BMJ

RESEARCH

Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review

Gloria Cordoba, researcher,[1] Lisa Schwartz, professor,[2] Steven Woloshin, professor,[2] Harold Bae, researcher,[2] Peter C Gøtzsche, professor[1]

[1]Nordic Cochrane Centre, Rigshospitalet and University of Copenhagen, Blegdamsvej 9, DK-2100 Copenhagen, Denmark

[2]Dartmouth Institute for Health Policy, Dartmouth Medical School, New Hampshire, USA

Correspondence to: P C Gøtzsche pcg@cochrane.dk

Cite this as: *BMJ* 2010;341:c3920
doi:10.1136/bmj.c3920

ABSTRACT

Objective To study how composite outcomes, which have combined several components into a single measure, are defined, reported, and interpreted.

Design Systematic review of parallel group randomised clinical trials published in 2008 reporting a binary composite outcome. Two independent observers extracted the data using a standardised data sheet, and two other observers, blinded to the results, selected the most important component.

Results Of 40 included trials, 29 (73%) were about cardiovascular topics and 24 (60%) were entirely or partly industry funded. Composite outcomes had a median of three components (range 2–9). Death or cardiovascular death was the most important component in 33 trials (83%). Only one trial provided a good rationale for the choice of components. We judged that the components were not of similar importance in 28 trials (70%); in 20 of these, death was combined with hospital admission. Other major problems were change in the definition of the composite outcome between the abstract, methods, and results sections (13 trials); missing, ambiguous, or uninterpretable data (9 trials); and post hoc construction of composite outcomes (4 trials). Only 24 trials (60%) provided reliable estimates for both the composite and its components, and only six trials (15%) had components of similar, or possibly similar, clinical importance and provided reliable estimates. In 11 of 16 trials with a statistically significant composite, the abstract conclusion falsely implied that the effect applied also to the most important component.

Conclusions The use of composite outcomes in trials is problematic. Components are often unreasonably combined, inconsistently defined, and inadequately reported. These problems will leave many readers confused, often with an exaggerated perception of how well interventions work.

INTRODUCTION

A composite outcome consists of two or more component outcomes. Patients who have experienced any one of the events specified by the components are considered to have experienced the composite outcome.[1] The main advantages supporting the use of a composite outcome are that it increases statistical efficiency because of higher event rates, which reduces sample size requirement, costs, and time; it helps investigators avoid an arbitrary choice between several important outcomes that refer to the same disease process; and it is a means of assessing the effectiveness of a patient reported outcome that addresses more than one aspect of the patient's health status.[1-6]

Unfortunately, composite outcomes can be misleading. This is especially true when treatment effects vary across components with very different clinical importance.[7] For example, suppose a drug leads to a large reduction in a composite outcome of "death or chest pain." This finding could mean that the drug resulted in fewer deaths and less chest pain. But it is also possible that the composite was driven entirely by a reduction in chest pain with no change, or even an increase, in death.

Studies show that treatment effects often vary, and typically, the effect is smallest for the most important component and biggest for the less important components.[3 5 8] Unless authors clearly present data for all components and take care in how they discuss composite findings, it is easy for readers to assume mistakenly that the treatment effect applies to all components. In this study, we systematically examined how composite outcomes were used and how well they were reported in recent randomised trials.

METHODS

We performed a systematic review of parallel group randomised clinical trials published in 2008 that had a primary composite outcome. We excluded studies where the composite was a secondary outcome measure and studies with more than two arms.

Search strategy

An iterative search strategy was developed, using various combinations of search terms and refining them based on the initial collection of trials. Furthermore, we identified relevant terms from a previous review of cardiovascular trials published between 2000 and 2006,[8] where the authors had hand searched 14 major journals. The final PubMed search was done on 26 January 2009. We limited the articles to those published in 2008 and combined "random*" with one or more of 31

Table 1 | Characteristics of 40 trials published in 2008 reporting composite outcomes

	No (%) of trials
Clinical area:	
Cardiovascular	29 (73)
Nephrology	3 (8)
Gynaecology	2 (5)
Other	6 (15)
Journal:	
New England Journal of Medicine	6 (15)
JAMA	4 (10)
American Heart Journal	3 (8)
Lancet	3 (8)
BMJ	2 (5)
Circulation	2 (5)
Transplantation	2 (5)
Other	18 (45)
Funding:	
Industry funding	16 (40)
Partly industry funding	8 (20)
No industry funding	7 (18)
Unclear	9 (23)

search terms (see webtable 1 on bmj.com). We dropped two additional terms, "composed of"[tiab] and "combination of"[tiab], as these were too unspecific, yielding 6255 and 14 633 hits, respectively, when combined with "random*."

Study selection and data extraction

The abstracts were reviewed by one person (GC), and potentially eligible articles were retrieved in full and assessed independently by two coders (GC and HB). Disagreements were resolved by discussion, and for ambiguous cases the other authors were involved. The two coders used a standard form to extract data independently and collected data on journals, clinical area, composite outcome and its components, and source of funding.

One composite outcome was included per article. When more than one such outcome was reported in an article, we used a hierarchical selection process of (*a*) authors' explicit declaration of primacy, (*b*) the composite outcome used to calculate the sample size, (*c*) authors' attribution of importance to the composite outcome in their description of the results, or (*d*) the composite outcome that appeared first in the methods section.

Content analysis

Two pairs of independent observers used standardised protocols to assess the definition and quality of reporting of the composite outcomes. Most judgments involved assessments of facts (such as whether the number of components making up the composite changed within the paper). Here, disagreement was almost entirely due to oversight, not a difference in opinion. For the few subjective judgments, we created simple and explicit rules to objectify the process as much as possible. For example, we judged the conclusion of abstracts as falsely suggesting that an effect on the composite also applied to the most important component (when it did not) if all components were listed using "and" or if the composite was named as a class of events. Our rules are provided when we present results. Also, we provide examples to allow readers to decide for themselves whether our judgments were reasonable. We resolved discrepancies involving facts and disagreements by discussion.

Composite definition Two observers (PCG, and LS or SW) independently and blinded to the results selected the most important component of each composite outcome, taking into account the hierarchy for analysing composite outcomes proposed by Lubsen et al,[9] and always choosing death (or disease specific death) if such a component had been used. The observers also rated the gradient of importance for components, and looked for any discussion of the rationale for the composite.

Reporting of composite We assessed the consistency of the components of the composite between the abstract, methods, and results; determined whether data were reported for all components (that is, so that they could be used in a meta-analysis); judged whether the components were of similar importance; and evaluated whether the conclusion presented in the abstract or the discussion section suggested that the intervention was effective for all the components of the composite outcome rather than just for the composite.

Data analysis

We present descriptive statistics and used Fisher's exact test for analysis of binary data. We had planned to estimate an average inflation factor, based on a comparison of the effect for the composite outcome and that for the most clinically important outcome, but realised that this was problematic (see Discussion).[8]

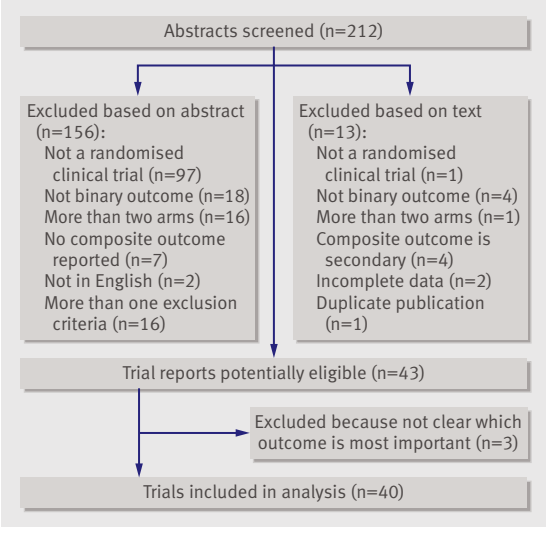

Fig 1 | Flow chart for inclusion of trials

Table 2 | Definition and reporting of composite outcomes used in 40 trials published in 2008

	No (%) of trials
Composite definition	
Most important component:	
Death (all causes or disease specific)	33 (83)
Clinical outcome (hospital admission or symptom)	7 (17)
Clinical importance of components:	
Similar	7 (18)
Might be similar	5 (13)
Not similar	28 (70)
Author discussion of composite outcome:	7 (18)
No discussion	33 (83)
Explains rationale for composite	1 (3)
Acknowledges problems with composite	6 (15)
Reporting of composite	
Components consistent between abstract, methods, and results	27 (68)
Components inconsistent:	13 (33)
Major inconsistency (components added or deleted)	5 (13)
Minor inconsistency (ambiguous wording change)	8 (20)
Data for components provided	31 (78)
Data for components not provided:	9 (23)
Missing data	2 (5)
Ambiguous data	7 (18)

RESULTS

Our searches identified 212 abstracts, 169 of which were ineligible as described in fig 1. The remaining 43 articles were potentially eligible, but we excluded three[w1–w3] because it was not clear to us which outcome was most clinically important (which needed to be identified for our reporting analysis). For example, a trial that compared two methods of vein stripping had a composite outcome that consisted of haematoma in the thigh, ecchymosis, seroma, wound healing complications, wound infections, and phlebitis.[w1]

Table 1 describes the characteristics of the 40 included trials,[w4–w43] which together randomised 110 080 patients, with a median sample size of 1486 (interquartile range 213–4460). The two most common journals of publication were the *New England Journal of Medicine* (n=6) and *JAMA* (n=4); 29 trials (73%) were on cardiovascular topics. In 24 reports (60%) it was declared that the trials were totally (n=16) or partially (n=8) industry funded, seven trials did not receive industry support, and for nine trials the funding was not clear.

Composite definition

The composite outcomes had a median of three components (range 2–9). The most important component, selected by us, was death or cardiovascular death in 33 trials (83%), clinical events (such as incontinence symptoms, respiratory distress, phlebitis, or arrhythmia) in six trials (15%), and hospital admission in one trial (3%) (table 2).

Our assessment of composite

We judged that the components were of similar importance in seven trials (18%): infiltration or phlebitis[w4]; death or chronic lung disease in preterm babies[w28]; no reflow, slow flow, and ventricular arrhythmia[w15]; death, graft loss, or acute rejection[w13 w23 w29]; and total mortality, clinical re-infarction, or disabling stroke.[w30] Five trials (13%) were questionable, as they combined death and non-fatal myocardial infarction without defining non-fatal myocardial infarction—so it might have included silent events.[w12 w17 w20 w26 w31]

In the remaining 28 trials (70%), the components were not of similar importance: 20 trials had combined death with hospital admission (or procedures that required hospital admission, such as revascularisation), and eight trials had other problems[w5 w6 w16 w19 w22 w25 w32 w33] (such as combining death and silent myocardial infarctions,[w22] combining death with new exertional angina and transient ischaemic attack,[w16] or combining death with a doubling of serum creatinine concentration from baseline[w33]).

Author discussion of composite

Seven trial reports (18%) included a discussion related to the rationale for the composite. Only one report, about intravenous catheters, provided a rationale supporting the construction of the composite: "It has been argued that infiltration (easy to diagnose) may result from unrecognised phlebitic changes to the vein wall (hard to diagnose) leading to under-reporting of phlebitis. It is perhaps more useful to use the composite measure of infiltration or phlebitis as it avoids any potential for misdiagnosis."[w4] The other six reports only mentioned problems with the composite: three noted that the components did not have similar clinical importance,[w6–w8] one that the composite had not been validated for clinical relevance,[w5] one that the composite was driven by the procedural outcome,[w9] and one was problematic because one of its five components (myocardial infarction) favoured one drug and another component (bleeding) favoured the other drug.[w10]

Other definition problems

In four trials (10%), the trial authors explicitly stated that they created the composite post hoc.[w22 w25–w27] In three cases, the prespecified composite was not statistically significant, but the new, post hoc composite was, suggesting cherry picking (see examples in box).

Reporting of composites

Inconsistent reporting of components

In 13 reports (33%), the definition of the composite outcome changed between the abstract, methods, and results sections. For eight trials,[w12 w14 w17–w22] the reporting problem was minor, involving inconsistent use of modifiers—for example, whether a myocardial infarction[w12 w17 w18 w20] or a stroke[w22] was lethal, whether deaths referred to those from all causes or from specific disease,[w14 w21] and reversal of the scale (a positive stress test was later reported as a negative stress test[w19]).

For five trials, the inconsistency was major, as the components were not the same throughout the trial report.[w7 w8 w16 w23 w24] For example, in one trial, death was added as a new component.[w7] In another trial,

Selected examples of problems in defining composite outcomes

Composite outcome was not prespecified (cherry picking)[w25]

Abstract—"a composite end point of myocardial infarction, stroke, or sudden death."

Text (methods section)—Describes a different composite: "a composite cardiovascular end point consisting of sudden death, myocardial infarction, angina, or chest pain." It adds another one: "We wrote a more detailed protocol before the analysis of vascular events, and this added an additional composite end point of myocardial infarction, stroke, or sudden death, as this has become the commonly used end point in such analyses."

Our comment—The abstract made no mention of the prespecified composite. The additional composite was reported, but its post hoc nature was not mentioned. Of note, the prespecified composite was not statistically significant (P=0.68), but the post hoc composite was (P=0.008).

Inconsistent definition throughout the paper[w8]

Methods section—"A composite including bone-grafting, implant exchange or removal because of a broken nail or deep infection, and debridement of bone and soft tissue because of deep infection, dynamization of the fracture (i.e., interlocking screw removal to allow fracture-site compression with weight bearing) in the operating room or in the outpatient clinic; removal of locking screws because of hardware breakage or loosening; autodynamization (spontaneous screw breakage leading to dynamization at the fracture site prior to healing); fasciotomy; and drainage of hematomas."

Results section—"Bone-grafting in a patient with full cortical continuity, implant exchange for union in a patient with full cortical continuity, implant removal for union in a patient with full cortical continuity, reoperation in response to a local infection, bone-grafting in a patient with a fracture gap of <1 cm, implant exchange in a patient with a fracture gap of <1 cm, implant removal in a patient with a fracture gap of <1 cm, dynamization in a patient with a fracture gap of <1 cm, removal of locking screws due to hardware breakage or loosening of screws, treatment of wound necrosis in the presence of infection, fasciotomy for the treatment of intraoperative compartment syndrome, fasciotomy for the treatment of postoperative compartment syndrome, autodynamization (failure of the screw-bone construct [i.e., broken or bent screws] that dynamizes the fracture), draining of a hematoma, failure of the construct (broken nail)."

Our comment—Change in the number of components (from 8 in methods section to 11 in results section) is a major inconsistency.

about whether corticosteroids could be stopped early after renal transplantation,[w23] the abstract concluded there was "no evidence of an increased risk of poorer performance" (based on 1 *v* 0 severe acute rejections). But, using the definition in the methods and data in a table, we found an increased risk of rejection (14 *v* 6 acute rejections, P=0.06). In a third trial[w16] the results table omitted data for two components, sudden death and newly developed exertional angina, while the table provided data for an outcome not mentioned in the definition of the composite, stable angina.

Missing data for components

In two cardiovascular trials, data on the most important component were missing. In one,[w14] two of us tried to calculate deaths from cardiovascular causes from the categories presented in a table, but we arrived at two different answers and cannot determine which set of numbers, if either, is correct (fig 2). The other trial provided a table with all the components,[w7] but, as noted in the trial report, only those events that occurred first were tabulated. It was therefore not possible to see how many patients died, as only those deaths that occurred before any other events (such as gastrointestinal, eye, or skin complications) were reported.

In three other cardiovascular trials, the number of events for the components added up exactly to the number of composite events (see webtable 2 on bmj.com). The reports provided no way of knowing whether these data reflected only the first events (as above) or that no patient had more than one event.[w15–w17] We believe that only first events were reported, as it is implausible, for example, that no one had angina or a transient ischaemic attack before dying from cardiovascular causes.[w16]

In another four trials, numerical data could not be extracted. In one trial, the authors reported 31 "combined events" in a group with only 29 patients.[w11] In another trial, there were vastly more events in the component outcomes than in the composite outcome (an impossibility since by definition patients experience the composite if they experience any of the components).[w5] In the third trial, the number of components increased from three to eight after an interim analysis showed fewer events than anticipated, but we could not figure out what the composite was, as the reporting was inconsistent.[w8] In the fourth trial,[w13] the data were given as percentages, which led to inconsistencies: 11 versus 12 died according to the percentages but 11 versus 14 according to a table, and graft losses were 23 versus 22 from the percentages but 15 versus 15 in the table.[w13]

Problems with reporting the role of chance

Consistent with problems about how clinical trials are reported in general,[10] we found errors in the P values reported. One of the trials reported the composite was statistically significant (P=0.037)[w15] when it was not (P=0.09 according to our calculation). In another trial, there was an error in the opposite direction: the authors reported that the most important outcome was not significant (P=0.192)[w33] when in fact it was; the intervention was harmful, as it increased mortality significantly (P=0.046, our calculation). Confidence intervals for the components were not reported in 22 trials (55%).

Inadequate interpretation

In 22 cases (55%), the conclusions of the abstract or the discussion did not remind readers that the outcome was a composite, and 33 conclusions (82%) did not specifically say if there was—or was not—a similar effect on the most important component (see examples in fig 3). Statistically significant results were reported in three trials for the most important component (death or cardiovascular death), in one trial for both the most important component and the composite outcome (but in opposite directions, as the effect was beneficial for the composite of death or non-fatal myocardial infarction and harmful for death[w12]), and in 16 trials for the composite outcome only. In 11 of these 16 trials, the abstract conclusions falsely implied the effect applied also to the most important component: two listed all components of the composite using "and" (see webtable 3 on bmj.com), and nine referred to the composite as a class of events (for example, "reduced the incidence of major cardiovascular events"[w14]).

The composite outcome consisted of myocardial infarction, stroke, arterial revascularisation, hospitalisation for unstable angina, or death from cardiovascular causes, but deaths from cardiovascular causes were not presented (in the table or the article)

How data were presented:

End point	Rosuvastatin (n=8901)	Placebo (n=8901)
Primary end point	142	251
Non-fatal myocardial infarction	22	62
Any myocardial infarction	31	68
Non-fatal stroke	30	58
Any stroke	33	64
Arterial revascularisation	76	143
Hospitalisation for unstable angina	16	27
Arterial revascularisation or hospitalisation for unstable angina	76	143
Myocardial infarction, stroke, or confirmed death from cardiovascular causes	83	157
Death on known date	190	235
Any death	198	247

Question: How many deaths were there from cardiovascular causes? (try for yourself before going to the possible solutions)

Possible solutions:

Solution 1	Rosuvastatin	Placebo
Myocardial infarction, stroke, or confirmed death from cardiovascular causes	83	157
minus Non-fatal myocardial infarction	-22	-62
minus Non-fatal stroke	-30	-58
Death from cardiovascular causes	31	37

Solution 2	Rosuvastatin	Placebo
Any myocardial infarction	31	68
minus Non-fatal myocardial infarction	-22	-62
plus Any stroke	+33	+64
minus Non-fatal stroke	-30	-58
Death from cardiovascular causes	12	12

Fig 2 | **Example of a confusing presentation of a composite outcome**[w14]

Overall evaluation

Accounting for inconsistencies in definition of components and in reported numbers, only 24 of the 40 trials (60%) provided reliable estimates for both the composite and its components. Of the 12 trials that had components of similar, or possibly similar, clinical importance, only six trials provided reliable estimates.[w4 w26 w28–w31]

DISCUSSION

Trials with composite outcomes are often problematic, characterised by a lack of logic behind the construction of the composites, inconsistent and unclear reporting, post hoc changes to the composites, and cherry picking of favourable outcomes or combinations of outcomes. Guidance for authors aimed at ensuring that the components are appropriate and avoid misleading results and statements[1 3 5-9] have existed for years but seem to have had little effect on the trials we examined, which were from 2008.

Composite outcomes create a substantial opportunity for post hoc changes. In a cohort of 102 trial protocols and subsequent publications, changes to at least one primary outcome had occurred in 63% of the trials, and not in a single case had the report acknowledged the modification.[11] It is therefore likely that many of the composite outcomes we studied, which were all primary outcomes, had been modified post hoc without acknowledging this. In fact, a survey of cardiovascular trials showed marked asymmetry in the distribution of P values around P=0.05, suggesting possible publication bias or that individual outcomes were selected for inclusion in the composite to ensure statistical significance.[8]

Because components can be combined in so many ways, it is easy to find significant results. In one of the trials we included,[w16] the composite consisted of eight cardiovascular end points, but there were also secondary composites that consisted of "combinations of primary end points as well as death from any cause." These combinations were not specified, but nine end points can be combined, as two or more components, in 502 possible ways (2^9-1(empty sample)-9(samples with only one component)). The result for the composite was not statistically significant, but the abstract noted that the hazard ratio was 0.10 for a combined end point of fatal coronary events and fatal cerebrovascular events (P=0.0037)—that is, a cherry picked result. One would expect 25 of 502 possible combinations to be significant purely by chance.

We found other examples of cherry picking. A trial of percutaneous coronary intervention had four components in the composite (death, myocardial infarction, urgent revascularisation of target vessel, and major bleeding), but the relative risk and the confidence interval were shown only for major bleeding, where the experimental drug had an advantage, and the last sentence in the conclusions in the abstract was: "it did significantly reduce the incidence of major bleeding."[w10]

We also encountered the most ingenious way of getting rid of dead patients that we have ever seen.[w7] Deaths in a cardiovascular trial were listed only if they occurred before anything else. Thus, one might avoid deaths by including a component that precedes death, such as chest pain.

It is also problematic that death was so commonly included in composites, as it provides the lowest event rates and the smallest treatment effects.[5] Furthermore, death can mean many things. It was total mortality in seven trials, some form of cardiovascular mortality in 17 trials, death with no further specification in seven trials, and sudden death in one trial. Since total mortality is the only mortality outcome that is guaranteed free from bias, we suggest that cardiovascular trialists use this outcome. A particularly revealing example of data dredging is the Anturane reinfarction trial.[12] After publication of positive results, researchers at the US Food and Drug Administration found that the trial's classification of cause of death was "hopelessly unreliable."[13] Cardiac deaths were classified into three groups—sudden deaths, myocardial infarction, or other cardiac event—and nearly all the errors in assigning cause of death favoured the

Remind readers that result is based on a composite outcome

Do **Conclusion:** "...a combination pill ... did not reduce a combined end point of total cardiovascular events among high-risk women."[w18]

Don't **Conclusion:** "...rosuvastatin significantly reduced the incidence of major cardiovascular events."[w14]

Our comment: Can mislead readers to believe that all components are equally important, even though there was an important gradient, from hospitalisation to cardiovascular death.

Report data for all components

Do Prespecified primary and secondary outcomes and death[w41]

	Placebo (n=929)	Simvastatin plus ezetimibe (n=8901)	Hazard ratio (95% CI)	P value
Primary outcome				
Patients with any event (could have >1 event)	355 (38.2)	333 (35.3)	0.96 (0.83 to 1.12)	0.59
Death from cardiovascular causes	56 (6.0)	47 (5.0)	0.83 (0.56 to 1.22)	0.34
Aortic valve replacement surgery	278 (29.9)	267 (28.3)	1.00 (0.56 to 1.22)	0.97
Congestive heart failure as a result of progression of aortic stenosis	23 (2.5)	25 (2.6)	1.09 (0.62 to 1.92)	0.77
Non-fatal myocardial infarction	26 (2.8)	17 (1.8)	0.64 (0.35 to 1.17)	0.15
Coronary-artery bypass grafting	100 (10.8)	69 (7.3)	0.68 (0.50 to 0.93)	0.02
Percutaneous coronary intervention	17 (1.8)	8 (0.8)	0.46 (0.20 to 1.06)	NA
Hospitalisation for unstable angina	8 (0.9)	5 (0.5)	0.61 (0.20 to 1.86)	NA
Non-haemorrhagic stroke	29 (3.1)	25 (2.6)	1.12 (0.68 to 1.87)	0.65

State whether the intervention has a similar effect on all components, or specify on which components there is an effect (specifically mentioning the most important component)

Do "There was no evidence that this treatment strategy increased mortality. Intensive glucose control significantly reduced the primary composite outcome of major macrovascular or microvascular events, mainly as a consequence of a reduction in nephropathy. There was no separately significant reduction in major macrovascular events, although a modest benefit could not be ruled out."[w36]

Don't PCI data and laboratory findings of the 73 patients

	Nicorandil (n=37)	Control (n=36)	P value
Composite endpoint:	2 (5.4%)	8 (22.2%)	0.037
No-reflow	1 (2.7%)	2 (5.2%)	0.538
Slow-flow	1 (2.7%)	4 (10.4%)	0.155
Ventricular arrhythmia	0	2 (5.2%)	0.146

Conclusion: "administration of intracoronary nicorandil reduced the occurrence of no-reflow, slow reflow, and reperfusion arrhythmia."[w15]

Our comment: Misleads readers to believe the effect is present, and the same, for all components. P value for composite is also wrong (see text)

Highlight inherent problems associated with composite outcomes

Do "A composite endpoint of grade 1–3 ongoing pain and either grade 3–4 induration (≥25 mm) or grade 2–4 nodules/cysts (>20 mm). The composite endpoint has not been validated for clinical relevance."[w5]

Fig 3 | Selected examples of composite outcomes being handled well ("Do") or poorly ("Don't")

conclusion that sulfinpyrazone decreased sudden death, the major finding of the trial.

The inclusion of clinician driven outcomes in the composite, such as admission to hospital, is problematic because they are far less important than dying and because they are highly vulnerable to bias in non-blinded trials. Nine of the 20 trials that had used hospital admission were not blinded for the clinicians.[w7 w11 w21 w27 w34–w38] Another survey showed that the inclusion of a clinician driven outcome was predictive of a statistically significant result for the primary composite outcome (odds ratio 2.24 (95% confidence interval 1.15 to 4.34), P=0.02).[3]

In addition to these problems, which we found equally often in the best general medical journals[w7 w10 w14 w16 w18 w20 w25] as in specialty journals, it is commonly difficult to explain what an effect on a composite outcome really means. This is particularly so when the effect on the composite outcome and on the most important single outcome go in different directions, as in the trial where the drug significantly decreased the composite end point of non-fatal myocardial infarction and death but increased significantly the number of deaths.[w12] A hypothetical conversation may illustrate the challenge:

"Mr Smith, here is a drug that will reduce your combined risk of getting a heart attack that will not kill you, or of dying."
"Doctor, I am not sure I quite understood you, but please give me this drug."
"But I should also mention that the drug will increase your risk of dying."
"Didn't you just tell me that the drug would decrease my risk of dying? I am confused."

Limitations of study

As we aimed at providing a general picture of the use of composite outcomes, we included all clinical areas. Because we relied on electronic searches, it is possible that hand searching journal articles would have yielded more trials. Most of the trials we identified were on cardiovascular topics, which is partly because composites are so common in this area and partly because all terms in our search strategy contained either "composite" or "combined" (see webtable 1 on bmj.com). For some diseases, composites may not be described as such. In cancer trials, for example, it is common to use disease-free survival, which means that the patients neither had tumour recurrence nor died. Such composites can be misleading, as some treatments reduce the risk of tumour recurrence while increasing the risk of death—for example, radiotherapy given to low risk patients such as women who had their breast cancer detected at screening.[14] Another example is HIV infection, where it is common to use a composite of death or time to first AIDS defining event. It would therefore be interesting to perform studies of composite outcomes in other disease areas.

We had planned to estimate an average "inflation" factor, comparing the effect for the composite with that for the most clinically important outcome, but it is not straightforward how one should analyse the data.[8 15] The observations are not independent, as the most important outcome contributes to the composite, and ratios between relative risks are very unstable when the denominator is close to zero (division almost by zero).[15] It is therefore not feasible to compare results within trials before pooling in a meta-analysis.

WHAT IS ALREADY KNOWN ON THIS TOPIC

- When trial results are reported as a composite outcome, the effect is often smallest for the most important component of the outcome and biggest for the less important components
- Clinician driven outcomes are predictive of a statistically significant result for the composite outcome
- Individual outcomes may be selected for inclusion in the composite to ensure statistical significance

WHAT THIS STUDY ADDS

- Changes in the definition of composite outcomes during a trial are common and suggest biased reporting
- Pivotal data are often missing, ambiguous, or uninterpretable

Implications of study results

For trialists—Composite outcomes should generally be avoided, as their use leads to much confusion and bias. If composites are used, trialists should follow published guidance[1 3 5-9]: only combine components of similar clinical importance, take care to define them consistently throughout the paper, analyse the prespecified composite, and list results for all components (not just the first occurring events) in a table with confidence intervals. Ideally, to avoid flaws in reporting and misleading perceptions about treatment effects,[9] every single combination of events should be shown in a table. Thus, for five components, there would need to be 31 (2^5-1) lines in the table of outcomes.

For meta-analysts—Meta-analysts should be careful when extracting data from trial reports with composite outcomes. We found many possibilities for data extraction errors—for example, subtle differences in wording may mean that what is being reported might not be what the meta-analyst thinks it means, or what was described in the methods section or elsewhere in the paper. Furthermore, it can be only those events that occurred first that are tabulated. Meta-analysis of composite outcomes is inappropriate, as the likelihood of cherry picking is too high; only the components should be used.

For editors—Composite outcomes are easily misunderstood by readers. Editors should insist that conclusions explicitly remind readers that the result is based on a composite outcome. To avoid misleading readers, editors should ensure that conclusions state whether the intervention had a similar effect on all components or specify on which components there was an effect, specifically mentioning the most important component (see fig 3). Finally, as the potential for post hoc changes is so large, editors should post the trial protocol and the raw data on the journal's website.

Conclusions

The use of composite outcomes in trials is problematic. Components are often unreasonably combined, inconsistently defined, and inadequately reported. These problems will leave many readers confused, often with an exaggerated perception of how well interventions work.

We thank Eric Lim and colleagues for supplying us with the included studies in their review of cardiovascular trials in an electronic format that helped us refine our search strategy.
Contributors: PCG, LS, and SW conceived and designed the study; all authors contributed to extraction, analysis and interpretation of data and drafting of the manuscript; PCG, LS, are SW are guarantors.
Funding: None.
Competing interests: None declared.
Ethical approval: Not required.
Data sharing: A full data set is available at www.cochrane.dk/research/data_archive/2010_2. These data may be used only for replication of the analyses published in this paper or for private study. Express written permission must be sought from the authors for any other data use.

1 Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *J Clin Epidemiol* 2007;60:651-7.
2 Ross S. Composite outcomes in randomized clinical trials: arguments for and against. *Am J Obstet gynecol* 2007;196:119e1-6.
3 Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA* 2003;289:2554-9.
4 Tomlinson G, Detsky AS. Composite end points in randomized trials: there is no free lunch. *JAMA* 2010;303:267-8.
5 Ferreira-Gonzalez I, Permanyer-Miralda G, Domingo-Salvany A, Busse JW, Heels-Ansdell D, Montori VM, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007;334:786.
6 FDA. Guidelines for industry: patient/reported outcome measures: use in medical product development to support labelling claims: draft guidance. *BMC Health and Quality of Life Outcomes* 2006;4:79.
7 Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse JW, Pacheco-Huergo V, Bryant D, et al. Validity of composite end points in clinical trials. *BMJ* 2005;330:594-6.
8 Lim E, Brown A, Helmy A, Mussa S, Altman DG. Composite outcomes in cardiovascular research: a survey of randomized trials. *Ann Intern Med* 2008;149:612-7.
9 Lubsen J, Kirwan BA. Combined endpoints: can we use them? *Stat Med* 2002;21:2959-70.
10 Gøtzsche PC. Believability of relative risks and odds ratios in abstracts: cross-sectional study. *BMJ* 2006;333:231-4.
11 Chan A-W, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
12 Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med* 1980;302:250-6.
13 Temple R, Pledger GW. The FDA's critique of the anturane reinfarction trial. *N Engl J Med* 1980;303:1488-92.
14 Early Breast Cancer Trialists' Collaborative Group. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. *Lancet* 2000;355:1757–70.
15 Krogsbøll LT, Hróbjartsson A, Gøtzsche PC. Spontaneous improvement in randomised clinical trials: meta-analysis of three-armed trials comparing no treatment, placebo and active intervention. *BMC Med Res Methodol* 2009;9:1.

Accepted: 29 May 2010