Ruth Owen MSc [1], Tim J. Collier MSc[1], Gregg W. Stone MD [3], Frank W. Rockhold PhD [4]

[1] Medical Statistics Department, London School of Hygiene & Tropical Medicine, London, UK
[2] Centro Nacional Investigaciones Cardiovasculares, Madrid, Spain
[3] The Zena and Michael A Wiener Cardiovascular Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4] Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, USA.

Disclosure statements:
Dr Pocock consults with AstraZeneca, Boehringer Ingelheim, Boston Scientific, Edwards, Medtronic, Vifor.
Dr Stone has received speaker honoraria from Cook and Terumo; has served as a consultant to Valfix, TherOx, Vascular Dynamics, Robocath, HeartFlow, Gore, Ablative Solutions, Miracor, Neovasc, V-Wave, Abiomed, Ancora, MAIA Pharmaceuticals, Vectorious, Reva, and Matrizyme; and owns equity/options in Ancora, Qool Therapeutics, Cagent, Applied Therapeutics, the Biostar family of funds, SpectraWave, Orchestra Biomed, Aria, Cardiac Success, the MedFocus family of funds, and Valfix.
Dr Rockhold reports personal fees from Novartis, personal fees from Eli Lilly, grants and personal fees from AstraZeneca, personal fees from Amgen, personal fees from Sanofi, personal fees from NovoNordisk, personal fees from Merck Research Labs, personal fees from UCB, personal fees from Tolerion, personal fees from Rhythm, personal fees from KLSMC Stem Cell, personal fees from Merck KGaA, personal fees from Janssen, personal fees from Phathom, personal fees from Clover Biopharma, personal fees from Sarepta, personal fees from Eidos, and personal fees from Icosavax outside the submitted work; and Equity Interest in GSK, Athira, DataVant, Spencer Healthcare.
Tim Collier, Ruth Owen and Xavier Rossello have nothing to disclose.

Corresponding Author:
Stuart Pocock,
Department of Medical Statistics
London School of Hygiene & Tropical Medicine
London WC1E 7HT
United Kingdom
Email: Stuart.Pocock@LSHTM.ac.uk
Tel: 020 7927 2413
Twitter: @pocock_stuart

ABSTRACT
Consensus as to best practices for the selection, reporting and interpretation of primary and secondary outcomes of randomized controlled trials (RCTs) is lacking. We thus review the strategies adopted in publications of RCTs for the analysis, presentation and interpretation of outcomes for treatment efficacy from a survey of all cardiovascular RCTs published in NEJM, Lancet and JAMA during 2019. We focus on the choice of primary (and co-primary) outcomes, the variety of approaches to selecting secondary outcomes, the options sometimes used to control type I error and the common practice to not correct for multiple testing in reporting of secondary outcomes. We comment on current practice across journals in the reporting of P-values and also how conclusions in trial reports frequently adhere to an undue reliance on $P<0.05$ as a basis for positive claims of treatment efficacy. A more nuanced interpretation based on the totality of evidence is warranted. We conclude with a set of recommendations for how future RCT reports could best select, report and interpret their findings on primary and secondary outcomes.

CONDENSED ABSTRACT
We survey recent practice in reporting of primary and secondary outcomes for cardiovascular randomised control trials (RCTs). Recommendations are made for how future RCT reports could best select, report and interpret findings on primary and secondary outcomes.

Key words: Randomized Controlled Trials, Strategies, Primary outcomes, Secondary outcomes

ABBREVIATIONS

JAMA	Journal of the American Medical Association
NEJM	New England Journal of Medicine
FDA	Food and Drug Administration
CV	Cardiovascular
RCT	Randomized Controlled Trial
MI	Myocardial Infarction
HF	Heart Failure
ITT	Intention-to-treat
CABG	Coronary Artery Bypass Grafting
PCI	Percutaneous Coronary Intervention

INTRODUCTION

Most randomized controlled trials (RCTs) pre-select a single primary outcome followed by one or more secondary outcomes for the evaluation of treatment efficacy (1–3). There may be additional safety outcomes and exploratory outcomes, but key to interpretation of trial findings is the primary outcome and the extent to which it provides evidence of a treatment difference.

The risk of a type I error runs high when examining multiple endpoints, so the interpretation of findings from secondary outcomes is a statistical challenge. If the primary outcome does not reveal a statistically significant treatment benefit it is common practice to report secondary outcomes in a spirit of data exploration without formal claims of positive findings, no matter how significant they may be (4), although some have argued for greater value from secondary endpoints even if the primary outcome failed (5).

Some trials pre-define a hierarchy of secondary outcomes for a sequence of formal statistical tests for treatment benefit (6), to be applied only if the primary outcome achieves a pre-specified level of significance, usually $P<0.05$. One adds in turn each secondary outcome that achieves $P<0.05$ to the set of claims for treatment efficacy until one of them fails to achieve this level of significance. Alternatively, the alpha for secondary endpoints (in the presence of a significant primary) may be preserved by an appropriate multiple testing algorithm (6).

Thus, consensus as to best practices for the selection, reporting and interpretation of primary and secondary outcomes from RCTs is lacking. The aim of this article is thus to describe current practices in the selection, analysis, presentation and interpretation of primary and secondary outcomes for treatment efficacy in RCTs, and to make recommendations to trialists, journals and regulators as to how these practices could be improved in the future.

A SURVEY OF CURRENT OUTCOME REPORTING PRACTICES

Methods

To examine the recent practices and policies of the reporting of primary and secondary outcomes, we surveyed trials published in the three main general medical journals: JAMA, Lancet and NEJM. We identified all RCTs published during Jan to Dec 2019, including both online and paper versions. We focused out attention on major RCTs in cardiovascular diseases, the area of our expertise, although the findings herein are likely generalizable to other specialties.

We extracted the following information for each trial:

1. Type of intervention: pharmaceutical, device, patient management surgery
2. Main source of funding: industry or public
3. Primary outcome(s): a single primary or 2 co-primaries
4. Was primary hypothesis test for superiority or non-inferiority?
5. Secondary outcomes: number reported
6. Any hierarchy of secondary outcomes? If so, how many were included
7. Any other multiple testing procedure used? If so, what was it
8. For every primary and secondary outcome we noted:
    a) the type of outcome: composite or single event; all-cause or cardiovascular death; non-fatal event types; quantitative measure; binary criteria or ordinal outcomes
    b) the level of statistical significance: $P \geq 0.05$, $P < 0.05$ or $P < 0.01$ or $P < 0.001$. When the P-value was not explicitly given it was calculated from the point estimate and 95% confidence interval.
9. Whether the Conclusions section of the Abstract and Discussion were confined to the primary outcome only, or also referred to secondary outcomes.

Each article was surveyed by 2 out of the 4 reviewers (TC, SP, RO, XR). Any inconsistencies found were resolved by consensus.

Survey Findings

We identified 84 RCT articles in total published during 2019: 20 in JAMA, 23 in Lancet and 41 in NEJM. The profile of these 84 trials is shown in Table 1. Approximately half of the trials evaluated a pharmaceutical intervention, with a quarter being trials of medical devices and just two of surgical interventions. Some device trials had surgery as the control arm. The remaining 18 trials evaluated various different forms of patient management.

The source of funding for trials was fairly evenly split between industry and public sources, 41 and 35 trials respectively, with both sources contributing in 8 trials.

The number of randomized participants ranged from 51 to 25,871 with a median of 1336 participants. The median or fixed follow-up time ranged from 6 hours to 10 years, with a median of 1 year (which was chosen in 22 trials). A follow-up of 90 days occurred in 11 trials. A broad range of disease conditions was studied (Table 1). Coronary artery disease (21 trials) was most frequent, followed by HF (11 trials) and stroke (9 trials).

Primary outcomes and components of composite primaries

Our findings regarding the primary outcome are summarized in Table 2. Among the 84 trials 75 (89%) had a single pre-defined primary outcome while 9 trials (11%) had two co-primary outcomes. Among these, 5 trials did not include a pre-defined multiplicity correction whereas 4 did. Amongst the latter, two RCTs split the alpha equally for the two outcomes, each requiring $P<0.025$ to achieve 5% significance, whereas the other two had unequal alpha split, 0.0372 vs 0.0094 and 0.045 vs 0.0119, the first adding up to less than 0.05 due to α-adjustment for interim analysis and the second adding up to more than .05 due to the two outcomes being correlated. In one trial without multiplicity correction the claim of significance for $P=0.04$ in one of the two co-primary outcomes could be challenged. For the other 8 trials with co-primary outcomes the presence or absence of multiplicity correction

would not have changed the conclusions, as the results were either highly significant (10 outcomes) or clearly non-significant (6 outcomes).

Among the 75 trials with a single primary outcome, 10 (13%) had a primary non-inferiority hypothesis, all but one of which (90%) demonstrated non-inferiority. For the remaining 65 trials (87%) with a primary superiority hypothesis, 38 (58%) achieved statistical significance at the 5% level. There was very strong evidence of superiority ($P<0.001$) in just 15 of the 65 (23%) superiority trials, while strong evidence was present ($P <0.01$ and ≥ 0.001) in 5 additional trials (8%). A modest strength of evidence ($P <0.05$ and ≥ 0.01) was present in 18 trials (28%). Of note, in two of these trials the positive finding was in the opposite direction to that hypothesized, i.e. offering evidence that the new treatment was harmful.

A composite endpoint was used in 37 of the 75 trials with a single primary outcome (49%). For the rest a disease outcome or binary criterion was used in 20 trials (27%), a quantitative measure was used in 12 trials (16%), all-cause death was the primary outcome in 4 trials (5%) and an ordinal outcome was used in 2 trials (3%). For trials with a composite primary (or co-primary) endpoint details regarding the number and type of components are shown in Table 3. A composite outcome with three components was the most common choice (20 trials), with two components next (11 trials). Four or five components were used in 11 trials and 1 trial had 8 components in the composite.

All composite outcomes included death as a component, evenly divided between cardiovascular death (24 composites) and all-cause death (23 composites). Amongst non-fatal components, MI was most common (29 trials), followed by stroke (27 trials), revascularisation (10 trials) and HF hospitalisation (6 trials). The most common choices of composites were death, MI or stroke (11 trials), death, MI or revascularisation (7 trials), and death or HF hospitalisation (5 trials). Most of these included cardiovascular death rather than all-cause death. For the 22 trials with a disease event, binary or ordinal outcome, the most

common choice was the Modified Rankin scale (6 trials), all-cause death (4 trials) and major bleed (4 trials). For the 15 trials with quantitative primary (or co-primary) outcomes these were most commonly a physiological measure (6 trials) followed by biomarker (4 trials), quality of life scale (2 trials) a risk score change (2 trials) and 1 other.

Secondary outcomes

All but one of the 84 trials pre-specified mutiple secondary outcomes. The statistical handling of secondary outcomes in these trials may be summarized as follows: A pre-declared hierarchy of secondary outcomes was reported in 18 trials (21%). Some other type of multiple testing procedure across secondary outcomes was used in 3 trials (4%). A list of secondary outcomes with no formal correction for multiple testing was observed in 63 trials (75%).

Table 4 summarizes the findings for the 18 trials with a hierarchical testing procedure of secondary endpoints. This practice was most common in industry sponsored trials. The number of secondary outcomes in the hierarchy varied considerably, ranging from one to nine outcomes with a mode of five outcomes. In 7 trials the hierarchy was not used because the primary (or co-primary) outcomes did not achieve $P<0.05$. In 3 trials hierarchical testing stopped at the first hurdle ($P<0.05$ was not achieved) while in 2 others only the first step in the hierarchy was significant. However, in 6 trials three or more hierarchical outcomes achieved $P<0.05$. In 2 trials, outcomes lower down the hierarchy would have achieved $P<0.05$, but formal testing had already stopped due to lack of significance for an outcome higher in the list.

The 3 trials using a correction for multiple testing employed: 1) use of the Holm procedure (20) for three secondary outcomes, with one of them being formally significant; 2) use of Bonferroni correction across 12 secondary outcomes; 3) use of a graphical approach for multiple comparisons (7) to “strongly control the overall type I error”.

The 63 RCTs that reported secondary outcomes without any pre-specified hierarchical testing or multiple testing procedure raised several issues:

1) It was often difficult to relate those reported to any pre-declared list of secondary outcomes though ICMJE and CONSORT recommend that they should match (8). This would have required inspection of the trial protocol and/or statistical analysis plan, which was beyond our remit. Hence, we evaluated those secondary outcomes that were reported, which may be more or less than any pre-defined list.
2) It is possible that some RCTs had a more limited set of key secondary outcomes with others being more exploratory. In practice, it was often not possible to make this distinction from the trial reports.
3) For many trial reports it was relevant to distinguish between efficacy outcomes and safety outcomes. Unless safety was the primary aim of an RCT (e.g. the primary outcome was a safety issue such as bleeding), we concentrated on efficacy issues only in studying secondary outcomes.

As shown in Table 5, the median number of secondary outcomes was 7 with a range from 1 to over 30. In Table 5 we also cross-classify the strength of evidence for a treatment effect in the primary outcome against the corresponding strength of evidence found for the "most significant" secondary outcome. Both are summarized in 4 categories: not significant (i.e. $P \geq 0.05$), $P<0.05$, $P<0.01$ and $P<0.001$. Overall a moderate association is noted.

Among 53 superiority trials, 11 (21%) reported non-significance for both the primary and all secondary outcomes. At the other extreme 11 RCTs achieved $P<0.001$ for both the primary and the most significant secondary outcome. A further 20 RCTs achieved a higher level of significance amongst the secondary outcomes than was achieved for the primary outcome. This phenomenon was more common as the number of secondary outcomes increased: 12 out of the 20 had 12 or more secondary outcomes.

For the 9 non-inferiority trials that actually demonstrated non-inferiority for the primary outcome, 2 of them showed a significant treatment difference (P<0.05) amongst the secondary outcomes, one in favour of the new treatment and the other with more deaths with the new treatment.

Relation of conclusions to outcomes

Table 6 relates the Conclusions of each manuscript (as summarized in the Abstract and also at the end of Discussion) to the actual outcomes of the trials. Of the 84 articles studied, 60 (71%) confined their conclusions to the primary (or co-primary) outcome only. This was the case for all 20 articles in JAMA, which may reflect journal policy. In 19 (23%) of the articles, one or more secondary outcomes were also included in the Conclusions, although in 5 cases these secondary outcomes appeared only in the Conclusions of the Discussion, not the Abstract. In 5 (6%) of the articles, safety outcomes were also mentioned in the Conclusions. In 2 articles the Conclusions referred to subgroup findings for the primary outcome.

Some interesting examples

All the following examples are selected from the 84 articles surveyed, being case studies that illustrate some of the challenges faced by authors and journals in interpreting primary and secondary outcome findings.

The DECLARE trial (9) of dapagliflozin versus placebo in high-risk diabetics published in NEJM had two co-primary outcomes and a hierarchy of eight secondary outcomes. To preserve the overall type I error both primaries needed to achieve P<0.05 for formal hierarchical testing of secondary outcomes to be undertaken. This was not the case since the major adverse cardiovascular events (MACE) co-primary outcome had P=0.17. Thus, the first secondary outcome, a renal composite, was only reported with hazard ratio 0.76 (95% CI 0.67 to 0.87), the nominal P<0.001 going unmentioned in line with current NEJM practice.

Thus, a formal claim of benefit for this renal outcome was not permitted in the article nor subsequently by the FDA. The renal finding was stated in the Abstract's Results and mentioned in the Discussion's Conclusions but not in the Abstract's Conclusions.

The DAPA-HF trial (10) of dapagliflozin versus placebo in HF with reduced ejection fraction had a composite primary outcome (worsening HF or CV death) and a hierarchy of five secondary outcomes. The fourth in the hierarchy (worsening renal function) did not achieve $P<0.05$ whereas all other primary and secondary outcomes did. Of note all-cause death was fifth in the hierarchy with hazard ratio 0.83 (95% CI 0.71 to 0.97). Its nominal $P=0.02$ does not appear in line with NEJM policy. While mortality does appear in the Abstract's Results it does not appear in any Conclusions.

The EXCEL trial of PCI or CABG for left main disease was powered for non-inferiority for a primary composite of death, stroke or MI at a median follow-up of 3 years (11). Evaluation of this composite outcome when all patients reached 5-year follow-up (the conclusion of the study) was not formally specified for hypothesis testing. The principal conclusion from the final 5-year report from the EXCEL trial was that there was no significant treatment difference based on the odds ratio and its 95% CI for the primary composite outcome, although strictly this observation should be considered hypothesis generating (12). Moreover, amongst 13 secondary outcomes there were some notable observed differences: all-cause death (3.1% greater after PCI), peri-procedural MI (2.1% greater after CABG), other MI (3.2% greater after PCI), cerebrovascular events (1.9% greater after CABG) and ischaemia-driven revascularisation (6.9% greater after PCI). The P values for these 5 outcomes would have been 0.04, 0.04, 0.001, 0.001, 0.05 and <0.001 respectively but were not published per NEJM policy. Although the 95% confidence intervals of the differences were published, in our opinion the inclusion of their P values would have helped the reader to concentrate on the potentially important exploratory findings amongst the numerous secondary outcomes

reported. This example also raises the issue of whether all the subtleties of a major trial's findings can be adequately captured by a Conclusion that only focusses on the primary composite outcome.

The THEMIS trial of ticagrelor versus placebo in patients with diabetes and stable coronary disease had two simultaneous articles: one in NEJM (13) on the whole trial population and the other in Lancet (14) devoted to the subgroup of patients with a history of PCI. Both reveal a significant reduction in the primary composite outcome (CV death, MI or stroke) although with increased risk of major bleeding. But the Lancet article's Conclusions go further stating that in those with prior PCI "ticagrelor provided a favourable net clinical benefit (more than in patients without a history of PCI)". This was derived from a post hoc exploratory composite outcome called irreversible harm events (all-cause death, MI, stroke, fatal bleed or intracranial haemorrhage). One might question whether an exploratory outcome in a subgroup merits inclusion as a key Conclusion. This is a good example of how reporting practices can vary between journals.

The CABANA trial (15) of catheter ablation versus drug therapy in atrial fibrillation was published in JAMA with the Conclusion that the former did not significantly reduce the primary composite of death, disabling stroke, serious bleed or cardiac arrest. The problems of crossovers and lower-than-expected events were pointed out. A more positive finding for the key secondary endpoint of death or CV hospitalization (hazard ratio 0.83, 95% CI 0.74-0.93, P=0.001) was presented but did not feature in the Conclusions. Neither did more positive as-treated and per-protocol analyses of the primary outcome. While the ITT primary outcome analysis preserves the advantage of randomization, the high crossover rate to ablation in the control group could dilute a true efficacy signal thereby complicating its interpretation. One might infer that the trial was unable to reach a robust conclusion regarding the merits of catheter ablation.

It is generally recognized that subgroup analyses should be deemed as exploratory or hypothesis generating and should not feature in an RCT article's Conclusions (16). Our survey encountered two exceptions to this principle: the PARAGON-HF and SYNTAX trials (17, 18) in NEJM and Lancet respectively. PARAGON-HF (17) compared sacubitril-valsartan and valsartan alone in HF with preserved ejection fraction. The primary outcome (total HF hospitalisations and CV death) had rate ratio 0.87, 95% CI 0.75 to 1.01, P=0.06. This lack of formal statistical significance was the basis for the Abstract's Conclusion. But the Discussion's concluding paragraph was more liberal ending with "future research should focus on the potential role of angiotensin-neprilysin inhibition in HF patients with ejection fraction below normal but not frankly reduced". This is based on evidence that ejection fraction is an apparent effect modifier, assessment of which by the reader was complicated by the NEJM's policy of not permitting interaction P-values (which would have been P=0.002). Based on this study the FDA recently expanded the indication for sacubitril-valsartan to reduce the risk of cardiovascular death and hospitalization for HF in adult patients with chronic HF regardless of ejection fraction. However, the FDA also acknowledged the importance of the aforementioned interaction, noting that its benefits are most clearly evident in patients with ejection fraction below normal.

The SYNTAX trial's 10-year follow-up report compared PCI and CABG in patients with three-vessel and left-main disease (18). The hazard ratio for the article's primary outcome of all-cause death (an originally unplanned endpoint that required patient reconsent) was 1.17 (95% CI 0.97 to 1.41, P=0.092), leading to the overall Conclusion that "no significant difference" existed. But based on an observed interaction with type of coronary artery disease (P=0.019), the Conclusions added: "However, CABG provided a significant survival benefit in patients with three-vessel disease, but not in patients with left main coronary artery

disease". These two examples are exceptions to the "rule" of no subgroup findings in the Conclusions. Whether such exceptions are appropriate is open to debate.

In addition, the SYNTAX trial has published its primary and secondary end points ranging in follow-up from 1 to 10 years. Does examining the same outcome at different time points increase the type I error and raise skepticism as the sequence unfolds? Although one could technically argue that multiple looks at different times require control of type 1 error, late outcomes are predicated on early results, and it would be unrealistic to capture both shorter-term and long-term data in a single trial report. The greatest reliance should be on the time point at which the primary endpoint was powered, with follow on reports interpreted according to their similarities or differences from the principal analysis.

DISCUSSION

Overall aims

The aim of this article was to review the strategies trialists adopt in the selection, presentation and interpretation of primary and secondary outcomes for evaluating treatment efficacy. This topic has received relatively little attention in the methodological literature (19, 20) We focused on major RCTs in cardiovascular diseases and their publications in the top three general medical journals, but many of our findings and recommendations can be generalized to: 1) RCTs in other diseases; 2) RCTs published in specialist medical journals; 3) RCT submissions to regulatory authorities; and 4) 'late-breaking' or 'hotline' presentations of RCTs at medical conferences. Our choice of journals - NEJM, Lancet and JAMA - inevitably provides a group of trials that are larger and focused on outcomes of greater clinical relevance than a less selected set of journals would provide.

There are several key aspects regarding primary and secondary outcomes: a) choosing the appropriate outcomes; b) defining what will be the primary and what will be secondary endpoints; c) statistical approaches to interpreting secondary outcome findings, especially

regarding type I error control; d) journal guidelines and policies on such matters; e) consequent Conclusions in each article's Abstract and at the end of the Discussion (sections that have the greatest impact on the reader). The first three issues should be clearly defined in the trial protocol and statistical analysis plan (SAP) whereas the latter two get resolved by mutual agreement between authors and journal editors.

Choice of outcomes

First let us consider the choice of outcomes, especially the primary. This is a huge topic, and for many RCTs decisions are largely based on prior benchmarks or findings from related studies. The FDA plays an important role in defining which outcomes are sufficient to warrant regulatory approval and labelling of a drug or device. Guidelines committees, providers, payors and of course patients may appraise varying outcome measures to be of greater or lesser relevance. For pivotal trials the focus is (nearly always) on outcomes reflecting key aspects of patient well-being that are most clinically meaningful, rather than surrogates. Most CV trials concentrate on "hard outcomes", usually disease events, although there is an increased interest in patient reported outcomes (PROs) in some fields e.g. angina and HF, and the PRO instruments can be used to assess outcome severity, such as the Modified Rankin scale for stroke. It is important to select the appropriate time period for the outcome to best capture treatment efficacy: some trials elect to evaluate the same outcome over two or more time periods, with one pre-defined as primary.

Composite outcomes

Some key outcomes, e.g. mortality, occur too infrequently to provide adequate statistical power in most trials; hence a composite of both fatal and non-fatal events is commonly chosen as primary - this occurred in approximately half the RCTs we surveyed. There can be much debate on what components to include in the composite (21, 22). For instance, do we include death from all-causes or just cardiovascular deaths? Also, how many components to

have, bearing in mind that including less clinically important items (e.g. unstable angina or revascularization) will add to the number of events, may or may not enhance statistical power, but will dilute the clinical impact and interpretability of the composite (23).

The ISCHEMIA trial (24) illustrates the dilemma: the initial protocol had a 5-component primary outcome of CV death, MI or hospitalisation for unstable angina, HF or cardiac arrest, the protocol was subsequently amended for the more meaningful outcome of CV death or MI, preserving a process to reflexively revert back to the original primary endpoint were enough events for the more ambitious 2-component outcome not accruing during enrolment. This indeed was the case. Though validity done on blinded interim evidence, this change did generate some controversy. In hindsight, the trial interpretation would have been nearly identical using either primary outcome. In general, changing the primary outcome during a trial while being blinded to treatment assignment is an acceptable practice and does not carry a statistical penalty but perhaps inevitably arouses suspicion as to whether the blind was truly preserved by everyone involved.

Whatever components one chooses, it is important to also provide secondary analyses for each separate component, so readers can see whether any overall treatment difference is consistent across components. In this regard it is increasingly recognized that there are clinical priorities amongst the components of a composite, which go unrecognized in a time-to-first-event analysis, and hence methods based on a pre-declared hierarchy of events (e.g. win ratio) may be more appropriate (25). Each clinical field has its own controversies. For instance, in chronic HF should one select time to first HF hospitalization or CV death, or should total HF hospitalizations (including repeats) be included in the primary outcome (26)? There are no universal right answers for such matters, but experience from past trials may elucidate what is the wisest choice, taking account of the relative statistical power and clinical meaningfulness of the options.

A single primary outcome

The great majority of trials (89% in our survey) pre-defined a single primary outcome. Obviously, the choice of primary outcome is crucial, balancing clinical and statistical issues: the measure(s) that best capture the overall efficacy of a treatment that can be adequately powered for a realistic treatment benefit. Wisely choosing the primary outcome is a huge responsibility as this single decision will greatly affect how the trial will be perceived. However, a single P value from hypothesis testing of the primary endpoint can never replace the need for a trial's interpretation to rely on the totality of evidence, including secondary and safety outcomes (2, 3).

Co-primary outcomes

A few trials (11% in our survey) elected to have two co-primary outcomes, in order to reflect two different aspects of potential treatment efficacy. As our survey examples illustrate, a variety of options were chosen for control of type I error. Some trials had a primary efficacy and a primary safety outcome, which is appropriate when past experience means informs what are the key efficacy and safety issues. Statistically they are usually handled separately (i.e. $P<0.05$ required for each, both of which must be met to declare trial success), and both should be included in the trial's conclusions.

Secondary outcomes

There is an immense diversity across trials in the reporting of secondary outcomes, as regards their number and kind, and what steps are taken to control type I error. Some trials have a pre-defined set of secondary outcomes which get prioritised over a broader set of other (exploratory) outcomes, whereas in other trials the selection and reporting of secondary outcomes is less formally prescribed. For all trials we see the need for a greater consistency of approach (20), whereby all publications should adhere to the protocol and SAP pre-

specifications regarding secondary endpoints, with authors and journals ensuring this happens.

One surprising feature of our survey is that the majority of trial reports (75%) reported secondary outcomes without any consideration of multiplicity. This leads to secondary outcomes being presented in a spirit of exploratory data analysis, which has implications for their findings not impacting on the overall trial conclusions. We feel there is room for improvement here, whereby more trials should formally integrate some key secondary outcomes into a statistical testing strategy.

Hierarchy of secondary outcomes

For control of type I error amongst secondary outcomes, our survey found the most common approach was to have a pre-declared hierarchy of outcomes (23% of trials): provided the primary outcome achieves $P<0.05$ each secondary outcome in turn is inspected to see if it also achieves $P<0.05$. The sequence stops once one fails to do so. This approach is more common in industry-sponsored trials (6), presumably with an eye on broadening a regulatory approval for extended label claims and promotion of other outcomes besides the primary. This practice has its problems: somewhat arbitrary choices are made regarding the hierarchical sequence (juggling what is best powered and what is most clinically important) and it gives undue emphasis to $P<0.05$ as a decision-making tool.

Correction for multiple testing

An alternative practice is to treat a pre-defined key set of secondary outcomes as being of equal importance, and to undertake some form of correction for multiple testing (again provided the primary outcome achieves $P<0.05$). The options are well explained in an FDA guidance (6) and include Bonferroni, Hochberg (27), Holm (28) and Bretz (7) graphical approaches. We feel the latter two are particularly helpful. At present such multiplicity correction appears relatively uncommon (only 3% of trials in our survey). We encourage a

greater uptake of these methods in future, and anecdotally this may already be underway, since they provide a means of making valid claims for secondary outcomes while paying due respect for type I error and avoiding "playing games" with hierarchies.

Interpretation of p-values

While control of type I error is a valuable means of inhibiting false positive claims of treatment efficacy, an unfortunate consequence is that P<0.05 gets used by authors, journals and readers as a decision-making tool in expressing any result as "positive" or "negative" (2, 3, 21). This is not in the spirit of good statistical science (29–32), and a more meaningful interpretation of P-values in a more continuous graded manner is warranted: that is, the smaller is P the stronger the evidence to contradict the null hypothesis of no treatment effect (33). In our survey, the primary outcome revealed P between 0.05 and 0.01 for treatment efficacy in 28% of trials (representing a modest level of evidence) whereas in 23% of trials P<0.001 was achieved, which then carries the opportunity to declare "strong evidence" of superiority. For those instances where P is only just below 0.05 a more cautious claim of "some evidence" is warranted bearing in mind that the magnitude of effect (if not a false positive) is imprecisely estimated with a wide confidence interval (34).

Also, evaluating strength of evidence simply on the basis of P values without taking into consideration the effect size, confidence interval and the clinical importance of the outcome is at best an incomplete exercise. Use of a Bayes factor may be a valuable alternative for assessing the strength and precision of evidence (35).

An article's conclusions

The Conclusions in any article (both in Abstract and end of Discussion) are usually confined to the primary outcome. However, in 23% of articles we surveyed one or more secondary outcomes featured in Conclusions (though more often in the Discussion rather than the Abstract). Most Conclusions are a straightforward re-statement of the primary result, e.g.

"treatment X significantly reduced the incidence of outcome Y". The emphasis on 5% significance as a "magic cut-off" for a positive statement is unfortunate, and we would like to see a more nuanced interpretation being used. For instance, if the hypothesized magnitude of treatment effect used in the original design's power calculation were true then P around 0.001 should be expected (36).

Regulators and journals

We feel there is an important distinction between the roles of regulators and medical journals. Regulators make decisions whether to approve a treatment or not (so do payers) based on the totality of evidence on treatment efficacy and safety, which often entails multiple studies. In contrast, the prime role of journals is to provide and interpret the scientific evidence from a specific RCT, often in the context of other relevant research. Hence, we would encourage journals to a more subtle view beyond the crude confines of a "positive" or "negative" conclusion.

Journal policies and guidelines

To address the common misuse and misinterpretation of P-values for hypothesis testing when multiplicity is not adjusted for, NEJM recently introduced a new policy restricting P- value reporting to tests that were pre-specified and for which type I error was acceptably controlled (37). Interpretation of secondary endpoints not adjusted for multiplicity is achieved solely through assessment of point estimates and their confidence intervals (without P values, although statisticians can still calculate the missing P-values if they so wish). While well-intentioned, this practice has caused consternation amongst readers and trialists, many of whom continue to include such P-values in the corresponding conference presentations, which often coincide with the NEJM article.

Conversely, the Lancet has guidelines for authors regarding randomized trials (38) with no restriction on the reporting of P-values. JAMA has a recent editorial on reporting and

interpretation of RCTs (39) in which they give particular attention to how endpoint results should impact on the Conclusions drawn both in the Abstract and the end of Discussion. The CONSORT Reporting Guidelines (40) are widely accepted by medical journals. As regards outcomes they declare: "Results should be reported for all planned primary and secondary endpoints, not just for analyses that were statistically significant or 'interesting'. Selective reporting within a study is a widespread and serious problem". Similar guidance appears in ICMJE Recommendations on publication in medical journals (8). Nowadays, more attention is being paid to trial protocols and statistical analysis plans, so any risk of such distortive reporting is substantially less. It should be acknowledged that when multiple secondary and exploratory outcomes are pre-specified, it may not be practical that all analyses appear in the principal report; several secondary manuscripts may be required to fully elucidate their findings.

An alternative approach to reporting p-values

Thus, while NEJM's intention to downgrade inferences regarding secondary outcomes (outside of any planned correction for multiple testing) is fundamentally wise, suppression of all nominal P-values may not be a sufficient solution (and in the case of subgroup interactions, promotes confusion). We suggest an alternative approach (for all journals) whereby P-values (and their treatment effects) for which type I error has been controlled are highlighted (perhaps in bold type) while other nominal P-values (and their treatment effects) are downgraded (in faint type). The principle is to inform the reader whether any treatment comparison exhibits *some potential evidence* of an effect beyond the realm of chance, information that is most succinctly expressed as a nominal P-value. One can add a qualifier in the footnotes of tables (based on recent NEJM examples) such as "The P-values and 95% confidence intervals for secondary outcomes have not been adjusted for multiple comparisons, and therefore inferences drawn from these intervals may not be reproducible"

(or "should be considered exploratory only"), but the practical utility of this disclaimer is arguable (41). The journal should further insist on a mature and cautious interpretation of these secondary hypothesis-generating findings in the Discussion, emphasizing a) the pitfalls of multiple testing without adjustment; b) the correct interpretation of P-values as conveying strength of evidence rather than misguidedly used as a decision tool; and c) interpretation of the clinical relevance of the magnitude of the observed treatment effects.

Observational studies

We note that in reports of observational studies such formal recognition of multiplicity issues often gets scant attention and it seems more permitted to "data dredge". While the STROBE guidelines for reporting of observational studies (42) have helped to raise standards, attempts to control type I error are often lacking.

Safety outcomes

Another topic is how to report on treatment safety and adverse events. Some trials will have a pre-defined primary safety outcome and possibly a few secondary safety outcomes. This is in addition to the broader reporting of serious adverse events, which is inevitably less structured. In this article we have focused on outcomes related to treatment efficacy, since statistical challenges regarding control of type I error and multiplicity of outcomes are particularly pertinent. Reporting of safety outcomes is just as important, but represents a separate topic meriting further guidance (43, 44).

Limitations

Our survey of trials published in three major medical journals during 2019 has some limitations. In principle we would have liked to extend the survey to more years and more journals. Also, it would have been useful to explore the links between actual reporting of primary and secondary outcomes and what was pre-specified in the protocol and SAP. Nevertheless, we feel the current one year's exploration of cardiovascular trials in the three

leading journals provides satisfactory evidence of current practice. We have concentrated on trial evidence regarding superiority hypotheses for treatment efficacy. While we have also made observations about non-inferiority trials (11% in our survey) and safety outcomes, these issues lie beyond the scope and principal objective of our present study and merit separate investigations.

Our conclusions and recommendations

In conclusion, our survey of current practice in the reporting of primary and secondary outcomes in cardiovascular RCTs has elucidated many interesting features regarding choice of outcomes and approaches to control of type I error, the use and interpretation of P-values, a great diversity or reporting practices, the impact of journal policies and how Conclusions in trials reports are conveyed. The key issues along with our recommendations are summarized in the **Central Illustration**. We hope these ideas provide a useful contribution to reaching an eventual consensus regarding future desirable conventions in the reporting of RCTs.

HIGHLIGHTS

- Consensus is lacking re best practice for reporting of primary and secondary outcomes of RCTs
- We survey recent publications re analysis, presentation and interpretation of primary and secondary efficacy outcomes
- Recommendations are made regarding future practice, especially re interpretation of P-values

References

1. Solomon SD, Pfeffer MA. The future of clinical trials in cardiovascular medicine. Circulation 2016;133:2662–70.

2. Pocock SJ, Stone GW. The primary outcome is positive - Is that good enough? N Engl J Med 2016;375:971–79.

3. Pocock SJ, Stone GW. The primary outcome fails - What next? N Engl J Med 2016;375:861–70.

4. O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. Control. Clin. Trials 1997;18:550–56.

5. Davis CE. Secondary endpoints can be validly analyzed, even if the primary endpoint does not provide clear statistical significance. Control Clin Trials 1997;18:557–60.

6. Food and Drug Administration. Multiple Endpoints in Clinical Trials Guidance for Industry | FDA. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry. Accessed September 23, 2020.

7. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Stat Med 2009;28:586–604.

8. ICMJE. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly work in Medical Journals. Available at: http://www.icmje.org/recommendations/. Accessed December 21, 2020.

9. Kato ET, Silverman MG, Mosenzon O, et al. Effect of Dapagliflozin on Heart Failure and Mortality in Type 2 Diabetes Mellitus. Circulation 2019;139:2528–36.

10. McMurray JJV, Solomon SD, Inzucchi SE, et al. Dapagliflozin in patients with heart failure and reduced ejection fraction. N Engl J Med 2019;381:1995–2008.

11. Stone GW, Sabik JF, Serruys PW, et al. Everolimus-Eluting Stents or Bypass Surgery for

Left Main Coronary Artery Disease. N Engl J Med 2016;375:2223–35.

12. Stone GW, Pieter Kappetein A, Sabik JF, et al. Five-year outcomes after PCI or CABG for left main coronary disease. N Engl J Med 2019;381:1820–30.

13. Gabriel Steg P, Bhatt DL, Simon T, et al. Ticagrelor in patients with stable coronary disease and diabetes. N Engl J Med 2019;381:1309–20.

14. Bhatt DL, Steg PG, Mehta SR, et al. Ticagrelor in patients with diabetes and stable coronary artery disease with a history of previous percutaneous coronary intervention (THEMIS-PCI): a phase 3, placebo-controlled, randomised trial. Lancet 2019;394:1169–80.

15. Packer DL, Mark DB, Robb RA, et al. Effect of Catheter Ablation vs Antiarrhythmic Drug Therapy on Mortality, Stroke, Bleeding, and Cardiac Arrest among Patients with Atrial Fibrillation: The CABANA Randomized Clinical Trial. J Am Med Assoc. 2019;321:1261–74.

16. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. N. Engl. J. Med. 2007;357:2189–2194.

17. Solomon SD, McMurray JJV, Anand IS, et al. Angiotensin–Neprilysin Inhibition in Heart Failure with Preserved Ejection Fraction. N Engl J Med 2019;381:1609–20.

18. Thuijs DJFM, Kappetein AP, Serruys PW, et al. Percutaneous coronary intervention versus coronary artery bypass grafting in patients with three-vessel or left main coronary artery disease: 10-year follow-up of the multicentre randomised controlled SYNTAX trial. Lancet 2019;394:1325–34.

19. Pocock SJ, Clayton TC, Stone GW. Design of Major Randomized Trials. J Am Coll Cardiol 2015;66:2757–66.

20. Rockhold F, Segreti T. Secondary Efficacy Endpoints. Wiley Encycl. Clin. Trials 2007.

21. Stolker JM, Spertus JA, Cohen DJ, et al. Rethinking composite end points in clinical trials insights from patients and trialists. Circulation 2014;130:1254–61.

22. Freemantle N, Calvert MJ. Interpreting composite outcomes in trials. BMJ 2010;341:354.

23. Pocock SJ, McMurray JJ V, Collier TJ. Statistical Controversies in Reporting of Clinical Trials: Part 2 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol 2015;66:2648–62.

24. Maron DJ, Hochman JS, Reynolds HR, et al. Initial Invasive or Conservative Strategy for Stable Coronary Disease. N Engl J Med 2020;382:1395–1407.

25. Redfors B, Gregson J, Crowley A, et al. The win ratio approach for composite endpoints: practical guidance based on previous experience. Eur Heart J 2020 Sep 9 [E-pub ahead of print]; doi: 10.1093/eurheartj/ehaa665.

26. Claggett B, Pocock S, Wei LJ, Pfeffer MA, McMurray JJV, Solomon SD. Comparison of time-to-first event and recurrent-event methods in randomized clinical trials. Circulation 2018;138:570–77.

27. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–02.

28. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat 1979;6:65–70.

29. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "$p < 0.05$." Am Stat 2019;73:1–19.

30. Ioannidis JPA. The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance. J Am Med Assoc 2019;321:2067–68.

31. Cook JA, Fergusson DA, Ford I, et al. There is still a place for significance testing in clinical trials. Clin Trials 2019;16:223–24.

32. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305–07.

33. Sterne JAC, Smith GD, Cox DR. Sifting the evidence—what's wrong with significance

tests? BMJ 2001;322:226.

34. Pocock SJ, Ware JH. Translating statistical findings into plain English. Lancet 2009;373:1926–28.

35. Lin R, Yin G. Bayes factor and posterior probability: Complementary statistical evidence to p-value. Contemp. Clin Trials 2015;44:33–35.

36. Hung HM, O'Neill RT, Bauer P, Köhne K. The behavior of the P-value when the alternative hypothesis is true. Biometrics 1997;53:11–22.

37. Harrington D, D'Agostino RB, Gatsonis C, et al. New guidelines for statistical reporting in the journal. N Engl J Med 2019;381:285–86.

38. The Lancet, Randomised trials in The Lancet: formatting guidelines. Available at: https://els-jbs-prod-cdn.jbs.elsevierhealth.com/pb/assets/raw/Lancet/authors/RCTguidelines.pdf. Accessed December 10, 2020.

39. Bauchner H, Golub RM, Fontanarosa PB. Reporting and Interpretation of Randomized Clinical Trials. J Am Med Assoc 2019;322:732–735.

40. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332.

41. Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ. Robust misinterpretation of confidence intervals. Psychon Bull Rev 2014;21:1157–64.

42. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147:573–7.

43. Phillips R, Hazell L, Sauzet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open 2019;9:e024537.

44. Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges

and opportunities. Trials 2012;13:138.

CENTRAL ILLUSTRATION: Strategies for Reporting of Primary and Secondary Outcomes from Randomized Controlled Trials

Caption: Central Illustration summarising the main survey findings and related recommendations for the choice, analysis, reporting and interpretation of primary and secondary outcomes in randomized controlled trials.

RCTs, randomized controlled trials; SAP, statistical analysis plan

Table 1. Profile of 84 Cardiovascular Randomized Controlled Trials Published in 3 Major Medical Journals During 2019

	Number of trials		Number of trials
Journal		**Median (or Fixed) Follow-up Times**	
NEJM	41	<3 months	9
Lancet	23	3 months - <6 months	15
JAMA	20	6 months - <1 year	6
		1 year precisely	22
Type of Intervention		>1 year - <2.5 years	11
Pharmaceutical	43	2.5 years - <5 years	12
Device	21	5 years - <10 years	7
Management strategy	18	≥10 years	2
Surgery	2		
		Disease Conditions	
Source of Funding		Coronary artery disease	21
Industry	41	Heart failure	11
Public	35	Stroke	9
Both	8	Diabetes	7
		Atrial fibrillation	7
Participants Randomized		Primary and/or secondary prevention	7
<200	7	Aortic stenosis	6
200 - <500	12	Hypertension	5
500 - <1000	13	Cardiac arrest	4
1000 - <2000	16	Other	7
2000 - <5000	19		
5000 - <10,000	11		
≥10,000	6		

Table 2. Findings About the Primary Outcomes in the 84 Randomized Trials Surveyed

	Number of trials
Number of Primary Outcomes	
Single primary outcome	75 (89%)
Two co-primary outcomes	9 (11%)
Type of Primary Outcome (75 trials with a single outcome only)	
Composite endpoint	37 (49%)
Disease event or binary outcome	20 (27%)
Quantitative measure	12 (16%)
All-cause death	4 (5%)
Ordinal measure	2 (3%)
Primary Hypothesis (75 trials with a single outcome only)	
Superiority	65 (87%)
Non-Inferiority	10 (13%)
Level of Statistical Significance (65 superiority trials only)	
$P \geq 0.05$ (i.e. "not significant")	27 (42%)
$P < 0.05$ and ≥ 0.01	18 (28%)
$P < 0.01$ and ≥ 0.001	5 (8%)
$P < 0.001$	15 (23%)

Table 3. Components of the Composite Primary (or Co-Primary) Endpoints

Number of Components	**Number of Composites**
Two	11
Three	20
Four	5
Five	6
Seven	1
Eight	1
Type of Component	
All-cause death	23
Cardiovascular death[+]	24
Myocardial infarction	29
Stroke	27
Revascularisation	10
Heart failure hospitalisation	6
Unstable angina	4
Major bleed	4
Hospitalisation	3
Other	21
Most Common Composites	
Death*, myocardial infarction, stroke	11
Death*, myocardial infarction, revascularisation	7
Death*, heart failure hospitalization	5

* of which 16 were cardiovascular death and 7 were all-cause death

[+] of which 4 were cardiac death

Table 4. Findings from 18 Randomized Trials with a Pre-Declared Hierarchy of Secondary Outcomes

Funding	**Number of Trials**
Industry	15
Public	3
Number of Outcomes in the Hierarchy	
One	3
Two	3
Three	2
Four	1
Five	4
Six	1
Seven	1
Eight	1
Nine	2
Results of the Hierarchical Testing	
Hierarchy not entered, i.e. primary outcome not significant	7*
First component in hierarchy not "significant"	3[+]
Only first component in hierarchy significant	2
Three or more components in hierarchy significant	6**

* in three of these "negative" trials there were very highly significant findings for secondary outcomes (P<0.001)

[+] in one trial others in the hierarchy were significant (P<00.05)

** in one trial the significantly lower all-cause mortality (P<0.05) was too far down the hierarchy to be claimed

Table 5. Findings from 63 Randomized Trials with Secondary Outcomes which were not in a Hierarchy or Otherwise Corrected for Multiplicity

Number of Secondary Outcomes	**Number of Trials**
1 to 3	13 (21%)
4 to 6	18 (29%)
7 to 10	9 (14%)
11 to 15	11 (17%)
16 to 20	8 (13%)
>20	4 (6%)

The Most Significant Secondary Outcome Classified by Significance of the Primary Outcome

	Most Significant Secondary Outcome			
Primary Outcome	Not significant	P<0.05 and ≥0.01	P<0.01 and ≥0.001	P<0.001
Non-inferiority	8	**2**	**-**	**-**
Not significant	11	**3**	**3**	**4**
P<0.05 and ≥0.01	1	4	**6**	**4**
P<0.01 and ≥0.001	**-**	1	2	**-**
P<0.001	**-**	3	**-**	11

Table 6. Relationships Between the Article Conclusions and the Outcomes of the 84 Randomized Trials*

Conclusions based on:	Number of articles
The primary outcome only	60 (71%)[+]
Also include one or more secondary outcomes	19 (23%)**
Also include safety outcome(s)	6 (7%)[++]

Articles have Conclusions in the Abstract and at the end of the Discussion. The latter is sometimes a bit longer.

* In 4 articles the Conclusion's wording is imprecise, but it is implicitly referring to the primary outcome.

[+] All 20 articles in JAMA only referred to the Primary Outcome in the Conclusions, which may thus reflect an explicit or implicit journal policy. In 2 articles, subgroup findings for the Primary Outcome were in the Conclusions.

** In 5 of these, reference to Secondary Outcomes only occurred in the Conclusions in the Discussion, not the Abstract. In one case, the Secondary claim was not statistically significant. In 5 cases, the secondary claims were from a pre-defined hierarchy or multiple testing procedure.

[++] This was mostly based on a key pre-defined safety outcome. One article's conclusions referred to both safety and secondary outcomes.

CENTRAL ILLUSTRATION: Strategies for Reporting of Primary and Secondary Outcomes from Randomized Controlled Trials

TOPIC	SURVEY FINDINGS	RECOMMENDATIONS
Choice of Primary Outcome	Single primary outcome in 89% of RCTs, mostly disease-related clinical events/death	Choose the most clinically meaningful primary outcome that can also be well powered with reasonable assumptions
Co-Primary Outcomes	Co-primary outcomes in 11% of RCTs, analyzed by a variety of statistical strategies	Greater clarity and consistency is needed in analysing co-primary outcomes
Composite Outcomes	Composite primary outcome in 50% of RCTs; 3 components were a common choice	Confine the composite outcome to key components; check results for consistency between the components
Choice of Secondary Outcomes	Great diversity in the number and type of secondary outcomes	Choice of key secondary outcomes needs attention; reports must be consistent with the protocol and SAP (although multiple publications may be required)
Lack of Control for Multiple Testing	75% of RCTS paid no regard to multiple testing of secondary outcomes	RCTs should include a clear statistical strategy for testing of secondary outcomes to control multiplicity, or otherwise clearly describe these as tertiary, exploratory outcomes
Hierarchical Testing of Secondary Outcomes	21% of RCTs had a pre-defined hierarchy (mostly industry-sponsored trials)	Choice of hierarchy may be arbitrary. Consider other methods if only 2 or 3 powered secondary endpoints. A hierarchy is acceptable for Sequentially testing a longer list.
Methods for Multiplicity Correction	Only 4% of RCTs had multiplicity correction across key secondary outcomes other than a hierarchy	Such methods, especially those of Bretz and Holm, should be more widely used
Interpretation of P-values	P<0.05 was commonly used as a decision-making tool: trials were crudely classified as "positive" or "negative"	This is a misuse of P-values. A graded interpretation is required, wherein the smaller is P the stronger the evidence
New Journal Policies	NEJM only permits P-values for outcomes with strict pre-specified type-I error control	P-values (when interpreted correctly) and effect sizes should be presented for all meaningful outcomes. Nominal P-values and effect sizes outside formal testing could be downgraded (e.g. in fainter type) and footnoted with a caveat cautioning against inferential conclusions
Strength of Evidence	For the primary outcome P<0.05 occurred in 58% of superiority trials, with P<0.001 in 23% of trials	P just <0.05 but >0.01 is at best modest evidence of efficacy. P<0.001 is strong evidence (i.e. proof beyond reasonable doubt)
Article's Conclusions	Most RCTs confine the Conclusions to the primary outcome, relying on P<0.05 for a positive statement	Conclusions need to be more nuanced, taking into account the totality of evidence from the RCT and not dependent on the "magic cut-off" of 5% significance for the primary outcome alone

RCTs, randomized controlled trials; SAP, statistical analysis plan