Bias was reduced through the removal of subjective elements from the outcome definition in an open-label randomised trial where blinded outcome assessment was not feasible

Brennan C. Kahan[a], Caroline J. Doré[b], Michael F. Murphy[c], Vipul Jairath[d,e,]

[a]Pragmatic Clinical Trials Unit, Queen Mary University, London, UK

[b]Comprehensive Clinical Trials Unit at UCL, UCL, London, UK

[c]NHS Blood and Transplant, Oxford, UK

[d] Department of Medicine, Western University and London Health Sciences Network, London, Ontario, Canada.

[f]Nuffield Department of Medicine, University of Oxford

Corresponding author:

Brennan C Kahan

b.kahan@qmul.ac.uk

58 Turner St

E1 2AB

Queen Mary, University of London

Abstract

Objective:

To determine whether modifying an outcome definition to remove subjective elements reduced bias in a trial that could not use blinded outcome assessment.

Study setting and design:

Re-analysis of an open-label trial comparing a restrictive vs. liberal transfusion strategy for gastrointestinal bleeding. The usual definition of the primary outcome, *further bleeding*, allows subjective clinical symptoms to be used alone for diagnosis, whereas the definition used in the trial required more objective confirmation by endoscopy. We compared treatment effect estimates for these two definitions.

Results:

Fewer subjective symptom-identified events were confirmed using more objective methods in the restrictive arm (18%) than in the liberal arm (56%), indicating differential assessment between arms. An analysis using all events (both subjective and more objective) led to an odds ratio of 0.83 (95% CI: 0.50-1.37). When only events confirmed using more objective methods were included, the odds ratio was 0.50 (95% CI: 0.32-0.78). The ratio of the odds ratios was 1.66, indicating that including unconfirmed events in the definition biased the treatment effect upwards by 66%.

Conclusion:

Modifying the outcome definition to exclude subjective elements substantially reduced bias. This may be a useful strategy for reducing bias in trials that cannot blind outcome assessment.

Keywords: Bias, open-label trial, blinding, outcome assessment, randomized controlled trial, cluster-randomized.

Running title: Reducing bias in open-label trials

What is new:

Key findings

- Blinded outcome assessment in clinical trials is not always feasible but including subjective elements can produce bias in estimated treatment effects.
- We modified the outcome definition of further bleeding in an open-label, cluster randomised trial of gastrointestinal bleeding to omit subjective elements, resulting in a definition using more objective measures.
- Re-analysis of the trial data-set found that the modified definition which excluded subjective elements reduced bias by up to 66%.

What this adds to what is known

- Modifying the outcome definition to exclude subjective elements can reduce bias when unblinded assessment is unavoidable.

What is the implication, what should change now

- Modifying trial endpoints to exclude subjective elements can be a useful strategy when designing open-label trials where blinded outcome assessment is not possible.

1. Background

Blinded outcome assessment is a key component of randomised controlled trials, as unblinded assessment can result in substantial bias in the estimated treatment effects [1-7]. However, blinded assessment can be difficult to achieve under some circumstances. For example, TRIGGER (Transfusion in Gastrointestinal Bleeding) was an open-label cluster-randomised trial that assessed the feasibility of implementing two red blood cell transfusion strategies (restrictive vs. liberal) in patients admitted to UK hospitals with acute upper gastrointestinal bleeding (AUGIB) [8-10]. The primary clinical outcome was an episode of further bleeding arising from the patient's upper gastrointestinal tract. In open-label trials, blinded outcome assessment can often be achieved by having a third party (e.g. another clinician at the hospital) who is unaware of treatment allocation assess the patient, or by sending the relevant information to a central (blinded) adjudication committee for assessment. However, neither of these options was feasible in TRIGGER [11]. Because cluster randomisation was used, every clinician within each trial site was aware of the treatment allocation in that hospital, and therefore having a third party assess the patient directly was impossible. Equally, it was not possible to compile relevant information in a blinded manner for review by an independent adjudication committee [11]. Therefore, assessment of further bleeding could not be done in a blinded manner in TRIGGER.

There is little guidance on methods to reduce the risk of bias when blinded outcome assessment is infeasible. One approach would be to modify the outcome to make it easier to implement blinded assessment. However, in some circumstances the only modification that is possible is to use a surrogate measure in place of a clinically important outcome. Surrogate measures are not always useful indicators of clinical benefits or harms, and may not be directly relevant to patients or clinicians [12, 13]. For example, the occurrence of a further bleeding episode in TRIGGER could have been based on biomarkers, such as a drop in a patient's haemoglobin count, which could easily have been adjudicated by a blinded committee. However, this may be a misleading surrogate for further bleeding since the count may not change during the acute phase of haemorrhage and therefore would have been of limited clinical relevance.

An alternative approach is to modify the outcome definition to make it less subjective, which can reduce bias while retaining the clinical value of the outcome measure [11]. This second approach was used in TRIGGER. The definition of further bleeding was modified from that used in routine clinical practice to exclude subjective elements. The resulting definition only included more objective measures of further bleeding [11].

We assessed whether modifying an outcome definition can reduce bias using TRIGGER as a case study. We re-analysed the TRIGGER trial to compare two outcome definitions, the usual definition, which included subjective elements, and the modified definition used in TRIGGER, which excluded subjective elements.

2. Methods

2.1 Choice of outcome measure in TRIGGER

TRIGGER (ISRCTN 85757829) was a cluster-randomised trial that compared the feasibility of implementing two red blood cell transfusion strategies for patients admitted to UK hospitals with AUGIB. In both strategies, patients received a transfusion when their haemoglobin dropped below a certain level, which was 8 g/dL under the restrictive transfusion policy and 10 g/dL under the liberal transfusion policy. It was impossible to blind the trial personnel as the intervention involved blood transfusion in an emergency setting. Our previous publication explains why a blinded outcome assessment was not feasible [11]. The standard of care for managing a patient with AUGIB is resuscitation and stabilisation, followed by direct visual inspection of the upper gastrointestinal tract with a fibre optic telescope, called an endoscopy, to identify and treat the source of bleeding.

The primary clinical outcome was further bleeding up to day 28. It was defined as either ongoing bleeding at the end of the initial endoscopy or a bleed that restarted after stopping, as per standard international consensus criteria [14]. Further bleeding can be assessed either using a patient's physical signs and symptoms, or by a visual inspection of the upper gastrointestinal tract using endoscopy. Assessment based on patient symptoms considers haemodynamic instability (e.g. low blood pressure and an increased heart rate), a drop in the haemoglobin concentration, whether the patient has vomited blood, and the passage of altered blood per rectum. In visual inspection, the patient's upper gastrointestinal tract is examined during an endoscopy to determine whether there is ongoing bleeding in the stomach.

We consider an outcome to be subjective if (a) its assessment depends on the judgement of the assessor; and (b) this judgement may be influenced by knowledge of the patient's treatment assignment [15]. Assessment of further bleeding based on patient symptoms inherently involves a degree of subjectivity. For example, signs of haemodynamic instability can be mimicked by other conditions such as sepsis, dehydration, or intercurrent illness. The clinician must judge whether the haemodynamic instability is caused by further bleeding or something else. The haemoglobin

concentration may drop due to a delayed response following the initial bleed, haemodilution after fluid infusion, another factor, or a combination. Again, the clinician must judge the cause of the concentration drop. If a patient vomits altered blood ("coffee grounds" rather than fresh red blood) or passes dark altered blood per rectum (which can often persist for up to 5 days after the initial bleed), the clinician must decide whether the blood is recent or old. Recent blood indicates a new or ongoing bleed, whereas old blood may originate from the initial bleed for which the patient attended the hospital.

In contrast, assessment of further bleeding using endoscopy is far less subjective, as it does not require the clinician to judge the "degree" of what they see. Instead, they simply determine the presence or absence of ongoing bleeding in the upper gastrointestinal tract. Assessment via endoscopy is therefore a more objective endpoint, as it should reduce the potential for ascertainment bias compared with an assessment based on clinical symptoms alone.

In clinical practice, further bleeding may initially be suspected based on clinical symptoms or may be identified directly during the patient's initial endoscopy (e.g. if there is still ongoing bleeding after a therapeutic procedure has been undertaken to try to stop the bleed). If further bleeding is suspected based on clinical symptoms, then an endoscopy is typically undertaken to confirm that the symptoms are the result of further bleeding, rather than the original bleed or another condition, and to attempt to treat the bleeding lesion.

There are therefore three possible outcomes of the assessment of a further bleeding episode:

1. ***Unconfirmed events suspected from patient symptoms:*** A further bleeding episode is suspected from patient symptoms, but is not confirmed through an endoscopy
2. ***Confirmed events suspected from patient symptoms:*** A further bleeding episode is suspected from patient symptoms and is confirmed with an endoscopy
3. ***Confirmed events identified during endoscopy:*** A further bleeding episode is identified directly during an endoscopy

Unconfirmed events suspected from patient symptoms typically occur when a subsequent endoscopy shows no evidence of further bleeding. The patient symptoms instead result from either the original bleed for which the patient presented to the hospital or some other medical condition. However, in rare cases a suspected event may not be confirmed because a follow-up endoscopy is not performed. This may occur if further intervention is deemed not to be in the patient's best

interests, such as on grounds of futility if the patient is likely to die in the near future due to co-morbidity. Confirmed events are true further bleeding events, as regardless of why they are initially suspected, they are confirmed during an endoscopy.

As TRIGGER was an open-label trial, participating clinicians may have assessed suspected further bleeding episodes differently depending on the treatment arm. For example, the clinicians may have had pre-conceptions about the effectiveness of one treatment policy over another and may therefore have been more likely to view patient symptoms such as a drop in the haemoglobin concentration or vomiting blood as indicative of a new bleeding episode, rather than an old one. This type of ascertainment bias for subjective outcomes has been recorded in a variety of situations [1-5]. In this scenario, including unconfirmed bleeding events suspected from patient symptoms (i.e. events that were not confirmed using more objective means) in the analysis would have led to biased estimates of the treatment effect. Further bleeding episodes in TRIGGER thus required more objective confirmation. The primary analysis for further bleeding only included events that were confirmed with an endoscopy, both those directly identified during an endoscopy and those that were initially suspected from patient symptoms.

This article presents a re-analysis of the TRIGGER trial to determine (a) the number of further bleeding episodes suspected from clinical signs that were not confirmed using more objective methods, and whether this number differed according to the treatment arm (indicating ascertainment bias); and (b) whether an analysis that included both confirmed and unconfirmed events led to bias in the estimated treatment effect, compared with the analysis that required more objective confirmation of every suspected event.

2.2 Statistical methods

We summarised the number of further bleeding episodes suspected from patient symptoms that were not confirmed using more objective methods separately for each treatment arm. We estimated the treatment effects (expressed as odds ratios), 95% confidence intervals, and p-values using two definitions for further bleeding episodes:

- ***Excluding unconfirmed events:*** This outcome included:
    - Confirmed events suspected from patient symptoms
    - Confirmed events identified during endoscopy
- ***Including unconfirmed events:*** This outcome included:

- Confirmed events suspected from patient symptoms
- Confirmed events identified during endoscopy
- Unconfirmed events suspected from patient symptoms

We analysed both definitions of further bleeding using a logistic regression model, adjusted for baseline covariates (age, the number of co-morbidities, presence of shock, and coagulopathy) [16]. We used generalised estimating equations to account for clustering by hospital [17]. The treatment effect estimates were expressed as the restrictive transfusion policy vs. liberal transfusion policy. An odds ratio less than 1 indicated a beneficial effect of the restrictive transfusion policy. An odds ratio greater than 1 indicated a beneficial effect of the liberal transfusion policy. We compared the odds ratios for the two outcome definitions by calculating a ratio of odds ratios (ROR), by dividing the odds ratio obtained from the analysis that included unconfirmed events by the odds ratio obtained from the analysis that excluded unconfirmed events.

The ROR gives a measure of the bias introduced by including episodes that were not confirmed using more objective methods. An ROR of 1 indicates no bias, an ROR<1 indicates bias in favour of the restrictive transfusion policy, and an ROR>1 indicates bias in favour of the liberal transfusion policy.

As discussed earlier, in rare cases, unconfirmed events suspected from patient symptoms can occur because a follow-up endoscopy was not performed after an episode of further bleeding was suspected based on patient symptoms. This situation can occur if further intervention is deemed to not be in the patient's best interests, such as on grounds of futility if the patient is likely to die in the near future due to co-morbidity. It is then unclear whether the initial assessment of further bleeding based on patient symptoms is correct or not. In TRIGGER, we did not record whether an endoscopy was explicitly performed to confirm a suspected bleeding event. We therefore cannot say with certainty that all unconfirmed events suspected from patient symptoms were because the endoscopy showed no signs of further bleeding. A small number of unconfirmed events may have occurred because no confirmatory endoscopy was undertaken due to the patient's health status, although this event is uncommon [18]. We therefore performed a sensitivity analysis [19], in which we reclassified unconfirmed events suspected from patient symptoms as confirmed events if the patient died within one week of the unconfirmed bleeding event.

A patient with severe bleeding may undergo interventional radiology or surgery to confirm and treat the lesion, rather than endoscopy. For simplicity, we have referred to further bleeding events

confirmed either by endoscopy, radiology, or surgery as events that were confirmed using more objective methods.

The re-analyses presented here are exploratory and should not be used to make inferences about the effectiveness of a restrictive or liberal approach to red blood cell transfusion in patients with AUGIB.

3. Results

3.1 TRIGGER data

Further bleeding data at 28 days were available for 905 patients, 393 of whom were on the restrictive policy and 512 on the liberal policy. As shown in Figure 1, there were 27 further bleeding events in the restrictive transfusion arm, of which 14 were suspected based on patient symptoms but were not confirmed using more objective methods, 3 were suspected based on patient symptoms and were confirmed using more objective methods and 10 were discovered directly during an endoscopy. In comparison, there were 42 further bleeding episodes in the liberal transfusion group, of which 11 were suspected from patient symptoms but were not confirmed using more objective methods, 14 were based on patient symptoms and were confirmed using more objective methods, and 17 were discovered directly during an endoscopy (Figure 1).

There were 17 occurrences of further bleeding that were initially suspected based on patient symptoms in the restrictive group and 25 in the liberal group. A significantly lower proportion of these events were confirmed using more objective methods in the restrictive arm than in the liberal arm (restrictive 3/17 (18%) vs. liberal 14/25 (56%), ratio 0.32, 95% CI 0.11 to 0.93).

3.2 Including unconfirmed events biased the treatment effect

The analysis that included unconfirmed events (i.e. included unconfirmed events suspected from patient symptoms in addition to events confirmed by an endoscopy) analysed 27/393 events (7%) in the restrictive group and 42/512 (8%) in the liberal group, giving an odds ratio of 0.83 (95% CI: 0.50 to 1.37, p = 0.47). The analysis that excluded unconfirmed events (i.e. included only events confirmed by an endoscopy) analysed 13/393 events (3%) in the restrictive group and 31/512 (6%) in the liberal group, giving an odds ratio of 0.50 (95% CI: 0.32 to 0.78, p = 0.002).

The definition which included unconfirmed events indicated a small difference between treatment groups that was not statistically significant. In contrast, the definition which excluded unconfirmed

events found a large reduction in further bleeding episodes, which was statistically significant. The ROR was 1.66, indicating that the analysis which included unconfirmed events as part of the primary outcome biased the treatment effect upwards by 66% in favour of the liberal transfusion arm.

3.3 Sensitivity analysis

Of the patients who had an unconfirmed further bleeding event that was initially suspected based on their symptoms, 2/11 (18%) in the liberal arm and 2/14 (14%) in the restrictive arm died within a week of the suspected bleed. These events were therefore more likely to have remained unconfirmed due to the patients' ill health, rather than being ruled out as further bleeding with an endoscopy. A sensitivity analysis that reclassified these bleeding episodes as confirmed found similar results to the original analysis; the odds ratio for confirmed bleeding episodes was 0.53 (95% CI 0.34 to 0.83, p = 0.005) and the ROR was 1.57.

4. Discussion

Blinded outcome assessment is sometimes difficult to achieve in open-label trials. A recent review found that only 26% of open-label trials reported using blinded outcome assessment [20]. Unblinded outcome assessment can lead to biased estimates of treatment effect when the assessment of the outcome measure is subjective. The outcome definition can be modified to remove subjective elements, as was done in the TRIGGER trial [11]. However, it is unclear how much of a difference this approach makes in practice. We re-analysed the TRIGGER trial to assess whether modifying the usual outcome definition for further bleeding to exclude subjective events based solely on patient symptoms had effectively reduced bias.

We found that a substantially higher proportion of subjective events were not confirmed using more objective methods in the restrictive arm than in the liberal arm, indicating differential assessment between the treatment groups. Clinicians' preconceived expectations of the efficacy of the two interventions may have driven this difference. Previous research has shown that bias associated with unblinded outcome assessment is often in the direction of the assessors' expectations. For example, if assessors expect an intervention to perform better than its control, they will typically assess the intervention patients as having done better than the control patients [21]. The clinicians involved in TRIGGER may have expected the liberal transfusion policy to lead to better patient outcomes than the restrictive policy, as the liberal policy more closely reflects usual clinical practice.

The differential assessment in the treatment arms led to substantially different results when unconfirmed events were included or excluded. Including unconfirmed events in the outcome definition biased the treatment effect upwards by 66%. The statistically significant difference found in favour of the restrictive transfusion arm thus became non-significant. The strategy of modifying outcome definitions to exclude subjective elements in open-label trials where blinded outcome assessment is not feasible is thus a useful option for researchers to consider in order to reduce bias.

It should be noted that modifying an outcome definition may have other implications for the trial apart from a reduction in bias. Excluding (or including) some elements in the outcome definition will have implications for precision and power. For example, in TRIGGER the number of events was reduced from 69 to 44 after redefining the endpoint. This can have implications for sample size calculations, which depend on the number of events (for time-to-event outcomes) or the proportion of patients experiencing an event (for binary outcomes). A reduction in the number or proportion of events will typically lead to larger sample size requirements. However, it is not necessarily the case that including additional elements in the definition will increase power; if these extra elements are not representative of the underlying 'true' outcome (e.g. are poor surrogates) then incorporating them in the outcome definition will actually reduce precision, and hence require larger sample sizes.

Another important consideration is the clinical relevance of the outcome definition. Excluding (or including) some elements in the outcome definition will have implications for how meaningful the outcome is to patients and clinicians. For example, in TRIGGER, further bleeding could have been defined based on a drop in the patient's haemoglobin level, which could have easily been assessed by a blinded committee. However, this is a poor surrogate outcome for further bleeding, and so although such a definition would have led to an unbiased estimate, it would not have accurately reflected the underlying clinical condition. Therefore, when modifying an outcome definition it is important to ensure that the clinical relevance of the outcome is retained.

One limitation of this work is that we have assumed that all episodes of further bleeding were identified, either based on clinical suspicion or directly during an endoscopy. However, it is possible that some events were missed; for example, a patient may have had a further bleeding episode, but the symptoms were sufficiently mild that the clinician did not suspect a further bleeding event. The estimated odds ratios from both analyses may have differed slightly had these events been included, and as such, the ratio of odds ratios may also have differed.

5. Conclusion

Modifying the outcome definition to exclude subjective elements led to a substantial reduction in bias in the TRIGGER trial. This may be a useful strategy to adopt for open-label trials where blinded outcome assessment is not feasible.

6. Acknowledgements

We wish to acknowledge Jennifer De Beyer, Centre for Statistics in Medicine, Oxford (UK) for her editorial assistance with this manuscript. We would also like to thank two reviewers and an associate editor for helpful comments on the manuscript.

References

1. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012;344:e1119.
2. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2013;185(4):E201-11.
3. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Rasmussen JV, Hilden J, et al. Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. International journal of epidemiology. 2014.
4. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology. 1994;44(1):16-20.
5. Poolman RW, Struijs PA, Krips R, Sierevelt IN, Marti RK, Farrokhyar F, et al. Reporting of outcomes in orthopaedic randomized trials: does blinding of outcome assessors matter? The Journal of bone and joint surgery American volume. 2007;89(3):550-8.
6. Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Annals of internal medicine. 2012;157(6):429-38.
7. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601-5.
8. Jairath V, Kahan BC, Gray A, Dore CJ, Mora A, Dyer C, et al. Restrictive vs liberal blood transfusion for acute upper gastrointestinal bleeding: rationale and protocol for a cluster randomized feasibility trial. Transfusion medicine reviews. 2013;27(3):146-53.
9. Jairath V, Kahan BC, Gray A, Dore CJ, Mora A, James MW, et al. Restrictive versus liberal blood transfusion for acute upper gastrointestinal bleeding (TRIGGER): a pragmatic, open-label, cluster randomised feasibility trial. Lancet. 2015.
10. Kahan BC, Jairath V, Murphy MF, Dore CJ. Update on the transfusion in gastrointestinal bleeding (TRIGGER) trial: statistical analysis plan for a cluster-randomised feasibility trial. Trials. 2013;14:206.
11. Kahan BC, Cro S, Dore CJ, Bratton DJ, Rehal S, Maskell NA, et al. Reducing bias in open-label trials where blinded outcome assessment is not feasible: strategies from two randomised trials. Trials. 2014;15(1):456.
12. Gotzsche PC, Liberati A, Torri V, Rossetti L. Beware of surrogate outcome measures. International journal of technology assessment in health care. 1996;12(2):238-46.
13. Grimes DA, Schulz KF. Surrogate end points in clinical research: hazardous to your health. Obstetrics and gynecology. 2005;105(5 Pt 1):1114-8.
14. Laine L, Spiegel B, Rostom A, Moayyedi P, Kuipers EJ, Bardou M, et al. Methodology for randomized trials of patients with nonvariceal upper gastrointestinal bleeding: recommendations from an international consensus conference. The American journal of gastroenterology. 2010;105(3):540-50.
15. Moustgaard H, Bello S, Miller FG, Hrobjartsson A. Subjective and objective outcomes in randomized clinical trials: definitions differed in methods publications and were often absent from trial reports. Journal of clinical epidemiology. 2014;67(12):1327-34.
16. Kahan BC, Jairath V, Dore CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials. 2014;15:139.
17. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22.

18. Hearnshaw SA, Logan RF, Lowe D, Travis SP, Murphy MF, Palmer KR. Acute upper gastrointestinal bleeding in the UK: patient characteristics, diagnoses and outcomes in the 2007 UK audit. Gut. 2011;60(10):1327-35.
19. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. BMC medical research methodology. 2014;14:11.
20. Kahan BC, Rehal S, Cro S. Blinded Outcome Assessment Was Infrequently Used and Poorly Reported in Open Trials. PloS one. 2015;10(6):e0131926.
21. Linde K, Witt CM, Streng A, Weidenhammer W, Wagenpfeil S, Brinkhaus B, et al. The impact of patient expectations on outcomes in four randomized controlled trials of acupuncture in patients with chronic pain. Pain. 2007;128(3):264-71.

Figure 1 – Treatment effect estimates in TRIGGER under different outcome definitions

*This figure shows the number of events for each treatment arm in TRIGGER under two different outcome definitions: (a) including all further bleeding events; and (b) including only further bleeding events confirmed by endoscopy or another more objective method.