Outcome assessment by central adjudicators in randomised stroke trials: simulation of differential and non-differential misclassification

Peter J Godolphin[1,2], Philip M Bath[3,4], Christopher Partlett[1], Eivind Berge[5], Martin M Brown[6], Misha Eliasziw[7], Per Morten Sandset[8], Joaquín Serena[9], Alan A Montgomery[1]

Corresponding Author: Peter J Godolphin

Corresponding Author's Email: p.godolphin@ucl.ac.uk

Corresponding Author's Phone Number: +44 (0)20 7670 4801

Corresponding Author's Address: MRC Clinical Trials Unit at University College London, Institute of Clinical Trials & Methodology, 90 High Holborn, London, WC1V 6LJ

Institutions:

1. Nottingham Clinical Trials Unit, University of Nottingham, Nottingham, UK

2: MRC Clinical Trials Unit at University College London, Institute of Clinical Trials and Methodology, London, UK

3: Stroke Trials Unit, Division of Clinical Neuroscience, University of Nottingham, Nottingham, UK

4: Stroke, Nottingham University Hospitals NHS Trust, Nottingham, UK

5: Department of Internal Medicine, Oslo University Hospital, Oslo, Norway

6: Stroke Research Group, UCL Institute of Neurology, UCL, London, UK

7: Department of Public Health and Community Medicine, Tufts University, Boston, USA

8: Department of Haematology, Oslo University Hospital and University of Oslo, Oslo, Norway

9: Department of Neurology. Stroke Unit, Hospital Josep Trueta, IDIBGI. Girona, Spain

Total number of tables and figures: 5 tables and 1 figure

Keywords: Adjudication, stroke, clinical trial, simulation, detection bias, misclassification

Word count: 4496 (main text) + 139 (declarations)

Abstract:

Introduction: Adjudication of the primary outcome in randomised trials is thought to control misclassification. We investigated the amount of misclassification needed before adjudication changed the primary trial results.

Methods: We included data from five randomised stroke trials. Differential misclassification was introduced for each primary outcome until the estimated treatment effect was altered. This was simulated 1000 times. We calculated the between-simulation mean proportion of participants that needed to be differentially misclassified to alter the treatment effect.

In addition, we simulated hypothetical trials with a binary outcome and varying sample size (1000-10000), overall event rate (10-50%), and treatment effect (0.67-0.90). We introduced non-differential misclassification until the treatment effect was non-significant at 5% level.

Results: For the five trials, the range of unweighted kappa values were reduced from 0.89-0.97 to 0.65-0.85 before the treatment effect was altered. This corresponded to 2.1%-6% of participants misclassified differentially for trials with a binary outcome. For the hypothetical trials, those with a larger sample size, stronger treatment effect and overall event rate closer to 50% needed a higher proportion of events non-differentially misclassified before the treatment effect became non-significant.

Discussion: We found that only a small amount of differential misclassification was required before adjudication altered the primary trial results, whereas a considerable proportion of participants needed to be misclassified non-differentially before adjudication changed trial conclusions. Given that differential misclassification should not occur in trials with sufficient blinding, these results suggest that central adjudication is of most use in studies with unblinded outcome assessment.

Conclusion: For trials without adequate blinding, central adjudication is vital to control for differential misclassification. However, for large blinded trials, adjudication is of less importance and may not be necessary.

1. Introduction:

In randomised trials, outcomes are commonly assessed by site investigators at each trial site. For studies with many sites, random error (non-differential misclassification) could be introduced into outcome assessment through inexperience of site investigators or varied practice across sites. Furthermore, for open-label trials with inadequate blinding of treatment allocation, there is the possibility of detection bias in the assessment of outcomes, with site investigators misclassifying outcomes differently between arms (differential misclassification). Hróbjartsson et al.[1] showed that, on average, unblinded site investigators exaggerate treatment effects for subjective binary outcomes by 36%.

To control for differential and non-differential misclassification, many trials use a central adjudication committee, made up of blinded independent experts who assess the trial outcome in addition to the site investigators. The central adjudicators' assessment of the outcome is often used in preference to that of the site investigators. Central adjudication is commonly included in vascular trials[2], including those that are investigating stroke[3, 4], although the value of adjudication has been questioned[5-7].

We have previously carried out a systematic review and meta-analysis that included 15 randomised stroke trials where both central adjudicators and site investigators assessed the primary outcome[8]. In this systematic review, we found no evidence that central adjudication of a trial's primary outcome altered the treatment effect estimate compared with the estimate obtained using site reported data (pooled ratio of treatment effects (RTE)=1.02, 95% C.I:[0.95, 1.09]). This result concurred with two other meta-analyses investigating the impact of adjudication of binary outcomes on the treatment effect estimates[6, 7].

The aim of the present simulation study was to investigate whether there are circumstances when central adjudication of a trial's primary outcome would change the primary treatment effect estimate.

2. Methods:

To investigate when central adjudication changes a trial's results we can explore how much differential misclassification by site investigators was necessary to alter the estimated treatment effect, i.e. the 95% confidence interval of the RTE excludes the null value, one. However, this investigation ignored the statistical significance of the treatment estimate pre- and post-misclassification and only identified when the treatment effect estimate differs significantly to the estimate obtained after central adjudication (RTE≠1). Therefore, for completeness, we considered situations where the RTE remained at one after misclassification (here non-differential), and the significance of the treatment estimate differed pre- and post-misclassification. Thus, in this study we (1) evaluated how much differential misclassification was needed to alter the estimated treatment effect; and (2) explored how much non-differential misclassification caused a significant treatment effect to become non-significant at 5% level.

2.1 Differential misclassification using real trial data:

For studies with adequate blinding, central adjudication should control for non-differential misclassification by reducing random 'noise' around the main estimate of interest. However, increasing this 'noise' in a simulation will not meaningfully shift the estimate of interest, because the amount of misclassification in a blinded trial should be equal in both treatment arms. Therefore, to explore the situation where central adjudication does alter the treatment effect estimate, we introduced differential misclassification. Previous studies have shown that site investigators often exaggerate treatment effect estimates[1], so we introduced differential misclassification for outcomes assessed by site investigators to make the treatment effect estimates more beneficial. The starting point for misclassification was the centrally adjudicated data, as this is the gold standard, and outcomes were misclassified to increasing extent. This misclassification differs for binary and ordinal variables, as explained below.

2.1.1 Data collection

Our systematic review of central adjudication in stroke trials included 15 trials totalling 69,650 participants[8]. All included trials had their primary outcome assessed by both site investigators and central adjudicators, and were asked to provide either summary results or individual patient data

(IPD). Of the 15 trials in our systematic review, we selected the five trials that provided IPD, as differential misclassification is introduced at a patient level. The five studies covered a variety of outcomes, number of participants randomised, and treatment effectiveness[9-14].

The five studies selected corresponded to seven unique populations as one study, NASCET, carried out separate analyses for patients with mild-, moderate- and severe-grade carotid artery stenosis (denoted as NASCET:mild, NASCET:moderate and NASCET:severe respectively). Throughout the remainder of this article these will be referred to as individual trials. Therefore, in this simulation study there were seven trials included (HAEST, ICSS, REVASCAT, TARDIS, and the three aforementioned NASCET subpopulations).

2.1.2 Misclassification for binary outcomes

For binary outcomes, differential misclassification was introduced by increasing the proportion of participants who (a) were in the control arm and had an event, and, (b) were in the treatment arm and did not have an event. For each trial, varying proportions of participants were randomly misclassified. Only participants in the control arm without the event and participants in the treatment arm with the event were misclassified, as the objective was to make the treatment effect estimates more beneficial.

2.1.3 Misclassification for ordinal outcomes

For ordinal outcomes, a similar approach was taken. In both trials where the outcome was analysed in an ordinal fashion, participants could be allocated one of six categories. To simulate increased differential misclassification, selected participants in the control arm had their outcome value increased (worse outcome) and those in the treatment group had their outcome value decreased (better outcome). As the proportion of participants misclassified in the simulation increased, the number of participants misclassified by one category, two categories and so on, increased proportionally. Outcomes were constrained by the minimum (0) and maximum (5) values.

2.1.4 The proportion of misclassification necessary to alter the estimated treatment effect

The number of participants misclassified was increased in 0.1% increments, and, for each increment the trial's primary analysis was repeated using the misclassified outcome. The treatment effect was then compared with the treatment effect based on central adjudicated data (remains constant for each trial) using the ratio of treatment effects (RTE). An RTE < 1 indicates that the misclassified data

produces a more beneficial treatment effect. For each 0.1% increment, we ran 1000 simulations, from which we then calculated the mean RTE and 95% confidence interval. We stopped increasing the increments when the upper bound of the 95% confidence interval was less than 1 (misclassified treatment effect is significantly different to the treatment effect based on centrally adjudicated data).

2.1.5 Statistical analysis

We calculated percent agreement and unweighted kappa between central adjudicators and site investigators for the primary outcome of each trial before misclassification. For trials with ordinal outcomes, weighted kappa used linear weights was also determined. Each trial was analysed as per the analysis specified in their main results paper, except for the three NASCET trials, where a univariate Cox proportional hazards model was fitted for each trial.

After simulation, the within-simulation mean and standard deviation of the treatment effect after misclassification, number of participants misclassified, crude percent agreement and unweighted (and weighted if appropriate) kappa were determined for each trial. All analyses, including those described in the following sections, were undertaken using Stata version 15.1.

2.2 Non-differential misclassification using hypothetical trial data:

For studies with adequate blinding, any misclassification of an outcome is expected to be equal between treatment and control arms, that is, non-differential. For these studies, the RTE will be close to one even with introduction of a large amount of non-differential misclassification. However, this could still impact on trial conclusions by introducing greater random error, resulting in wider 95% confidence intervals around the estimated treatment effect. Thus, we can estimate the amount of non-differential misclassification required to cause a loss of precision such that the 95% confidence interval for a real treatment effect no longer excludes the null.

2.2.1 Data generation

Data was generated using Stata to represent a simple parallel group trial with a binary primary outcome. We estimated the treatment effect using relative risk and significance level was set at 5%. We aimed to establish how much non-differential misclassification was required for a previously significant treatment effect to become non-significant.

2.2.3 Characteristics to vary

Three different treatment effects were chosen: relative risks of 0.67 (for example, events in a ratio of 3:2 between control and treatment groups respectively), 0.82 (ratio of 11:9) and 0.90 (ratio of 21:19) to represent strong, moderate and modest treatment effects respectively. In stroke trials overall event rate is usually low, so we explored situations where the overall event rate was ≤50%.The overall event rate was simulated in 10% intervals, from 10% to 50% and additionally at 15%. Finally, the overall trial sample size was simulated to be either 1000, 2000, 3000, 5000 or 10000. Thus, by varying sample size, overall event rate and treatment effect there were 90 distinct scenarios. This is summarised in Table 1. The simulation code is provided in the supplementary material to enable further, more specific, scenarios to be explored.

2.2.4 Misclassifying events

For each scenario, events were misclassified proportionately in each arm in order to preserve the relative risk and thus keep the RTE equal to one. The amount of misclassification required for the 95% confidence interval of the relative risk to include the null value of one was expressed as a percentage of the total number of events in the original dataset.

3. Findings:

3.1 Differential misclassification using real trial data:

For five of the trials, the primary outcome was binary, whereas for the remaining two trials the primary outcome was analysed on an ordinal scale (Table 2). The number of participants randomised varied between 206 (REVASCAT) and 3096 (TARDIS). Using the real data, agreement was high between central adjudicators and site investigators, with crude agreement ranging from 93.2% to 99.6% and kappa ranging from 0.89 to 0.97 (Table 3, see Supplementary Tables 1a-1b and 2a-2b).

After simulation of differential misclassification, as planned, the treatment effect was more beneficial for every trial such that the upper bound of the confidence interval for the RTE was 0.99 (Table 4). For trials with a binary outcome, between 2.1% and 6% of participants needed to be differentially misclassified to alter the estimated treatment effect, with the amount of misclassification inversely associated with study size (Table 4). In the two trials with ordinal primary outcomes, there was substantial variation in the proportion of participants that needed to be misclassified (1.9% and 27.8%). However, these studies did represent the trials with the largest and smallest number of participants respectively. Following misclassification, crude agreement remained high for all but one of the trials, but the kappa values were reduced in the range of 0.65 to 0.85 (Table 5, see Supplementary Tables 1c-1d and 2c-2d).

3.2 Non-differential misclassification using hypothetical trial data:

For 26 of the scenarios, the initial risk ratio was not significant at 5% before misclassification, so these cases are not given (displayed as NA in Supplementary Table 3). As expected, more events were required to be misclassified to change a significant treatment effect to non-significant at the 5% level when the original treatment effect was strongest (Figure 1, see Supplementary Table 3 and Supplementary Figures 1-2).

Greater sample size and higher overall event rate both required a larger proportion of events to be misclassified before significant treatment effects become non-significant (Figure 1). For example, in a hypothetical blinded trial with 5000 participants, overall event rate of 20% and a modest treatment effect (relative risk=0.82), 649 (64.9%) of the events would need to be misclassified non-differentially before a significant treatment would become non-significant.

4. Discussion:

In this simulation study based on seven distinct stroke trial populations we found that only a small amount of differential misclassification was needed before central adjudication would have altered the estimated treatment effect. Larger trials appeared to be most vulnerable to this bias, in part due to their larger sample size being able to detect a smaller difference in treatment effect. However, for blinded studies where differential misclassification should not occur, an implausible amount of random error is required to alter trial conclusions.

Whilst ordinal outcomes could be misclassified by more than one level (i.e. mRS of 1 to 3), it can be argued that this would be less severe than misclassification of a binary event (0 to 1 or vice versa). Therefore, the results from binary and ordinal outcomes should not be compared. Overall, we found that a relatively small amount of differential misclassification was needed to alter the estimated treatment effect. This suggests that central adjudication is important to control for differential misclassification in randomised trials. However, three of the five trials included had blinded outcome assessment, so the plausibility of this amount of differential misclassification happening in practice to these studies is far less than the unblinded trials. In our review[8] we did not see any indication of detection bias through differential misclassification, so even the small proportion needed before the treatment effect changes may be a rare occurrence in trials. One reason for this finding in our review could be due to 9 (60%) of the included studies having the site investigators blind to treatment allocation and the majority of the studies had stroke as their primary outcome, which is well defined and accurately measured[15]. We found no significant interaction between blinding status and RTE, but this may have been due to the reviews small sample size. A Cochrane review[7] that included 47 trials which adjudicated subjective binary events did find an interaction between blinding status of the site investigators and the ratio of odds ratios (RORs), with the suggestion that unblinded site investigators exaggerate treatment effect estimates (two trials, ROR=0.76, 95% C.I: [0.46, 1.12]). Furthermore, unblinded site assessors have been shown to exaggerate treatment effect estimates in multiple studies by Hróbjartsson[1, 16, 17]. Thus, differential misclassification is a real possibility in medical research, and adjudication can control for this.

However, for blinded studies, we would not expect central adjudication to control for differential misclassification, and instead only reduce random noise around the effect of interest. As expected,

the proportion of events needed to be misclassified before a significant treatment effect becomes non-significant increases with trial size, overall event rate and strength of treatment effect. This can also vary with method of adjudication, but this is not something we explored in our study. We have shown for a trial with a binary outcome that a large amount of non-differential misclassification is necessary before even a modest treatment effect is missed. For the five stroke trials included in the first part of this study, the largest agreement for a trial with a binary outcome was 98.8%. Far higher disagreement would have been needed before central adjudication ensures that a modest and significant treatment effect does not become non-significant through random error. In a previous simulation study that explored central adjudication of stroke type in a stroke trial with blinded outcome assessment[18], the agreement between the adjudicators and site investigators was 98% and kappa had to reduce from 0.92 to 0.46 before a true subgroup effect by stroke type was missed. This amount of random error is not plausible for many trial settings. Other studies investigating adjudication in stroke trials found agreement between adjudicators and site investigators of 91% for all stroke[19], and 90% for stroke[4]. Thus, for large blinded trials, central adjudication could be an unnecessary expenditure to control for non-differential misclassification. However, it is important to note that for other non-stroke outcomes commonly assessed in stroke trials, such as coronary events or fatal vascular events, agreement may not be as high as described above. Adjudication of these outcomes, especially if they are part of a primary composite outcome such as major adverse cardiovascular events, could still be warranted in these settings. One alternative approach to site-assessment followed by adjudication could be to assess outcomes centrally, taking away the need for site-assessment. However, this approach would only be suitable for those studies with central follow-up.

One limitation of our study is that we have only focused on adjudication of the primary outcome, and the high level of agreement we found across the included studies may be lower for different outcomes. For example, a study exploring adjudication of serious adverse events found agreement between site investigators and central adjudicators for likely causality of event of 56%[20]. However, we have chosen a variety of stroke trials that represent acute stroke, primary and secondary prevention studies as well as including the majority of common primary outcomes in these studies. Another limitation is that we only explored non-differential misclassification through binary outcomes. Our justification for this is that the majority of stroke trials included in our review[8] had a binary primary

outcome. Furthermore, it is possible that trials with ordinal outcomes would need greater misclassification than those with binary outcomes, due to the ordinal scale the outcome is measured on.

To conclude, we found that central adjudication is important for stroke trials without sufficient blinding for outcome assessment through its control of differential misclassification. However, for randomised stroke trials that do have adequate blinded outcome assessment, central adjudication is less important and may not be necessary.

References:

1. Hróbjartsson, A., 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**: p. e1119.
2. Dechartres, A., et al., *Inadequate planning and reporting of adjudication committees in clinical trials: recommendation proposal.* J Clin Epidemiol, 2009. **62**(7): p. 695-702.
3. Lopez-Cancio, E., et al., *Phone and Video-Based Modalities of Central Blinded Adjudication of Modified Rankin Scores in an Endovascular Stroke Trial.* Stroke, 2015. **46**(12): p. 3405-10.
4. Ninomiya, T., et al., *Effects of the end point adjudication process on the results of the Perindopril Protection Against Recurrent Stroke Study (PROGRESS).* Stroke, 2009. **40**(6): p. 2111-5.
5. Granger, C.B., et al., *Do we need to adjudicate major clinical events?* Clin Trials, 2008. **5**(1): p. 56-60.
6. Pogue, J., S.D. Walter, and S. Yusuf, *Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs.* Clin Trials, 2009. **6**(3): p. 239-51.
7. Ndounga Diakou, L.A., et al., *Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates.* Cochrane Database Syst Rev, 2016. **3**: p. Mr000043.
8. Godolphin, P.J., et al., *Outcome assessment by central adjudicators versus site investigators in randomised stroke trials: A systematic review and meta-analysis.* Stroke, 2019.
9. Bath, P.M., et al., *Antiplatelet therapy with aspirin, clopidogrel, and dipyridamole versus clopidogrel alone or aspirin and dipyridamole in patients with acute cerebral ischaemia (TARDIS): a randomised, open-label, phase 3 superiority trial.* The Lancet, 2018. **391**(10123): p. 850-859.
10. Barnett, H.J.M., et al., *Benefit of Carotid Endarterectomy in Patients with Symptomatic Moderate or Severe Stenosis.* New England Journal of Medicine, 1998. **339**(20): p. 1415-1425.
11. Barnett, H.J.M., et al., *Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis.* N Engl J Med, 1991. **325**(7): p. 445-53.
12. Berge, E., et al., *Low molecular-weight heparin versus aspirin in patients with acute ischaemic stroke and atrial fibrillation: a double-blind randomised study. HAEST Study Group. Heparin in Acute Embolic Stroke Trial.* Lancet, 2000. **355**(9211): p. 1205-10.
13. Bonati, L.H., et al., *Long-term outcomes after stenting versus endarterectomy for treatment of symptomatic carotid stenosis: the International Carotid Stenting Study (ICSS) randomised trial.* The Lancet, 2015. **385**(9967): p. 529-538.
14. Jovin, T.G., et al., *Thrombectomy within 8 Hours after Symptom Onset in Ischemic Stroke.* New England Journal of Medicine, 2015. **372**(24): p. 2296-2306.
15. Hicks Karen, A., et al., *2017 Cardiovascular and Stroke Endpoint Definitions for Clinical Trials.* Circulation, 2018. **137**(9): p. 961-972.
16. Hrobjartsson, A., 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.* Int J Epidemiol, 2014. **43**(3): p. 937-48.
17. Hrobjartsson, A., et al., *Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors.* Cmaj, 2013. **185**(4): p. E201-11.
18. Godolphin, P.J., et al., *Central masked adjudication of stroke diagnosis at trial entry offered no advantage over diagnosis by local clinicians: Secondary analysis and simulation.* Contemporary clinical trials communications, 2018. **12**: p. 176-181.
19. Easton, J.D., et al., *Estimated treatment effect of ticagrelor versus aspirin by investigator-assessed events compared with judgement by an independent event adjudication committee in the SOCRATES trial.* Int J Stroke, 2019: p. 1747493019851282.
20. Godolphin, P.J., et al., *Central adjudication of serious adverse events did not affect trial's safety results: Data from the Efficacy of Nitric Oxide in Stroke (ENOS) trial.* PLOS ONE, 2018. **13**(11): p. e0208142.

Tables and Figures:

Table 1: Summary of parameters used in the simulation of non-differential misclassification

Description	Values
Treatment effect	0.67, 0.82, 0.90
Overall event rate	10%, 15%, 20%, 30%, 40%, 50%
Sample size	1000, 2000, 3000, 5000, 10000

Table 2: Summary of included trials

Trial name	Population	Intervention	Comparator	Primary outcome	Were site investigators blind to treatment?	Adjudication information
HAEST	Patients with acute ischaemic stroke and atrial fibrillation (n=449).	Dalteparin (n=224)	Aspirin (n=225)	Recurrent ischaemic Stroke (binary). Analysed using logistic regression.	Yes	Two clinicians assessed medical notes (including reports from cranial scans) and original case report forms with diagnosis concealed.
ICSS	Patients with symptomatic carotid stenosis (n=1713)	Stenting (n=855)	Carotid endarterectomy (n=858)	Fatal or disabling stroke (binary). Analysed using cox regression.	No	Two clinicians assessed outcome without knowledge of site assessment
NASCET	Patients with non-disabling stroke and carotid stenosis of 30-99% in the internal carotid artery. There were three populations: mild (<50%, n=1368); moderate (50-69%, n=858); and severe (70-99%, n=659) stenosis.	Carotid endarterectomy. In addition, patients received medical care, including antiplatelet therapy. Mild (n=678), moderate (n=430), severe (n=328).	Medical care, including antiplatelet therapy. Mild (n=690), moderate (n=428), severe (n=331)	Fatal or non-fatal ipsilateral stroke (binary). Analysed using Mantel–Haenszel chi-square test. To obtain an estimate, we analysed NASCET trials using univariate cox regressionl.	No	Neurologists and surgeons assessed original case report forms and cranial scans without knowledge of site assessment.
REVASCAT	Patients with acute ischaemic stroke who could be treated within 8 hours (n=206)	Medical therapy (including alteplase if eligible) and thrombectomy (n=103)	Medical therapy (including alteplase if eligible) (n=103)	Functional outcome at 90 days (mRS, ordinal). Patients who scored 5 or 6 were grouped in a single category. Analysed using ordinal logistic regression (6 point scale)	Yes	Neurologists assessed audio-tape or video recording of patient evaluation of the primary outcome.
TARDIS	Patients with acute ischaemic stroke or TIA (n=3096).	Aspirin, clopidogrel and dipyridamole (n=1556)	Aspirin and dipyridamole, or clopidogrel alone (n=1540)	Functional outcome and recurrent stroke and TIA (ordinal). Analysed using ordinal logistic regression (6 point scale)	Yes	Clinicians assessed medical notes, original case report forms and cranial scans, if requested.

mRS refers to modified Rankin Scale

Table 3: Agreement between central adjudicators and site investigators on the primary outcome using original trial data

Trial	Central adjudicator data			Site investigator data			Agreement between central adjudicators and site investigators			Crude agreement	Kappa
HAEST		Treated	Control		Treated	Control	SI \ CA	No event	Event	99.6%	0.97*
	No event	205	208	No event	203	208	No event	411	0		
	Event	19	17	Event	21	17	event	2	36		
ICSS		Treated	Control		Treated	Control	SI \ CA	No event	Event	98.8%	0.89*
	No event	808	801	No event	812	802	No event	1600	14		
	Event	49	52	Event	44	50	event	7	87		
NASCET: mild		Treated	Control		Treated	Control	SI \ CA	No event	Event	98.9%	0.96*
	No event	589	580	No event	592	584	No event	1165	11		
	Event	89	110	Event	86	106	event	4	188		
NASCET: moderate		Treated	Control		Treated	Control	SI \ CA	No event	Event	99.2%	0.97*
	No event	373	348	No event	374	354	No event	721	7		
	Event	57	80	Event	56	74	event	0	130		
NASCET: severe		Treated	Control		Treated	Control	SI \ CA	No event	Event	99.2%	0.97*
	No event	300	264	No event	299	268	No event	563	4		
	Event	28	67	Event	29	63	event	1	91		
REVASCAT	see Supplementary Table 1a						see Supplementary Table 1b			93.2%	0.91* 0.96†
TARDIS	see Supplementary Table 2a						see Supplementary Table 2b			98.8%	0.91* 0.91†

SI refers to Site investigators; CA refers to Central adjudicators

*Unweighted kappa

†Weighted kappa using linear weights

Table 4: Number and proportion of participants required to be differentially misclassified to alter estimated treatment effect

Trial (N)	Treatment effect before misclassification (95% CI)	Mean treatment effect after misclassification (SD)	Mean number of participants misclassified (SD)	Mean percentage of participants misclassified (SD)	RTE (95% CI)
HAEST (n=449)	1.13 (0.57, 2.24)	0.45 (0.07)	20.4 (4.3)	4.5% (1.0%)	0.40 (0.16, 0.99)
ICSS (n=1710)	0.94 (0.64, 1.39)	0.54 (0.04)	35.9 (5.9)	2.1% (0.3%)	0.59 (0.34, 0.99)
NASCET: mild (n=1368)	0.83 (0.72, 0.95)	0.55 (0.03)	55.0 (7.0)	4.0% (0.5%)	0.68 (0.46, 0.99)
NASCET: moderate (n=858)	0.71 (0.60, 0.84)	0.43 (0.03)	51.4 (6.7)	6.0% (0.8%)	0.63 (0.39, 0.99)
NASCET: severe (n=659)	0.36 (0.28, 0.43)	0.19 (0.02)	39.3 (5.9)	6.0% (0.9%)	0.53 (0.28, 0.99)
REVASCAT (n=206)	0.57 (0.35, 0.95)	0.28 (0.02)	57.2 (5.8)	27.8% (2.8%)	0.48 (0.23, 0.99)
TARDIS (n=3096)	0.90 (0.67, 1.20)	0.59 (0.03)	60.3 (7.3)	1.9% (0.2%)	0.66 (0.44, 0.99)

Data from 1000 simulations (starting seed 2206). Treatment effects lower than one indicates treatment is beneficial. SD refers to standard deviation. RTE refers to ratio of treatment effects.

Table 5: Agreement between central adjudicators and site investigators on primary outcome after differential misclassification

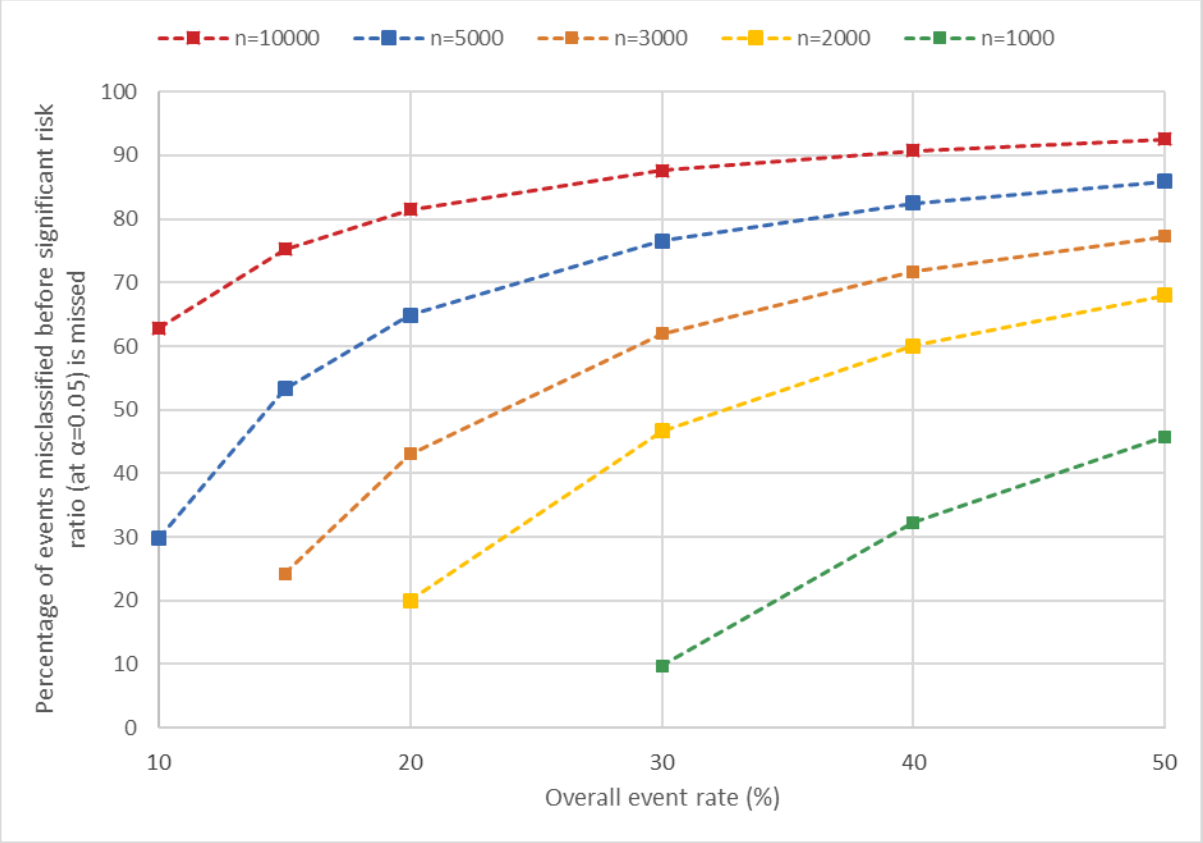
Trial	Central adjudicator data			Example misclassified site investigator data			Example agreement between central adjudicators and misclassified site investigators			Mean crude agreement (SD)	Mean kappa (SD)
HAEST		Treated	Control		Treated	Control	SI \ CA	No event	Event	95.4% (1.00%)	0.75 (0.05)*
	No event	205	208	No event	206	189	No event	394	1		
	Event	19	17	Event	18	36	event	19	35		
ICSS		Treated	Control		Treated	Control	SI \ CA	No event	Event	97.9% (0.34%)	0.84 (0.02)*
	No event	808	801	No event	811	769	No event	1577	3		
	Event	49	52	Event	46	84	event	32	98		
NASCET: mild		Treated	Control		Treated	Control	SI \ CA	No event	Event	96.0% (0.51%)	0.85 (0.02)*
	No event	589	580	No event	594	534	No event	1123	5		
	Event	89	110	Event	84	156	event	46	194		
NASCET: moderate		Treated	Control		Treated	Control	SI \ CA	No event	Event	94.0% (0.78%)	0.80 (0.02)*
	No event	373	348	No event	383	320	No event	683	10		
	Event	57	80	Event	47	128	event	38	127		
NASCET: severe		Treated	Control		Treated	Control	SI \ CA	No event	Event	94.0% (0.88%)	0.79 (0.03)*
	No event	300	264	No event	303	231	No event	531	3		
	Event	28	67	Event	25	100	event	33	92		
REVASCAT	see Supplementary Table 1c						see Supplementary Table 1d			72.2% (2.80%)	0.65 (0.03)* 0.84 (0.02)†
TARDIS	see Supplementary Table 2c						see Supplementary Table 2d			98.1% (0.24%)	0.85 (0.02)* 0.87 (0.02)†

Crude agreement and kappa are from 1000 simulations (starting seed 2206). Example site investigator data and example agreement are taken from one of the 1000 simulations. SI refers to Site investigators; CA refers to Central adjudicators

*Unweighted kappa

†Weighted kappa using linear weights

Figure 1: Amount of non-differential misclassification required such that treatment effect (relative risk=0.82) is no longer significant at 5% level for various sample sizes and overall event rates



Missing scenarios are due to the initial treatment effect before misclassification being non-significant (p>0.05). n refers to hypothetical trial sample size