OPEN ACCESS Freely available online

PLOS ONE

Quantification of Errors in Ordinal Outcome Scales Using Shannon Entropy: Effect on Sample Size Calculations

Pitchaiah Mandava[1,3]*, Chase S. Krumpelman[3], Jharna N. Shah[3], Donna L. White[2], Thomas A. Kent[1,3]

1 The Michael E. DeBakey Veterans Affairs Medical Center Comprehensive Stroke Program, Houston, Texas, United States of America, **2** Clinical Epidemiology and Comparative Effectiveness Program, Houston Veterans Affairs Health Services Research Center of Excellence, Houston, Texas, United States of America, **3** The Stroke Outcomes Laboratory, Department of Neurology, Baylor College of Medicine, Houston, Texas, United States of America

Abstract

Objective: Clinical trial outcomes often involve an ordinal scale of subjective functional assessments but the optimal way to quantify results is not clear. In stroke, the most commonly used scale, the modified Rankin Score (mRS), a range of scores (''Shift'') is proposed as superior to dichotomization because of greater information transfer. The influence of known uncertainties in mRS assessment has not been quantified. We hypothesized that errors caused by uncertainties could be quantified by applying information theory. Using Shannon's model, we quantified errors of the ''Shift'' compared to dichotomized outcomes using published distributions of mRS uncertainties and applied this model to clinical trials.

Methods: We identified 35 randomized stroke trials that met inclusion criteria. Each trial's mRS distribution was multiplied with the noise distribution from published mRS inter-rater variability to generate an error percentage for ''shift'' and dichotomized cut-points. For the SAINT I neuroprotectant trial, considered positive by ''shift'' mRS while the larger follow-up SAINT II trial was negative, we recalculated sample size required if classification uncertainty was taken into account.

Results: Considering the full mRS range, error rate was 26.1%±5.31 (Mean±SD). Error rates were lower for all dichotomizations tested using cut-points (e.g. mRS 1; 6.8%±2.89; overall $p<0.001$). Taking errors into account, SAINT I would have required 24% more subjects than were randomized.

Conclusion: We show when uncertainty in assessments is considered, the lowest error rates are with dichotomization. While using the full range of mRS is conceptually appealing, a gain of information is counter-balanced by a decrease in reliability. The resultant errors need to be considered since sample size may otherwise be underestimated. In principle, we have outlined an approach to error estimation for any condition in which there are uncertainties in outcome assessment. We provide the user with programs to calculate and incorporate errors into sample size estimation.

Citation: Mandava P, Krumpelman CS, Shah JN, White DL, Kent TA (2013) Quantification of Errors in Ordinal Outcome Scales Using Shannon Entropy: Effect on Sample Size Calculations. PLoS ONE 8(7): e67754. doi:10.1371/journal.pone.0067754

Editor: Robert K. Hills, Cardiff University, United Kingdom

Received January 28, 2013; **Accepted** May 22, 2013; **Published** July 5, 2013

Funding: This work was funded in part by the Baylor College of Medicine Institutional Clinical and Translational Research Grant Program (TAK) and Department of Veterans Affairs (VISN 16 PRG; PM). The Stroke Outcomes Laboratory was established through a grant from the pilot grant program of the Institute for Clinical and Translational Research at the Baylor College of Medicine to TAK. Dr. White's effort was supported in part by National Institutes of Health (NIH)/National Institute of Diabetes and Digestive and Kidney Diseases (K01 DK078154-05) and the Houston VA HSR&D Center of Excellence (HFP90-020). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist. Regarding the issue of a patent: The authors applied for a patent in 2008 for an idea that has no relevance to the manuscript under consideration. The patent application was denied.

* E-mail: pmandava@bcm.tmc.edu

Introduction

In the analysis of new therapeutic approaches to disease, it is essential that the effects of treatment be captured in a reliable manner. Measures for many conditions include scales that involve subjective assessment of a subject's well-being comparing two different treatments. In the case of stroke, the modified Rankin Score (mRS) is the most widely adopted measure of recovery of function in stroke trials [1]. As an ordinal scale, this instrument provides an ordering of possible outcomes, ranging from near complete recovery (e.g., 0 in mRS) to death (e.g., 6 in mRS). Analysis of outcome results can be performed by two methods: 1) Full-scale analysis where results for each group (treatment and placebo) are depicted as a proportion of patients in some or all ascending grade, and, 2) ''Dichotomization'' where results for each group are depicted as proportion of patients into two collapsed or binned grade categories (e.g. mRS 0–1 indicating excellent recovery, mRS 2–5, a dependant state), with an added ''safety'' category of mortality (mRS 6).

Dichotomization of outcome scales including dichotomization of mRS at cut-point of 1 (e.g, mRS 0–1 vs. 2–6) was used successfully in the NINDS trial of intravenous alteplase for ischemic stroke [2]. Of note alteplase is the first and only medication approved by FDA for use in ischemic stroke. More recently dichotomization at higher cut-points of mRS 3 and 4 have been employed in three randomized stroke trials of hemicraniectomy (DECIMAL, DESTINY, and HAMLET), which had patients with high baseline stroke severity, all of which were positive with relatively low number of subjects [3–5].

There remains discussion as to which method of analysis is the most appropriate approach for outcome measures in stroke trials. For example, the European Medical Agency issued guidance that when ordinal scales are used for testing the efficacy of novel medicines or devices, the full-scale be analyzed [6]. The impetus for this guidance came from the work of Whitehead [7], and Campbell et al [8], which showed that when number of categories is increased from two to six, sample size requirements are reduced by 23% because of a gain in the amount of information available [9]. Along these lines, several authors have suggested abandoning dichotomization in favor of ordinal scale analysis [10–13]. Proponents of full-scale analysis (also known as "shift"-analysis or "sliding dichotomization") support its use by invoking Shannon's seminal work on information systems and Altman's and Royston's work on the advantage of ordinal scale analysis vis-à-vis dichotomization [14–18]. Their central argument is that the loss of information inherent in switching from full-scale analysis to dichotomization may obscure important treatment effects [13,19]. The 'Shift' approach as suggested by Saver and Gornbein [12] and used in SAINT I [20], SAINT II [21], and IST-3 [22] was conceived as the difference in distributions between treatment and control groups as an ordinal/categorical analysis of outcome classification across all ranks, grades or a major part of the ordinal scale. This ordinal scale analysis is similar to that suggested by Whitehead [7] and Campbell et al [8]. It assumes a common proportional odds ratio applied to mRS 0, mRS 0–1, mRS 0–2, mRS 0–3, etc. Note that this "shift' differs from a change in modified Rankin score from baseline for each patient, as suggested by Lai and Duncan [23].

On both sides of this discussion (i.e., use of dichotomization vs. 'shift' analysis), there has not been explicit consideration of uncertainties regarding how well the recorded mRS scores reflect each patient's true recovery state. However, from the work of van Sweiten et al and others we know that inter-rater reliability of mRS is relatively low [24–27], particularly for mid-range (mRS scores of 2–4) values. Quinn et al have also shown that uncertainties in mRS assessment persist in spite of certification and re-education of assessors and do not depend on the assessors' field of specialization, educational background, country of origin, native language or length of patient interview [27–29]. These findings indicate that uncertainty or "noise" in the Rankin scoring may not be negligible, and indicate a need for closer examination of the patient-observer-score model that is the foundation of stroke outcome measurement.

In information processing terminology, dichotomization with an efficacy measure (mRS 0–1) and a safety measure (mRS 6) can be considered as an implementation of a band-stop filter. A central concept in information theory is the communication system which consists of a transmitter, a channel, and a receiver [15]. The transmitter produces a signal/symbol which is then passed on through the channel to the receiver for interpretation. In real-world situations, the channel is susceptible to noise which may corrupt the transmitted signal/symbol such that the receiver sees a different signal than was originally sent. This model is applicable to the situation of an observer evaluating a stroke patient, where the patient (transmitter) has a true Rankin score (signal) which is transmitted through the noisy channel of human assessment (observer) and is ultimately recorded as the outcome score for that patient (receiver).

In this paper, we hypothesized that uncertainties in assessment of this subjective outcome scale could be modeled and that errors will be higher if the entire scale is used compared to dichotomous measures. We calculate the error introduced by the channel (i.e., observer) during the transmission of the ordinal scale and dichotomized outcomes to an observer. Van Swieten's inter-rater variability matrix in mRS classification by different observers is used as a characterization of the noise introduced by the observation channel [24]. The inter-rater variability matrix has been termed the 'confusion matrix' in various sub-fields of information theory [30]. Using the confusion matrix, the error rate for each approach was calculated. We then demonstrate the effect of the noise/error on sample size calculations using the SAINT I trial as our working example [20]. SAINT I is a particularly interesting test case because this earlier phase trial reported positive results with the "Shift" approach as the primary outcome measure, with unspecified positive dichotomizations. The SAINT trials tested a putative neuroprotectant, NXY-059, in acute ischemic stroke with hopes that it would improve outcome or reduce the hemorrhage rate after thrombolysis. While SAINT I was considered positive using a 'shift' analysis to compare the range of ordinal mRS scale 0 to 4 and collapsing scale 5 and 6 in treated patients vs. the placebo control group, the subsequent larger SAINT II trial did not demonstrate benefit with respect either to the "shift" or the commonly used mRS 0–2 dichotomous score [21]. We investigated whether increased error due to noise in the mRS indicated that the sample size targeted in SAINT I was smaller than calculated in the absence of noise. If true, then the likelihood for a spurious result is increased given an inadequate sample size.

Methods

Literature Search to Identify Stroke Randomized Clinical Trials

Two investigators (PM and JNS) independently performed structured searches in Medline to identify potentially eligible clinical stroke trials using keywords 'acute', 'ischemic', 'stroke' and 'Rankin Scale (or Score)' and reviewed all abstracts and retrieved articles for study inclusion. Studies were eligible for inclusion if they: 1) were randomized controlled stroke trials with at least 10 subjects in each study arm, 2) reported full range of mRS (0–6) outcome data in both the intervention and control group(s) at least 3 months or beyond, and 3) were published as original research manuscripts in English in a peer-reviewed journal. Two hundred and ninety-six articles were retrieved by this keyword search and subsequently reviewed, from which we identified 35 RCTs that met our inclusion criteria. Thirty-eight control arms from these 35 RCTs were then evaluated using our model to estimate misclassification error rates. For this study, we selected the control arms of these RCTs because sample size estimates for testing a novel treatment are calculated using the control arm ordinal scale outcome such as mRS along with treatment effect size [7].

Misclassification Rates with Ordinal, Collapsed Ordinal and at Various Dichotomization and Trichotomization Cut-points

To calculate the misclassification or error rates in different scenarios a custom MATLAB® program was created. Error rates are computed in three sequential steps.

Step 1: For each of the 38 placebo/control arm distributions, simulated patient populations were generated and each patient's mRS was stored as 'mRS-Observed'. Due to wide variability in number of patients from 15 to >1500 in the trials, an arbitrarily large number (n = 10000) of patients were simulated as previously used in similar studies [31,32]. A single one-step command 'repmat' is able to accomplish this task in Matlab®. This step essentially creates 10000 patients and reflects the mRS distribution

reported for each trial. Each of these 10000 patients is then assigned a Rankin score, termed mRS-Observed(j) (see File S1).

Step 2. Results of this step were passed through the Shannon's noisy channel model with van Swieten's confusion matrix serving the role of noise (Figure 1). For example a patient may have been assigned a mRS grade of 2 in step 1 but due to the effect of noise in the system may be assigned an mRS grade other than 2. The output of this step for each patient is termed the "mRS-true". At the end of this step, each of the 10000 patients will be assigned a "mRS-true(j)".

Step 3. Misclassifications are counted for each patient when there is a mismatch between the input (mRS-Observed(j) by step 1) and output (mRS-true(j) after passing through Shannon's noisy channel by step 2; Figure 1). The equation below summarizes this step for each subject 'A(j)'.

A(j) is assigned a 1 if $mRS_{observed}(j) = mRS_{true}(j)$ otherwise it is a 0.

Step 4. Total misclassification is then computed by summing across all subjects and dividing by the number of patients. This step is summarized in the form of an equation given below.

$$1-(\sum_{i=1}^{n} A(i))/n$$

Misclassification/error percentages were calculated for all 38 control arms for different scenarios that have been used in various stroke trials: full range of mRS (i.e., a full shift analysis); collapsing the higher grades of mRS 4–6 into one grade and considering mRS 0,1,2,3 as independent grades; 'dichotomizing' at four different cut-point of mRS 1,2,3,4 and for two different trichotomizations (mRS 0–1, 2–4, 5–6 and mRS 0–2, 3–4, 5–6 [33]).

A user-driven MATLAB® program is provided in File S1 that takes the mRS 0–6 distribution of a control/standard treatment arm along with a user selected confusion matrix (default of van Swieten or user-entered) and provides error percentages for the full range of ordinal scale, collapsed scale, and various dichotomizations and trichotomizations. The equation is flexible and can accommodate scales with different number of categories.

Note that van Swieten's inter-rater variability matrix was tabulated for the Rankin scale ranging from 0–5, while recent trials use the modified Rankin scale ranging from 0–6 with 6 representing death. Since there is likely low inter-rater variability in the diagnosis of death, a corresponding noise-free element was added to the van Swieten matrix.

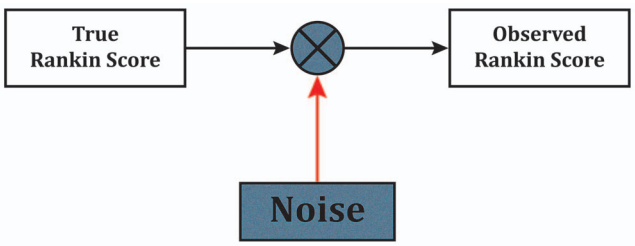

Figure 1. Shannon's information transmission model adapted to scoring of a patient on the 7 point modified Rankin Scale. A noise or error source is assumed to be in the channel between the sender represented by the 'True Rankin' score and the receiver represented by the 'Observed Rankin' score.
doi:10.1371/journal.pone.0067754.g001

Sample Sizes for SAINT I Based on Consideration of mRS Errors

Sample sizes for full ordinal scale analysis is based on an assumption of a common proportional odds across the whole range [7,8,14,32,34]. Lees et al reported that they used a common proportional odds ratio (OR) of 1.3 to derive the sample size for SAINT I [20]. Whitehead [7] and Campbell et al [8] provided equations to calculate sample sizes when a common proportional odds ratio model is used with the full ordinal scale analysis [14]. Here, the common proportional odds model is applied to the placebo/control arm to derive the sample size [7,8,14].The equation provided by Campbell et al [8] and the initial equation of Whitehead [7] does not incorporate an error term for assessment of categories.

In the presence of an error in classification in ordinal scale analysis, Whitehead [7] provides an additional subsequent example to calculate sample sizes. This example requires that the complete distribution for the reference/control arm be available. Given that the distribution of subjects in the control arm of SAINT I for different grades of mRS is available, sample size was calculated using example worked out in Section 4 of Whitehead [7] but collapsing grades mRS 5 and 6 as was done in SAINT I [20]. Since the example provided by Whitehead is quite detailed, a custom MATLAB® implementation is provided in File S1.

Other Statistical Tests

Tests of means were done by ANOVA routine supplied by Matlab®. Results of the ANOVA testing were used in post-hoc tests with a Matlab® routine 'multcompare'. This routine implements Tukeys 'honestly significant difference' criterion [35].

Results

Error Rates

The placebo/control arms of the 35 trials were processed by steps described in the Methods section and error rates for different scenarios calculated. The median NIH stroke scale, a measure of baseline stroke severity from 0 (no deficit) to 38 (coma/dead), of the 35 trials with 38 control arms ranged from 3 to 24.

If the full range of mRS is used, the misclassification error rates ranged from 7.8% to 44.4% (Table 1 and Figure 2; Mean±SD: 26.1%±5.31). Collapsing mRS grades 4 to 6 into one grade, as employed in the recently completed IST-3 trial [22] and considering the other grades as independent grades produced misclassification errors ranging from 5.9% to 44.0% (22.5%±5.66). If mRS 1 was chosen as the cut-off point, then the error rates ranged from 0% to 13.2% (6.8%±2.89). Error rates when using mRS 2 as a cut-off point were 1.7% to 24.8% (9.0%±3.33); for mRS 3 as a cut-off point the error rates ranged from 4.3% to 14.1% (7.8%±1.81); and for cut-off point of 4 the error rates ranged from 0.4% to 8.7% (3.5%±1.70). Comparison of means of error rates by ANOVA and post-hoc testing shows that the error rates were significantly different ($p<0.0001$) and all dichotomous errors lower than full range, with mRS 0–4 dichotomization error the lowest.

Error rates for dichotomization mRS 0–1 and mRS 0–2 and two trichotomizations (mRS 0–1, 2–4, 5–6:10.3%±2.75 and mRS 0–2, 3–4, 5–6:12.6%±3.13) are shown in Figure 3. Post-hoc testing showed that the trichotomizations error rate was higher when compared to the corresponding dichotomization error ($p<0.05$).

There was a wide range of calculated error rates among the different trials, from 7.8%-44.4%. Error rate in DECIMAL trial

Table 1. Error percentages for 38 studies for the full ordinal scale (mRS 0.6), partially collapsed ordinal scale (mRS 0.3, 4–6) and dichotomization (mRS 0–1, 2–6; mRS 0–2, 3–6; mRS 0–3, 4–6; mRS 0–4, 5–6) and trichotomization (mRS 0–1, 2–4, 5–6; mRS 0–2, 3–4, 5–6) cut-points.

Study [Reference Number]	mRS 0.6	mRS 0.3, 4–6	mRS 0–1, 2–6	mRS 0–2, 3–6	mRS 0–3, 4–6	mRS 0–4, 5–6	mRS 0–1, 2–4, 5–6	mRS 0–2, 3–4, 5–6
ABESTT [36]	25.64	23.37	9.41	8.62	6.24	2.27	11.68	10.89
ABESTTII [37]	23.19	20.8	8.04	7.6	6.18	2.39	10.43	9.99
ABESTTIICo [37]	27.26	24.8	9.29	9.77	6.99	2.46	11.75	12.23
ABESTTIIW [37]	31.48	29.73	9.39	14.64	7.68	1.75	11.14	16.39
ARTIS [38]	31.92	30.07	11.07	12.88	8.2	1.85	12.92	14.73
CAIST [39]	26.7	25.22	11.71	8.73	5.84	1.48	13.19	10.21
CASTA-Cereb [40]	30.4	27.44	9.39	11.49	8.24	2.96	12.35	14.45
Camerlingo [41]	24.71	18.09	5.54	7.26	6.2	6.62	12.16	13.88
Cereb-rt-pa [42]	23.45	21.94	9.73	7.3	6.07	1.51	11.24	8.81
DECIMAL [3]	7.76	5.93	0	1.68	4.25	1.83	1.83	3.51
DESTINY [4]	18.98	15.19	0	8.17	7.02	3.79	3.79	11.96
DP-b99 [43]	25.84	20.41	4.87	7.75	9.07	5.43	10.3	13.18
DP-b99-MACSI [44]	27.15	23.23	7.67	8.9	7.52	3.92	11.59	12.82
ECASSII [45]	25.4	22.53	6.48	8.59	8.2	2.87	9.35	11.46
ECASSIII [46]	22.09	19.37	8.93	6.56	4.96	2.72	11.65	9.28
EPITHET [47]	27.83	23.9	7.85	9.1	8.16	3.93	11.78	13.03
EPO [48]	26.13	20.76	7.08	7.1	7.42	5.37	12.45	12.47
Edaravone [49]	29.37	26.69	13.2	9.22	6.16	2.68	15.88	11.9
Enlimomab [50]	24.61	21.55	7.02	8.75	6.91	3.06	10.08	11.81
FIST [51]	24.14	20.39	5.15	9.33	7.03	3.75	8.9	13.08
GAIN [52]	26.19	21.55	5.61	8.02	8.93	4.64	10.25	12.66
HAMLET [5]	16.99	15.62	1.79	7.43	7.18	1.37	3.16	8.8
ICTUS [53]	25.14	21.24	4.85	8.53	8.93	3.9	8.75	12.43
IMS-III [54]	25.54	22.49	7.35	9.26	6.9	3.05	10.4	12.31
INSULINFARCT [55]	29.59	27.27	8.66	11.46	9.11	2.32	10.98	13.78
IST-3 [22]	23.05	18.38	6.31	7.56	5.52	4.67	10.98	12.23
MELT [56]	32.31	25.52	7.45	10.83	8.58	6.79	14.24	17.62
MR-RESCUE-Pen [57]	32.33	26.17	4.45	11.1	11.22	6.16	10.61	17.26
MR-RES-Non-Pen [57]	25.97	17.26	1.89	6.44	9.28	8.71	10.6	15.15
Minocycline [58]	44.37	43.99	6.99	24.82	14.12	0.38	7.37	25.2
NEST-1 [59]	27.22	22.53	7.2	8.53	7.43	4.69	11.89	13.22
NEST-2 [60]	29.17	25.13	5.69	10.94	9.84	4.04	9.73	14.98
NINDS [61]	23.96	20.57	6.19	7.88	7.45	3.39	9.58	11.27
PROACTII [62]	24.09	20.31	2.85	8.89	9.16	3.78	6.63	12.67
SAINTI [20]	25.08	21.71	7.6	7.13	7.96	3.37	10.97	10.5
SAINTII [21]	26.24	22.5	7.3	9.02	7.26	3.74	11.04	12.76
Synthesis [63]	25.16	22.32	5.02	7.32	10.25	2.84	7.86	10.16
Synthesis Exp [64]	26.51	22.88	7.12	8.18	8.35	3.63	10.75	11.81

doi:10.1371/journal.pone.0067754.t001

[3] using the full scale mRS 0–6 was the lowest (7.8%). This is likely due to lower proportion of patients (22%) in the most uncertain grades (mRS 2–4) and the remaining (78%) being in a non-uncertainty- prone state of mRS 6 (i.e., deceased). Error rate in the Minocycline trial [58], for the full scale, was the highest (44.4%), possibly since only 14% were in the low uncertainty-prone grades (mRS 0–1), while, 81% were in the higher uncertain grades (mRS 2–4).

In place of van Swieten's confusion matrix, the Wilson et al [26] matrix was applied and the above steps repeated resulting in higher error rates (Figure 4).

Post-hoc testing shows that each error rate with this matrix is higher than the corresponding error using van Swieten's confusion matrix except that the error rates for mRS 4 dichotomizations have overlapping confidence intervals. Wilson et al proposed a modification of the mRS called the mRS-Structured Interview (mRS-SI) and also provided a confusion matrix [26]. The errors

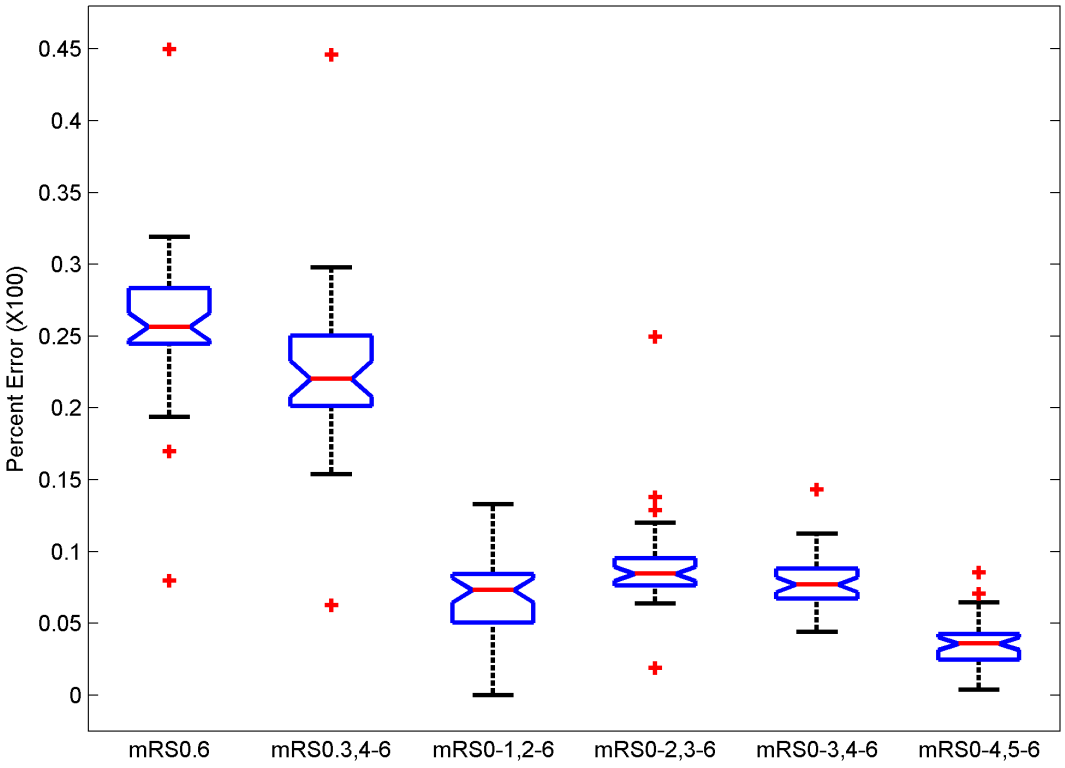

Figure 2. Box plots of error rates for the full ordinal scale of mRS (mRS 0.6), considering mRS 0 to 3 as individual grades and collapsing mRS grades 4 to 6 (mRS 0.3,4–6), dichotomizing at various cut-points of mRS 1 (mRS 0–1, 2–6), mRS 2 (mRS 0–2, 3–6), mRS 3 (mRS 0–3, 4–6) and mRS 4 (mRS 0–4, 5–6). van Swieten's inter-rater reliability matrix used as confusion matrix. ($p<.001$ ANOVA; post-hoc testing shows that all dichotomization errors are lower than either full scale errors with mRS 0–4 dichotomization the lowest; $p<.05$).
doi:10.1371/journal.pone.0067754.g002

calculated with this confusion matrix were lower than with the original mRS, however, the errors for the full range and collapsed ranges are still significantly higher than the errors with dichotomization ($p<0.001$; data shown as Figure S2 in File S2).

Calculation of Sample Size Incorporating mRS Measurement Errors

SAINT I [20] trial reported that, by applying ordinal analysis, the treatment arm showed efficacy vis-à-vis the placebo arm. A total of 1722 subjects were enrolled into two arms (861 in each arm) of SAINT I. However, applying the transformation matrix from van Swieten to account for misclassification and utilizing the available SAINT I placebo arm distribution,1070 subjects would be required in each arm to reliably estimate effects -nearly 24% more subjects than actually randomized (compare blue star to red star in Figure S3 in File S2). These calculations were repeated for SAINT II [21] employing their assumptions of a proportional odds ratio of 1.2 and power of 80%. SAINT II randomized 1621 patients to the placebo arm. If mRS misclassification was taken into account and using their mRS distribution, 1665 subjects would be needed, a difference of only 2.7%.

Discussion

Clinical trials with subjective functional assessments have presented a variety of challenges. In the case of stroke, many clinical trial difficulties stem from issues such as heterogeneity of baseline factors, spontaneous recovery and subjective nature of assessing stroke severity and outcomes particularly given uncertainties in classification of outcomes [65,66]. We show here that one such uncertainty, an asymmetrical distribution of misclassification in the mRS, introduces the need for more subjects to accommodate the potential biases in inferences about study effects that may occur if these uncertainties are not equally distributed**.** We include a set of Matlab programs (in File S1) that can be used in the future to estimate error rates and sample sizes using outcome scales. These programs are flexible in terms of categories and can be used with other outcome scales as long as the confusion matrix or equivalent is known. Note that while error estimates are important in estimating sample size, the lowest error configuration is not necessarily the best one if it does not capture the necessary range of expected outcomes. So for example, in a study of mild stroke, mRS 0–4, 5, 6 might be the lowest error, but miss important changes at the excellent outcome (mRS 0–1) range.

We performed an analysis of the influence of mRS misclassification on the expected error rates and applied this model to the empirical data derived from actual stroke clinical trials. We determined the influence of variability in mRS assessments on the overall misclassification error rates calculated for 38 individual control arms and showed that the error rates were highest when either the full-scale or collapsed full-scale (as in IST-3 [22], SAINT I [20], and SAINT II [21]) of mRS was considered as compared to dichotomization at cut-off points of mRS 1 and mRS 2. Using the SAINT I trial as an example, we demonstrated that when mRS misclassification uncertainties are taken into account, a higher sample size is required using the "shift" approach. Hence, SAINT I may have randomized 24% too few patients taking errors into consideration, thus, possibly accounting for the discrepant results between SAINT I and the larger SAINT II trial. There are other possible explanations for discrepant results between the two trials

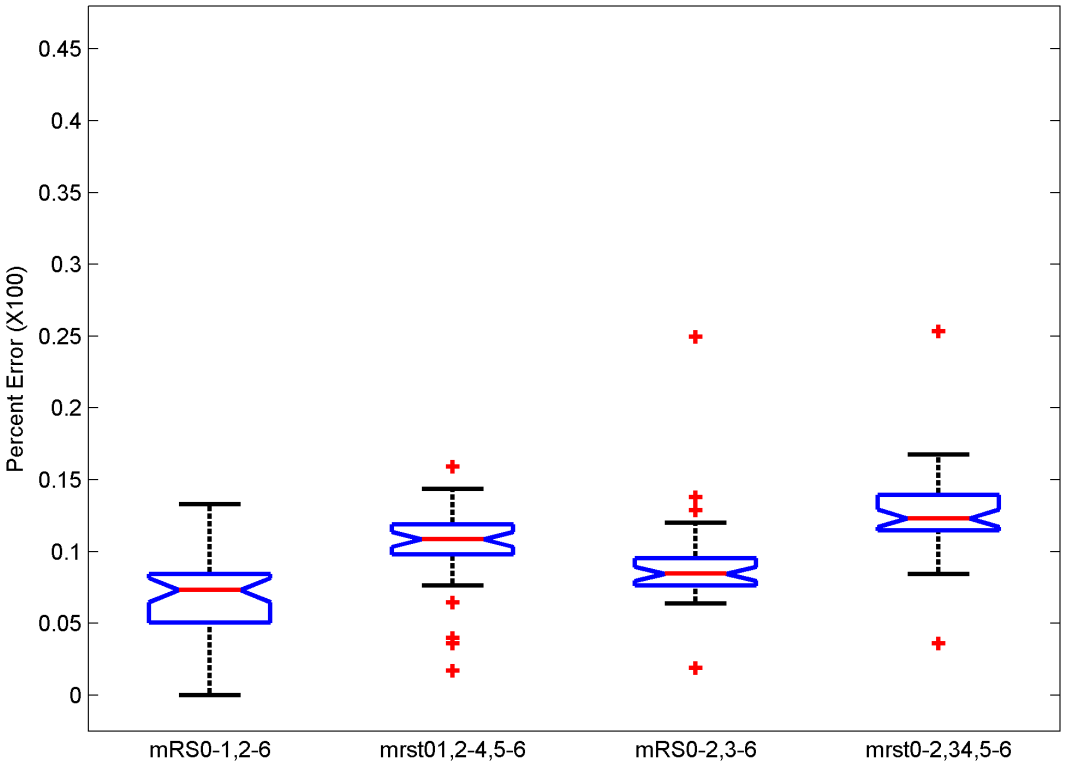

Figure 3. Box plots of error rates for dichotomizing at cut-point of mRS 1 (mRS 0–1, 2–6), trichotomizing at cut-points 1 and 4 (mRS 0–1, 2–4, 5–6), dichotomizing at mRS 2 (mRS 0–2, 3–6), and trichotomizing at cut-points 2 and 4 (mRS 0–2, 3–4, 5–6). van Swieten's inter-rater reliability matrix used as confusion matrix. Post-hoc testing shows that both trichotomization errors are higher than dichotomization ($p<.05$).
doi:10.1371/journal.pone.0067754.g003

and we cannot prove that inadequate sample size was the primary factor. However, the larger SAINT II employed a lower proportional odds ratio of 1.2 with a lower power (80%) and there was a marginal difference of 2.7% between actual sample size and that required by taking into account misclassification in mRS.

The actual error rates found depend on the range of the mRS in each trial because the uncertainty in misclassification is not evenly distributed across the entire range. While there is considerable evidence that there is loss of information when a scale is dichotomized at the median [14], it is not clear that the advantage of use of the wider range will always overcome the noise that it appears to generate.

Our results echo the concept put forward by Whitehead [7], that the advantage of decreased sample size with ordinal scale is lost if there are errors even modest in classification. He calculated that a uniform error of 20% in a hypothetical four-category scale increased the sample size requirements by more than 60%. Whitehead's projection was qualitatively confirmed here with real world mRS uncertainties and data derived from clinical trials. Misclassification of ordinal scale data leading to loss of power in statistical tests has been known for several decades [67].

It can be argued, from a strict information theory perspective, that misclassification error rates obtained by analyzing with the full-scale are not directly comparable with error rates obtained with the dichotomized approaches, since, there are different numbers of variables or "bits". To address this potential criticism, a normalized error per bit of information transmitted (or entropy) was calculated [see details in File S2]. After normalization with entropy, rates, while overall lower, were still higher with full-scale analysis approach. Note, however, that entropy normalization reflects the error per bit of information transmitted, but does not influence the error factor that needs to be considered for sample size determination, that is the much higher value shown here.

The inter-rater reliability matrix proposed by van Swieten et al [24] was derived from an assessment of 100 patients by pairs of physicians selected from a pool of 34. These 34 physician raters were either senior neurologists or resident physicians. This situation may not reflect the actual clinical trial environment where typically there are 100 s of patient subjects and raters with various educational backgrounds spread across several continents [27–29]. Wilson et al [26] study used two neurologists, one stroke physician, seven nurses and four physiotherapists. The inter-rater reliability in the Wilson et al study was lower than van Swieten's and resulted in higher error rates (Figure 4) when analyzed with the Shannon Entropy model compared to the van Swieten confusion matrix (Figure 2).

Other alternatives to van Swieten's inter-rater reliability table are not without limitations. Some of these publications did not report evaluations at the lower and higher ends of mRS [25,28], while others had fewer clinical assessors [25–27,68,69], and fewer patients [27,68,69] or, lacked face-to-face interviews. More attention needs to be given to the reliability of different implementation methods for rating outcomes, including centralized rating methods and incorporation of their errors into sample size estimation. Ideally, a comparison between a 'typical' assessor and a gold-standard 'expert' could be used. However, it is unclear if two 'experts' would agree on the assignment of a mRS grade to a patient given that studies on inter-rater reliability have reported kappa values ranging from 0.25 to 0.95 [27]. Additionally,

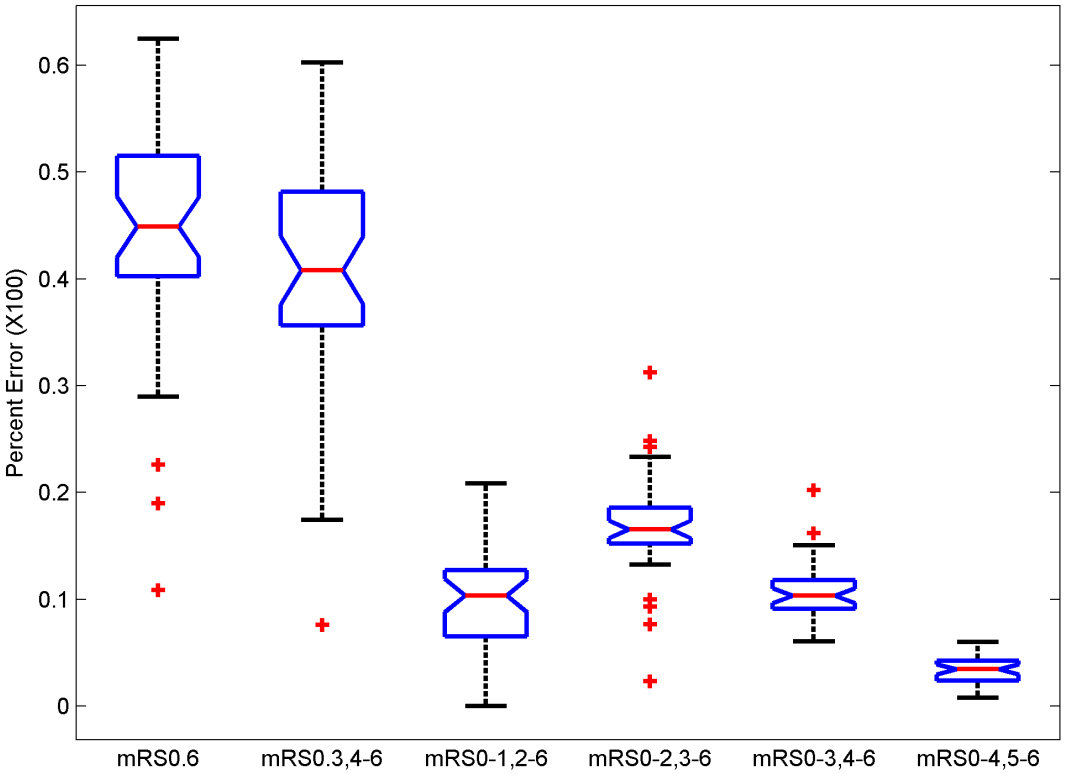

Figure 4. Inter-rater reliability matrix for mRS from Wilson et al [26] used as confusion matrix. Box plots of error rates for the full ordinal scale of mRS (mRS 0.6), considering mRS 0 to 3 as individual grades and collapsing mRS grades 4 to 6 (mRS 0.3,4–6), dichotomizing at a cut-point of mRS 1 (mRS 0–1, 2–6), dichotomizing at a cut-point of mRS 2 (mRS 0–2, 3–6), mRS 3 (mRS 0–3, 4–6) and mRS 4 (mRS 0–4, 5–6). Post-hoc testing shows that each error rate with this matrix is higher than the corresponding error using van Swieten's confusion matrix except the error rates for mRS 4.
doi:10.1371/journal.pone.0067754.g004

evidence suggests that disagreement can still persist after training the typical assessor, and then, comparing his/her score against that of an expert [28]. Over the last decade there have been attempts at improving the reliability of mRS assessment with the aid of a structured interview [26], questionnaire [68], and a focused assessment [69], although replication of these improvements have been inconsistent [26,29].

While our focus in this study has been the mRS, this same analysis can be extended to other ordinal scales employed in clinical trials. For example, Glasgow Outcome Scale (GOS), used in traumatic brain injury trials and infrequently in stroke trials, also demonstrate comparable error rates in the mid-range of the scale [70–72].

In conclusion, using stroke trials as an example, we demonstrated that misclassification error rates are overall higher with variations on the 'shift' analysis compared to dichotomization approach. We also demonstrated that the 'shift' analysis can lead to the need for higher sample size in the setting of misclassification. Selecting an appropriate sample size, while important, is difficult in the setting of uncertainties in measurement [73]. We found that in the case of mRS as the outcome measure, dichotomous outcomes are more reliable. Therefore, if ordinal analysis is employed, errors should be explicitly considered in sample size determination. In principle, we have outlined an approach to error estimation for any condition in which there are uncertainties in outcome assessment and provided the user with a set of Matlab programs to incorporate errors into sample size estimation. The relative advantage of dichotomizing vs. ordinal analysis will depend on the distribution of these uncertainties and the frequency of their occurrence under the specific conditions of the trial.

Supporting Information

File S1.
(DOCX)

File S2.
(DOCX)

Author Contributions

Conceived and designed the experiments: PM CSK TAK. Performed the experiments: PM CSK JNS TAK. Analyzed the data: PM CSK JNS. Wrote the paper: PM CSK JNS DLW TAK. Designed Software: CSK PM.

References

1. Quinn TJ, Dawson J, Walters MR, Lees KR (2009) Functional outcome measures in contemporary stroke trials. Int J Stroke 4: 200–205.
2. Tilley BC, Marler J, Geller N, Lu M, Legler J, et al. (1996) Using a global test for multiple outcomes in stroke trials with application to the NINDS t-PA Stroke Trial. Stroke 27: 2136–2141.
3. Vahedi K, Vicaut E, Mateo J, Kurtz A, Orabi M, et al. (2007) Sequential-design, multicenter, randomized, controlled trial of early decompressive craniectomy in malignant middle cerebral artery infarction (DECIMAL Trial). Stroke 38: 2506–2517.

4. Jüttler E, Schwab S, Schmiedek P, Unterberg A, Hennerici M, et al. (2007) Decompressive Surgery for the Treatment of Malignant Infarction of the Middle Cerebral Artery (DESTINY): a randomized, controlled trial. Stroke 38: 2518–2525.
5. Hofmeijer J, Kappelle LJ, Algra A, Amelink GJ, van Gijn J, et al. (2009) Surgical decompression for space-occupying cerebral infarction (the Hemicraniectomy After Middle Cerebral Artery infarction with Life-threatening Edema Trial [HAMLET]): a multicentre, open, randomised trial. The Lancet Neurol 8: 326–333.
6. Points to consider on clinical investigation of medicinal products for the treatment of acute stroke. The European Agency for the Evaluation of Medicinal Products (2001) Available: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003342.pdf.Accessed 2013 May 29.
7. Whitehead J (1993) Sample size calculations for ordered categorical data. Statistics in Medicine 1993;12: 2257–2271.
8. Campbell MJ, Julious SA, Altman DG (1995) Estimating sample sizes for binary, Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ 311: 1145–1148.
9. Tilley BC (2012) Contemporary outcome measures in acute stroke research. Stroke 43: 935–937.
10. Savitz SI, Lew R, Bluhmki E, Hacke W, Fisher M (2007) Shift analysis versus dichotomization of the modified Rankin Scale outcome scores in the NINDS and ECASS-II trials. Stroke 38: 3205–3212.
11. Saver JL (2007) Novel-end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. Stroke 38: 3055–3062.
12. Saver JL, Gornbein J (2009) Treatment effect for which shift or binary analyses are advantageous in acute stroke trials. Neurology 72: 1310–1315.
13. Bath PMW, Lees KR, Schellinger PD, Altman H, Bland M, et al. (2012) Statistical Analysis of the Primary Outcome in Acute Stroke Trials. Stroke 43: 1171–1118.
14. Altman DG, Royston P (2006) The cost of dichotomizing continuous variables. BMJ 332: 1080.
15. Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana IL: University of Illinois Press. Pp4–28 and 31–80.
16. Cover TM, Thomas JA (1991) Introduction and Preview. In: Elements of Information Theory. New York: John Wiley & Sons. Pp 1–11.
17. Outcomes Working Group. The Optimizing Analysis of Stroke Trials (OAST) (2007) Can We Improve the Statistical Analysis of Stroke Trials? Statistical Reanalysis of Functional Outcomes in Stroke Trials Collaboration. Stroke 38: 1911–1915.
18. Saver JL (2011) Optimal end points for acute stroke therapy trials. Best ways to measure treatment effects of drugs and devices. Stroke 42: 2356–2362.
19. Cobo E, Secades JJ, Miras F, González JA, Saver JL, et al. (2010) Boosting the chances to improve stroke treatment. Stroke 41: 3143–3150.
20. Lees KR, Zivin JA, Ashwood T, Davalos A, Davis SM, et al. (2006) NXY-059 for acute ischemic stroke. N Engl J Med 354: 588–600.
21. Shuaib A, Lees KR, Lyden P, Grotta J, Davalos A, et al. (2007) NXY-059 for the treatment of acute ischemic stroke. N Engl J of Med 357: 562–571.
22. IST-3 collaborative group, Sandercock P, Wardlaw JM, Lindley RI, Dennis M, Cohen G, et al. (2012) The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. Lancet 379: 2352–2363.
23. Lai SM, Duncan PW (2001) Stroke Recovery Profile and the Modified Rankin Assessment. Neuroepidemiology. 20: 26–30.
24. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J (1988) Interobserver agreement for the assessment of handicap in stroke patients. Stroke 19: 604–607.
25. Wilson JT, Hareendran A, Grant M, Baird T, Schulz UG, et al. (2002) Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin scale. Stroke 33: 2243–2246.
26. Wilson JT, Hareendran A, Hendry A, Potter J, Bone I, et al. (2005) Reliability of the modified Rankin scale across multiple raters: benefits of a structured interview. Stroke 36: 777–781.
27. Quinn TJ, Dawson J, Walters MR, Lees KR (2009) Exploring the reliability of the modified Rankin scale. Stroke 40: 762–766.
28. Quinn TJ, Dawson J, Walters MR, Lees KR (2008) Variability in modified Rankin scoring across a large cohort of international observers. Stroke 39: 2975–2979.
29. Quinn TJ, Macarthur K, Dawson J, Walters MR, Lees KR (2010) Reliability of Structured Modified Rankin Scale Assessment. Stroke 41: e602.
30. Erdogmus D, Xu D, Hild II K (2010) Classification with EEC, Divergence Measures and Error Bounds. In: Principe JC Editor. Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives. New York: Springer. 219–261.
31. Koziol JA, Feng AC (2006) On the analysis and interpretation of outcome measures in stroke clinical trials: lessons from the SAINT I study of NXY-059 for acute ischemic stroke. Stroke 37: 2644–2647.
32. Howard G, Waller JL, Voeks JH, Howard VJ, Jauch EC et al. (2012) A simple, assumption-free, and clinically interpretable approach for analysis of modified Rankin outcomes. Stroke 43: 664–669.
33. Haley EC, Thompson JL, Grotta JC, Lyden PD, Hemmen TG, et al. (2010) Phase IIB/III trial of tenecteplase in acute ischemic stroke: Results of a prematurely terminated randomized clinical trial. Stroke 41: 707–711.
34. Hall CE, Mirski M, Palesch YY, Diringer MN, Qureshi AI, et al. (2012) Clinical Trial Design in the Neurocritical Care Unit. Neurocrit Care 1: 6–19.
35. Statistics Toolbox Users Manual. The Mathworks Corporation. Available: http://www.mathworks.com/help/stats/multcompare.html. Accessed 2013 May 29.
36. Abciximab Emergent Stroke Treatment Trial (AbESTT) Investigators. (2005) Emergency administration of abciximab for treatment of patients with acute ischemic stroke: results of a randomized phase 2 trial. Stroke 36: 880–890.
37. Adams HP, Effron MB, Torner J, Dávalos A, Frayne J, et al. (2008) Emergency administration of abciximab for treatment of patients with acute ischemic stroke: results of an international Phase III trial: Abciximab in Emergency Treatment of Stroke Trial (AbESTT-II). Stroke 39: 87–99.
38. Zinstok SM, Roos YB; ARTIS Investigators (2012) Early administration of aspirin in patients treated with alteplase for acute ischaemic stroke: a randomised controlled trial. Lancet 380;731–737.
39. Lee YS, Bae HJ, Kang DW, Lee SH, Yu K, et al. (2011) Cilostazol in Acute Ischemic Stroke Treatment (CAIST Trial): a randomized double-blind non-inferiority trial. Cerebrovasc Dis 32: 65–71.
40. Heiss WD, Brainin M, Bornstein NM, Tuomilehto J, Hong Z, et al. (2012) Cerebrolysin in patients with acute ischemic stroke in Asia: results of a double-blind, placebo-controlled randomized trial. Stroke 43: 630–636.
41. Camerlingo M, Salvi P, Belloni G, Gamba T, Cesana BM, et al. (2005) Intravenous heparin started within the first 3 hours after onset of symptoms as a treatment for acute nonlacunar hemispheric cerebral infarctions. Stroke 36: 2415–2420.
42. Lang W, Stadler CH, Poljakovic Z, Fleet D; Lyse Study Group (2013) A prospective, randomized, placebo-controlled, double-blind trial about safety and efficacy of combined treatment with alteplase (rt-PA) and Cerebrolysin in acute ischaemic hemispheric stroke. Int J Stroke 8: 95–104.
43. Diener HC, Schneider D, Lampl Y, Bornstein NM, Kozak A, et al. (2008) DP-b99, a membrane-activated metal ion chelator, as neuroprotective therapy in ischemic stroke. Stroke 39: 1774–1778.
44. Lees KR, Bornstein N, Diener HC, Gorelick PB, Rosenberg G, et al. (2013) Results of membrane-activated chelator stroke intervention randomized trial of DP-b99 in acute ischemic stroke. Stroke 44: 580–584.
45. Hacke W, Kaste M, Fieschi C, von Kummer R, Davalos A, et al. (1998) Randomized double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). Second European–Australasian Acute Stroke Study Investigators. Lancet 352: 1245–1251.
46. Hacke W, Kaste M, Bluhmki E, Brozman M, Davalos A, et al. (2008). Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. N Engl J Med 359: 1317–1329.
47. Davis SM, Donnan GA, Parsons MW, Levi C, Butcher KS, et al. (2008) Effects of alteplase beyond 3 h after stroke in the Echoplanar Imaging Thrombolytic Evaluation Trial (EPITHET): a placebo-controlled randomised trial. Lancet Neurol 7: 299–309.
48. Ehrenreich H, Weissenborn K, Prange H, Schneider D, Weimar C, et al. (2009) Recombinant human erythropoietin in the treatment of acute ischemic stroke. Stroke 40: e647–656.
49. Shinohara Y, Saito I, Kobayashi S, Uchiyama S (2009) Edaravone (radical scavenger) versus Sodium Ozagrel (anti-platelet) in cardioembolic ischemic stroke (EDO trial). Cerebrovasc Dis 27: 485–492.
50. Enlimomab Acute Stroke Trial Investigators (2001) Use of anti-ICAM-1 therapy in ischemic stroke: results of the Enlimomab Acute Stroke Trial. Neurology 57: 1428–1434.
51. Franke CL, Palm R, Dalby M, Schoonderwaldt HC, Hantson L, et al. (1996) Flunarizine in stroke treatment (FIST): a double blind placebo-controlled trial in Scandinavia and the Netherlands. Acta Neurol Scand 93: 56–60.
52. The North American Glycine Antagonist in Neuroprotection (GAIN) Investigators (2000) Phase II Studies of the Glycine Antagonist GV150526 in Acute Stroke The North American Experience. Stroke 31: 358–365.
53. Dávalos A, Alvarez-Sabin J, Castillo J, Díez-Tejedor E, Ferro J, et al. (2012) Citicoline in the treatment of ischaemic stroke: an international, randomized, multicentre, placebo-controlled study (ICTUS trial). Lancet 380(9839): 349–57.
54. Broderick JP, Palesch YY, Demchuk AM, Yeatts SD, Khatri P, et al. (2013) Endovascular therapy after t-PA alone versus t-PA alone for stroke. N Engl J Med 368: 893–903.
55. Rosso C, Corvol JC, Pires C, Crozier S, Attal Y, et al. (2012) Intensive versus subcutaneous insulin in patients with hyperacute stroke: results from the randomized INSULINFARCT trial. Stroke 43: 2343–2349.
56. Ogawa A, Mori E, Minematsu K, Taki W, Takahashi A, et al. (2007) Randomized trial of intraarterial infusion of urokinase within 6 hours of middle cerebral artery stroke: the middle cerebral artery embolism local fibrinolytic intervention trial (MELT) Japan. Stroke 38: 2633–2639.
57. Kidwell CS, Jahan R, Gornbein J, Alger JR, Nenov V, et al. (2013) A trial of imaging selection and endovascular treatment for ischemic stroke. N Engl J Med 368: 914–923.
58. Padma Srivastava MV, Bhasin A, Bhatia R, Garg A, Gaikwad S, et al. (2012) Efficacy of minocycline in acute ischemic stroke: a single blinded, placebo-controlled trial. Neurol India 60: 23–28.

59. Lampl Y, Zivin JA, Fisher M, Lew R, Welin L, et al. (2007) Infrared laser therapy for ischemic stroke: a new treatment strategy: results of the Neuro Thera Effectiveness and Safety Trial-1 (NEST-1). Stroke 38: 1843–1849.
60. Zivin JA, Albers GW, Bornstein N, Chippendale T, Dahlof B, et al. (2009) Effectiveness and safety of transcranial laser therapy for acute ischemic stroke. Stroke 40: 1359–1364.
61. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group (1995) Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 333: 1581–1587.
62. Furlan AJ, Higashida R, Wechsler L, Gent M, Rowley H, et al. (1999) Intra-arterial prourokinase for acute ischemic stroke. The PROACT II study: a randomized controlled trial Prolyse in Acute Cerebral Thromboembolism. JAMA 282: 2003–2011.
63. Ciccone A, Valavassori L, Ponzio M, Ballabio E, Gasparotti R, et al. (2010). Intra-arterial or intravenous thrombolysis for acute ischemic stroke? The SYNTHESIS pilot trial. J Neuro Intervent Surg 2: 74–79.
64. Ciccone A, Valavassori L, Nichelatti M, Sgoifo A, Ponzio M, et al. (2013) Endovascular treatment for acute ischemic stroke. N Engl J Med 368: 904–913.
65. Mandava P, Kalkonde YV, Rochat R, Kent TA (2010) A matching algorithm to address imbalances in study populations: application to the National Institute of Neurological Diseases and Stroke Recombinant Tissue Plasminogen Activator acute stroke trial. Stroke 41: 765–770.
66. Mandava P, Krumpelman CS, Murthy SB, Kent TA (2012) A critical review of stroke trial analytical methodology: outcome measures, study design. In Lapchak PA and Zhang JH Editors. Translational Stroke Research: From Target Selection to Clinical Trials. New York: Springer. pp833–60.
67. Mote VL, Anderson RL (1965) An investigation of the effect of misclassification on the properties of $\chi 2$-tests in the analysis of categorical data. Biometrika 52: 95–109.
68. Bruno A, Shah N, Lin C, Close B, Hess DC, et al. (2010) Improving modified Rankin Scale assessment with a simplified questionnaire. Stroke 41: 1048–1050.
69. Saver JL, Filip B, Hamilton S, Yanes A, Craig S, et al. (2010) Improving the reliability of stroke disability grading in clinical trials and clinical practice: the Rankin Focused Assessment (RFA). Stroke 41: 992–995.
70. Maas AIR, Braakman R, Schouten HJA, Minderhoud JM, van Zomeren AH (1983) Agreement between physicians on assessment of outcome following severe head injury. J. Neurosurg. 58: 321–325.
71. Wilson JT, Pettigrew LEL, Teasdale GM (1998) Structured interviews for Glasgow Outcome scale and the extended Glasgow Outcome Scale: Guidelines for their use J Neurotrauma15: 573–585.
72. Wilson JT, Slieker FJ, Legrand V, Murray G, Stocchetti N, et al. (2007) Observer variation in the assessment of outcome in traumatic brain injury: experience from a multicenter, international randomized clinical trial. Neurosurgery 61: 123–128.
73. Norman G, Monteiro S, Salama S (2012) Sample size calculations: should the emperor's clothes be off the peg or made to measure? BMJ 345; e5278.