Ordinal outcome analysis improves the detection of between-hospital differences in outcome

Iris E. Ceyisakar (iris.kohen@gmail.com)
Erasmus MC https://orcid.org/0000-0002-9081-6278
Nikki van Leeuwen
Erasmus MC
Diederik W.J. Dippel
Erasmus MC
Ewout W. Steyerberg
Leids Universitair Medisch Centrum
Hester F. Lingsma
Erasmus MC

Research article

Keywords: between-hospital variation, observational data, comparative effectiveness research, statistical power, ordinal outcome analysis, proportional odds analysis, benchmarking

DOI: https://doi.org/10.21203/rs.3.rs-18822/v2

Abstract

Background There is a growing interest in assessment of the quality of hospital care, based on outcome measures. Many quality of care comparisons rely on binary outcomes, for example mortality rates. Due to low numbers, the observed differences in outcome are partly subject to chance.

Methods We aimed to quantify the gain in efficiency by ordinal instead of binary outcome analyses for hospital comparisons. We analyzed patients with traumatic brain injury (TBI) and stroke as examples. We sampled patients from two trials. We simulated ordinal and dichotomous outcomes based on the modified Rankin Scale (stroke) and Glasgow Outcome Scale (TBI) in scenarios with and without true differences between hospitals in outcome. The potential efficiency gain of ordinal outcomes, analyzed with ordinal logistic regression, compared to dichotomous outcomes, analyzed with binary logistic regression was expressed as the possible reduction in sample size while keeping the same statistical power to detect outliers.

Results In the IMPACT study (8,799 patients in 265 hospitals, mean number of patients per hospital = 36), the analysis of the ordinal scale rather than the dichotomized scale ('unfavorable outcome'), allowed for up to 32% less patients in the analysis without a loss of power. In the PRACTISE trial (1,657 patients in 12 hospitals, mean number of patients per hospital = 138), ordinal analysis allowed for 13% less patients. Compared to mortality, ordinal outcome analyses allowed for up to 37% to 63% less patients.

Conclusions Ordinal analyses provide the statistical power of substantially larger studies which have been analyzed with dichotomization of endpoints. We advise to exploit ordinal outcome measures for hospital comparisons, in order to increase efficiency in quality of care measurements.

Background

There is an ever-growing demand for information on performance of hospitals to improve quality of care[1]. Clinical outcomes are commonly used to determine which hospitals are allegedly performing better or worse, and which are to be labelled as potential outliers[2,3]. However, comparing outcomes between hospitals has its limitations. The observed differences in outcome between hospitals are often partly due to chance[4] and are only partly explained by actual differences in the quality of care[5]. Lack of power to detect differences between hospitals is a common problem for several clinically relevant outcome indicators. For example, complication rates are generally low and the small number of events leads to underpowered statistical analyses[6,7]. Furthermore one of the most commonly used clinical measures is the (standardized) mortality ratio (SMR)[8,9], which has a variety of disadvantages and methodological issues when used as a quality of care measure[10–12]. The main issue being that mortality is an especially rare outcome in many patient groups, leading to low power when trying to detect hospitals with aberrant outcomes[13].

Many clinical continuous or ordinal outcome scales do exist and are recorded, but these are often dichotomized (favorable and unfavorable) in quality of care comparisons, for reasons of simplicity.

Examples of ordinal outcome measures are the modified Rankin Scale (for stroke), the Glasgow Outcome Scale (for Traumatic Brain Injury (TBI), the Guillain Barré syndrome disability score, the NYHA Functional Classification (for heart failure) and the Rutherford Classification (for peripheral artery disease). Dichotomization has been shown to lead to a loss of clinically and statistically relevant information in several studies[14–17] while analysis on the full ordinal scale with proportional odds analysis, prevents this loss of information[18,19]. Simulation studies and empirical validation studies in clinical trials have demonstrated that ordinal analysis increases statistical power compared to binary outcome analysis[18–21]. For clinical trials it has already been advised not to dichotomize ordinal outcome scale but to exploit the full ordinal nature of the scale, to allow for detection of smaller treatment effects[15,19]. However, this potential gain in efficiency has not been assessed for hospital comparisons.

Our aim, therefore, is to quantify the gain in power, or reduction in sample size, that can be achieved by using ordinal compared to dichotomous outcomes as a measure of quality of care for hospital comparisons.

Methods

Simulation studies were performed with patients sampled from two databases. The databases consisted of hospital data of patients with either TBI in the International Mission on Prognosis And Clinical Trial Design in Traumatic Brain Injury (IMPACT) study[22], and stroke patients in the PRomoting ACute Thrombolysis in Ischemic StrokE (PRACTISE) trial[23].

The sampled patients were appointed to one of the 250 fictitious hospitals. The simulations included two scenarios, one in which the hospital influenced the outcome of the patient (A) and one in which outcome of the patient was completely independent of the hospital (B). In the first scenario (A) hospitals were given a “center effect”; a coefficient for the effect of hospital on outcome drawn from a normal distribution with m=0 and SD= 0.35. The true outcomes of the hospitals all differed from 0, as can be seen in Figure 1a. This meant that patients from one hospital had a higher chance of a good outcome than those of another[20], i.e. that ‘true’ hospital differences in outcome existed. Analysis of scenario A was done to determine the sensitivity (type II error) which could be achieved using either ordinal outcomes or dichotomized outcomes. Since all hospitals were assigned a center effect, the analysis which found the most hospitals with performances deviating from the mean had the best sensitivity for an effect.

In scenario B specificity was tested (Figure 1b) by checking if the analyses did not find more than 5% of differently performing hospitals when there was no true difference in hospital performance.

To simulate outcomes, a multinomial generalized logit regression model was fitted to predict the probability for outcomes for each patient based on the given baseline covariates. Furthermore, in scenario A the probability for outcome was either increased or decreased depending on the hospital the patient was in.

For the baseline covariates of the TBI patients well know prognostics baseline characteristics were used: Glasgow Coma Scale (GCS) motor score, age and pupillary reactivity (both pupils reactive, one pupil reactive, no pupil reactivity)[22]. In the stroke data the following covariates were used: baseline National Institute of Health Stroke Scale (NIHSS) score, age, history of ischemic stroke, atrial fibrillation, an diabetes mellitus[23].

In TBI, we used the 5-point ordinal Glasgow Outcome Scale (GOS) at 6-months as an outcome measure (Figure 2a). For stroke, we used the modified Rankin Scale (mRS) at 3 months, a 7-point ordinal scale (Figure 2b). These are the most commonly used outcome measures for these conditions. In both scales, the worst disability state and death were combined for ethical reasons, resulting in a 4-point outcome scale for TBI, and a 6-point outcome scale for stroke[19,23,24]. Both ordinal outcome measures were thereafter dichotomized, into favorable (good recovery or moderate disability) and unfavorable outcome (severe disability, vegetative state and death) as well as dichotomized for mortality (including severe disability). Dichotomization for mortality was done to illustrate the case in which only mortality rates are measured.

To demonstrate the differences in sensitivity to detect hospital outliers the simulation was repeated with different number of patients per hospital, ranging from 25 to 200, which were distributed over 250 hospitals. Simulations were run 500 times.

Analysis

Outcomes were analyzed on (1) an ordinal scale, (2) dichotomized as favorable vs. unfavorable outcomes, and (3) dichotomization for death vs. alive. The binary outcomes were analyzed with standard fixed effect logistic regression models, the ordinal outcomes were analyzed with proportional odds fixed effect logistic regression models[25,26]. All models were adjusted for previously mentioned baseline covariates based on which the outcomes had been predicted, and included hospital as a categorical variable. This yielded an estimated center effect per hospital compared to mean center effects. Hospitals with predicted center effect values outside the 95% confidence intervals (CIs) of the overall mean were scored as outliers.

The ability of the model to determine which hospitals were outliers was measured by counting how many outliers the analysis would find in different scenarios. This means that in scenario A, the analysis which found the most outliers was determined as the most sensitive, and in scenario B the analyses were meant to have less than 5% outliers.

Higher power in the analyses results in higher rate of correctly identified outliers. Therefore, we could translate the ability to find outliers to the possibility of sample size reduction. The ability of regression models to determine which hospitals had aberrant outcomes, given dichotomized and ordinal outcomes was expressed in potential efficiency gains. The difference between ordinal and dichotomized outcomes was expressed as potential efficiency gain: the possible reduction in sample size while keeping the same

statistical power to detect outliers. All analyses were done using R Statistical Software 3.3.0. The script can be found in Appendix 1[27–33].

Results

The IMPACT study included data from eight randomized controlled trials and three observational studies[22]. Data from 8,799 patients was used, which came from 265 different centers, which admitted between 1 and 453 patients, which were mostly (78%) male, and had a median age of 30 (interquartile range (IQR): 21 - 45) (Table 1).

The PRACTISE trial was a cluster randomized trial of studying the implementation of IV thrombolytic treatment in the Netherlands. It included observational data of 1,657 patients in 12 centers[23]. Hospitals had a minimum of 28 and maximum of 310 patients, who had a median age of 73 (IQR: 62 - 80).

Table 1. Baseline characteristics of patients enrolled in the IMPACT study.

		n= 8799
Age (median, IQR)		30 (21 – 45)
Sex		
	Male (N, %)	6836 (78%)
Pupillary reactivity		
	Reactive to light (N, %)	5721 (82%)
	Not reactive to light (N, %)	1273 (18%)
Motor score		
	Makes no movements (N, %)	1335 (16%)
	Extension to painful stimuli (N, %)	1082 (13%)
	Abnormal flexion to painful stimuli (N, %)	1119 (14%)
	Flexion / Withdrawal to painful stimuli (N, %)	2034 (25%)
	Localizes to painful stimuli (N, %)	2440 (30%)
	Obeys commands (N, %)	269 (3%)

Table 2. Baseline characteristics of patients enrolled in the PRACTICE trial.

		n= 1657
Age (median, IQR)		73 (62 – 80)
Sex		
	Male (N, %)	902 (54%)
Atrial fibrillation		
	Present (N, %)	296 (18%)
	Not present (N, %)	1361 (82%)
Diabetes mellitus		
	Present (N, %)	274 (17%)
	Not present (N, %)	1383 (84%)
History of ischemic stroke		
	Present (N, %)	331 (20%)
	Not present (N, %)	1326 (80%)
NIHSS*		8

*NIHSS indicates National Institutes of Health Stroke Scale; indicator of Stroke severity

In the IMPACT study 4,544 (52%) of the patients had a favorable outcome and 4,255 (48%) had an unfavorable outcome. Of these, 2,780 (32%) were in vegetative state or died (Figure 2). In the PRACTISE trial 933 (56%) of the patients had a favorable outcome and 724 (44%) had an unfavorable outcome. Of these, 351 (21%) were in severely disabled state or died.

Sensitivity

More patients per hospital increased the percentage of hospitals which are correctly found to be deviant from the mean (Figure 3). Further, use of ordinal outcomes instead of dichotomized for favorable versus unfavorable outcome allowed for less patients in the analysis without loss of power; the use of ordinal outcomes compared to dichotomized outcomes allowed for up to 13% less patients in the analysis without a loss of power in the IMPACT study (Figure 3a) and for up to 32% less patients in the PRACTISE trial (Figure 3b). For example, a mean of 73 patients per hospital was needed to detect the same percentage deviant hospitals when ordinal outcomes were used compared to on average 134 patients per hospital when the dichotomization favorable versus unfavorable was used in the PRACTISE trial. Moreover, dichotomization for mortality required even more patients in the analysis compared to dichotomization for favorable versus unfavorable outcome, in this example 200 patients per hospital. This meant that the required number of patients could be reduced by 63% for the PRACTISE trial and up to 37% for the IMPACT study. The variation across simulations was relatively small (Appendix 2).

Specificity

To determine specificity, the simulations were performed without simulating true center effects. For all analyses an increase in sensitivity was not associated with a decrease in specificity: the type I error did not differ between analytical approached and was in all cases below 1%.

Discussion

This study aimed to assess how much power could be gained by using ordinal analysis instead of dichotomous analysis to detect between center differences in outcome. Use of ordinal outcomes in both stroke and TBI hospital comparisons, increased statistical efficiency of the estimation of differences between centers. The increase in statistical power resulted in a substantial reduction in required sample size when using ordinal instead of dichotomous outcomes. This sensitivity increase came without loss of specificity.

Our results are in line with previous studies on estimating treatment effects in RCTs[34–36]. Previous studies on ordinal outcome analysis in trials, showed an increase in power, and higher potential of detecting treatment effects[15,19,20,37], with sample size reductions up to 40%. The current study shows the use of ordinal data is not only of added value in RCTs that assess treatment effects, but also in observational data to assess differences between centers in outcome. It illustrates to what extent sample size can be reduced without loss of power compared to the use of a dichotomous outcome. In the example databases on stroke and TBI, a reduction in sample size of 37% and 63% was achieved. The difference in power gain between the two examples could be partly explained by the fact that the mRS is used as a 6-point ordinal scale (originally 7) while the GOS is used as a 4-point ordinal scale (originally 5). An ordinal scale with a higher number of levels may contain more information, and may provide more discriminability. In addition, the efficiency gain of an ordinal outcome is optimal if the proportional odds assumption perfectly holds[38,39].

In our analysis we used odds based on the true data from the IMPACT study and PRACTICE trial, in which the proportional odds assumption is not perfectly met. It has however been shown that even if proportional odds assumptions are violated, analysis of the ordinal scale is still beneficial over dichotomization and results are robust regardless of the violation [20,40,41]. In the past the importance of assumption of proportionality might have been stressed too much. More important than the proportional odds assumption is the ordering of the adjacent outcomes. If there is agreement among stakeholders that each score on a certain scale is more favorable than a one point lower score, testing for proportional odds assumptions can be considered redundant[41]. If not, a potential solution is to combine adjacent categories of the scale that are not perceived ordinal, e.g. dead and vegetative state.

This study illustrates how much information is lost, not only by discarding the ordinal outcomes but when dichotomization leads to low event rates. This is the case when only mortality ratios are considered, and especially when mortality at a fixed time point is used. Compared to ordinal outcomes mortality as outcome requires much larger sample sizes, in order to find potential differences in quality of care.

Using ordinal outcomes, when available, instead of dichotomous outcomes to compare hospitals is therefore strongly recommended. For stroke and TBI this is easily done, as most centers will be familiar with the use of these scales in research projects and clinical practice. We do however recognize that several medical conditions or fields do not have a relevant ordinal outcome scales. Ideally relevant ordinal scales for important conditions should be developed or refurbished and implemented.

The benefits of the use of ordinal scales have also been shown to have their limits. The chance of misclassification, even by extensively trained medical staff, is higher with the use of ordinal scales. This phenomenon is represented as the inter-rater reliability[42]. Misclassification has been included in the simulation, if misclassification is however larger than expected it can possibly lead to an underestimation of the error rates and an overestimation of the statistical power of the ordinal analyses[42–46]. Furthermore, in our analysis we collapsed vegetative state and mortality for the GOS, and similarly we collapsed mRS 5 (severe disability) with mRS 6 (death) into one state. For the GOS it is more of a common practice since it is questionable whether vegetative state is a better outcome than mortality. This has also been done for mRS, although patients in mRS 5 are awake and aware, and on average this is clearly a preferred health status over death. Clinically this might be a debatable choice, it is however done on occasion and in our analysis it makes comparison to the GOS easier and yields a more conservative estimate of the gain in power[23,24].

Dichotomization was done on the collapsed scale which adds possible misclassification to the dichotomized outcome scale, while true mortality ratios would not have any misclassification. At the same time including vegetative state and severe disability cases increases incidence rates and therefore the power of the analysis.

In this paper we repeatedly refer to reduction in sample size using ordinal instead of dichotomous outcomes. However, since statistical power is a major challenge in hospital comparisons, we would like to stress that by this we point out efficiency gain by using ordinal outcome analysis. Most (studies on) hospital comparisons are underpowered, and thus we do not advise aiming for smaller sample sizes when using ordinal outcomes. In this paper only one aspect in performing hospital comparisons is addressed. In general, to be able to perform valid and efficient hospital comparisons one should focus on 1) using larger sample sizes[47], 2) use ordinal outcome analyses and 3) sufficient case-mix adjustment.

Strengths and limitations

The advantage of performing simulations on quality of care data is that we have a priori knowledge of which hospitals deviate from the mean. A limitation of basing the simulation on real datasets is that it limits the variety of situations which are simulated. Furthermore, the number of patients per hospital was constant in our study, instead of a mix of smaller and larger hospitals as one would see in reality.

Conclusion

Use of the ordinal outcomes instead of the binary outcomes for hospital comparisons, results in considerable efficiency gains. In quality of care research, where lack of power is a substantial problem, using ordinal clinical outcomes could be a way to increase possibilities to find outliers when comparing hospitals. In cases where an ordinal scale is available we strongly advise to exploit the ordinal scale and to not dichotomize in any way.

Abbreviations

TBI: traumatic brain injury

SMR: standardized mortality ratio

IMPACT: International Mission on Prognosis And Clinical Trial Design in Traumatic Brain Injury

PRACTISE: PRomoting ACute Thrombolysis in Ischemic StrokE

GCS: Glasgow Coma Scale

mRS: modified Rankin Scale

NIHSS: National Institute of Health Stroke Scale

CIs: confidence intervals

IQR: interquartile range

Declarations

Ethics approval and consent to participate

PRomoting ACute Thrombolysis in Ischemic StrokE (PRACTISE) : The medical ethics committees in each participating center assessed the study protocol. The protocol has been set up according to the revised Consolidated Standards of Reporting Trials (CONSORT) statement for cluster-randomized trials and has been published earlier

International Mission on Prognosis And Clinical Trial Design in Traumatic Brain Injury study: The study has been approved by the Ethical Committees of all participating centres and have been performed in accordance with the ethical standards laid down in the Declaration of Helsinki and its later amendments.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and/or analysed during the current study are not publicly available because participants gave no consent for data sharing.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was performed without funding.

Authors' contributions

HFL and IEC planned the study, developed the model and IEC and NvL wrote the paper, IEC built the final version of the model and analysed the data. DWJD provided data and insight into clinic processes. DWJD, HFL and EWS provided valuable input in the writing process and expert knowledge, EWS and HFL statistical and analytical, DWJD from a medical point of view. All authors revised and approved the final version of the manuscript.

Acknowledgements

We'd like to thank all contributors to the *International Mission on Prognosis And Clinical Trial Design in Traumatic Brain Injury* study and the *PRomoting ACute Thrombolysis in Ischemic StrokE trial*. Foremost: Andrew I.R. Maas, Anthony Marmarou, Gordon D. Murray, Sir Graham M. Teasdale and Maaike Dirks , Louis W. Niessen, Jeroen D.H. van Wijngaarden, Peter J. Koudstaal, Cees L. Franke, Robert J. van Oostenbrugge, Robbert Huijsman, Mirella M.N. Minkman

References

1 Hibbard JH. What can we say about the impact of public reporting? Inconsistent execution yields variable results. *Ann Intern Med* 2008;**148**:160–1. doi:10.7326/0003-4819-148-2-200801150-00011

2 Bilimoria KY, Cohen ME, Merkow RP, *et al.* Comparison of Outlier Identification Methods in Hospital Surgical Quality Improvement Programs. *J Gastrointest Surg* 2010;**14**:1600–7. doi:10.1007/s11605-010-1316-6

3 Krumholz HM, Lin Z, Normand S-LT. *Measuring hospital clinical outcomes*. 2013.

4 Lingsma HF, Steyerberg EW, Eijkemans MJC, *et al.* Comparing and ranking hospitals based on outcome: Results from The Netherlands Stroke Survey. *QJM* 2009;**103**:99–108. doi:10.1093/qjmed/hcp169

5 Lingsma HF, Roozenbeek B, Li B, *et al.* Large Between-Center Differences in Outcome After Moderate and Severe Traumatic Brain Injury in the International Mission on Prognosis and Clinical Trial Design in Traumatic Brain Injury (IMPACT) Study. *Neurosurgery* 2011;**68**:601–8. doi:10.1227/NEU.0b013e318209333b

6 Van Dishoeck AM, Koek MBG, Steyerberg EW, *et al.* Use of surgical-site infection rates to rank hospital performance across several types of surgery. *Br J Surg* 2013;**100**:628–37. doi:10.1002/bjs.9039

7 Seaton SE, Barker L, Lingsma HF, *et al.* What is the probability of detecting poorly performing hospitals using funnel plots? *BMJ Qual Saf* 2013;**22**:870–6. doi:10.1136/bmjqs-2012-001689

8 Jarman B, Pieter D, Van Der Veen AA, *et al.* The hospital standardised mortality ratio: A powerful tool for Dutch hospitals to assess their quality of care? *Qual Saf Heal Care* 2010;**19**:9–13. doi:10.1136/qshc.2009.032953

9 Jarman B, Gault S, Alves B, *et al.* Explaining differences in English hospital death rates using routinely collected data. *BMJ* 1999;**318**:1515–20. doi:10.1136/BMJ.318.7197.1515

10 Tu YK, Gilthorpe MS. Revisiting the relation between change and initial value: A review and evaluation. Stat. Med. 2007;**26**:443–57. doi:10.1002/sim.2538

11 Van Den Bosch WF, Kelder JC, Wagner C. Predicting hospital mortality among frequently readmitted patients: HSMR biased by readmission. *BMC Health Serv Res* 2011;**11**:57. doi:10.1186/1472-6963-11-57

12 van Gestel YRBM, Rutten HJT, de Hingh IHJT, *et al.* The standardised mortality ratio is unreliable for assessing quality of care in rectal cancer. *Neth J Med* 2013;**71**:209–14.

13 Van Gestel YRBM, Lemmens VEPP, Lingsma HF, *et al.* The hospital standardized mortality ratio fallacy: A narrative review. Med. Care. 2012;**50**:662–7. doi:10.1097/MLR.0b013e31824ebd9f

14 MacCallum RC, Zhang S, Preacher KJ, *et al.* On the practice of dichotomization of quatitative variables. Psychol. Methods. 2002;**7**:19–40. doi:10.1037/1082-989X.7.1.19

15 Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;**332**:1080. doi:10.1136/bmj.332.7549.1080

16 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat Med* 2006;**25**:127–41. doi:10.1002/sim.2331

17 Maas AI, Murray G, Henney H, *et al.* Efficacy and safety of dexanabinol in severe traumatic brain injury: results of a phase III randomised, placebo-controlled, clinical trial. *Lancet Neurol* 2006;**5**:38–45. doi:10.1016/S1474-4422(05)70253-2

18 Valenta Z, Pitha J, Poledne R. Proportional odds logistic regression - Effective means of dealing with limited uncertainty in dichotomizing clinical outcomes. *Stat Med* 2006;**25**:4227–34. doi:10.1002/sim.2678

19 Roozenbeek B, Lingsma HF, Perel P, *et al.* The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. *Crit Care* 2011;**15**:R127–R127. doi:10.1186/cc10240

20 McHugh GS, Butcher I, Steyerberg EW, *et al.* A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: Results from the IMPACT Project. *Clin Trials* 2010;**7**:44–57. doi:10.1177/1740774509356580

21 Saver JL. Novel end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. *Stroke* 2007;**38**:3055–62. doi:10.1161/STROKEAHA.107.488536

22 Marmarou A, Lu J, Butcher I, *et al.* IMPACT database of traumatic brain injury: Design and description. *J Neurotrauma* 2007;**24**:239–50. doi:10.1089/neu.2006.0036

23 Dirks M, Niessen LW, Van Wijngaarden JDH, *et al.* Promoting thrombolysis in acute ischemic stroke. *Stroke* 2011;**42**:1325–30. doi:10.1161/STROKEAHA.110.596940

24 Goyal M, Menon BK, Van Zwam WH, *et al.* Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. *Lancet* 2016;**387**:1723–31. doi:10.1016/S0140-6736(16)00163-X

25 Li B, Lingsma HF, Steyerberg EW, *et al.* Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. *BMC Med Res Methodol* 2011;**11**:77. doi:10.1186/1471-2288-11-77

26 Brant R. Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics* 1990;**46**:1171. doi:10.2307/2532457

27 RStudio Team. RStudio: Integrated Development Environment for R. 2015.

28 R Core Team. R: A Language and Environment for Statistical Computing. 2016.

29 Feng D. miscF: Miscellaneous Functions. 2016.

30 Yee TW. VGAM: Vector Generalized Linear and Additive Models. 2016.

31 Elff M. memisc: Tools for Management of Survey Data and the Presentation of Analysis Results. 2016.

32 Harrell Jr. FE. Hmisc: Harrell Miscellaneous. 2016.

33 Harrell, Jr. FE. rms: Regression Modeling Strategies. 2016.

34 Savitz SI, Benatar M, Saver JL, *et al.* Outcome analysis in clinical trial design for acute stroke: Physicians' attitudes and choices. *Cerebrovasc Dis* 2008;**26**:156–62. doi:10.1159/000139663

35 Saver JL, Gornbein J. Treatment effects for which shift or binary analyses are advantageous in acute stroke trials. *Neurology* 2009;**72**:1310–5. doi:10.1212/01.wnl.0000341308.73506.b7

36 Bath PMW, Lees KR, Schellinger PD, *et al.* Statistical analysis of the primary outcome in acute stroke trials. *Stroke* 2012;**43**:1171–8. doi:10.1161/STROKEAHA.111.641456

37 Machado SG, Murray GD, Teasdale GM. Evaluation of designs for clinical trials of neuroprotective agents in head injury. *J Neurotrauma* 1999;**16**:1131–8. doi:10.1089/neu.1999.16.1131

38 Bolland K, Sooriyarachchi MR, Whitehead J. Sample size review in a head injury trial with ordered categorical responses. In: *Statistics in Medicine*. 1998. 2835–47. doi:10.1002/(SICI)1097-0258(19981230)17:24<2835::AID-SIM933>3.0.CO;2-8

39 Bath PMW, Gray LJ, Collier T, *et al.* Can we improve the statistical analysis of stroke trials? Statistical reanalysis of functional outcomes in stroke trials. *Stroke* 2007;**38**:1911–5. doi:10.1161/STROKEAHA.106.474080

40 Maas AR, Steyerberg E, Marmarou A, *et al.* IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. *Neurotherapeutics* 2010;**7**:127–34. doi:10.1016/j.nurt.2009.10.020

41 Senn S, Julious S. Measurement in clinical trials: A neglected issue for statistician. *Stat Med* 2009;:3189–3209. doi:10.1002/sim.3603

42 Van Swieten JC, Koudstaal PJ, Visser MC, *et al.* Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 1988;**19**:604–7. doi:10.1161/01.STR.19.5.604

43 Wilson JTL, Hareendran A, Grant M, *et al.* Improving the assessment of outcomes in stroke: Use of a structured interview to assign grades on the modified Rankin Scale. *Stroke* 2002;**33**:2243–6. doi:10.1161/01.STR.0000027437.22450.BD

44 Quinn TJ, Dawson J, Walters MR, *et al.* Reliability of the modified rankin scale: A systematic review. *Stroke* 2009;**40**:3393–5. doi:10.1161/STROKEAHA.109.557256

45 Lu J, Murray GD, Steyerberg EW, *et al.* Effects of Glasgow Outcome Scale Misclassification on Traumatic Brain Injury Clinical Trials. *J Neurotrauma* 2008;**25**:641–51. doi:10.1089/neu.2007.0510

46 Choi SC, Clifton GL, Marmarou A, *et al.* Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. *J Neurotrauma* 2002;**19**:17–22. doi:10.1089/089771502753460204

47 Van Dishoeck AM, Lingsma HF, Mackenbach JP, *et al.* Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf* 2011;**20**:869–74. doi:10.1136/bmjqs.2010.048058

Appendix Figure Legend

Appendix Figure 1. Results of the simulation based on the IMPACT database (a) and results of the simulation based on the PRACTICE trial (b). The graph shows the variability in number of patients which need to be included per hospital in order to be able to find the number better or worse performing hospitals, set out for data which has been dichotomized, dichotomized for mortality/severe disability, and which was analyzed respectively on the full ordinal GOS scale (a), the modified Rankin scale(b).

Figures

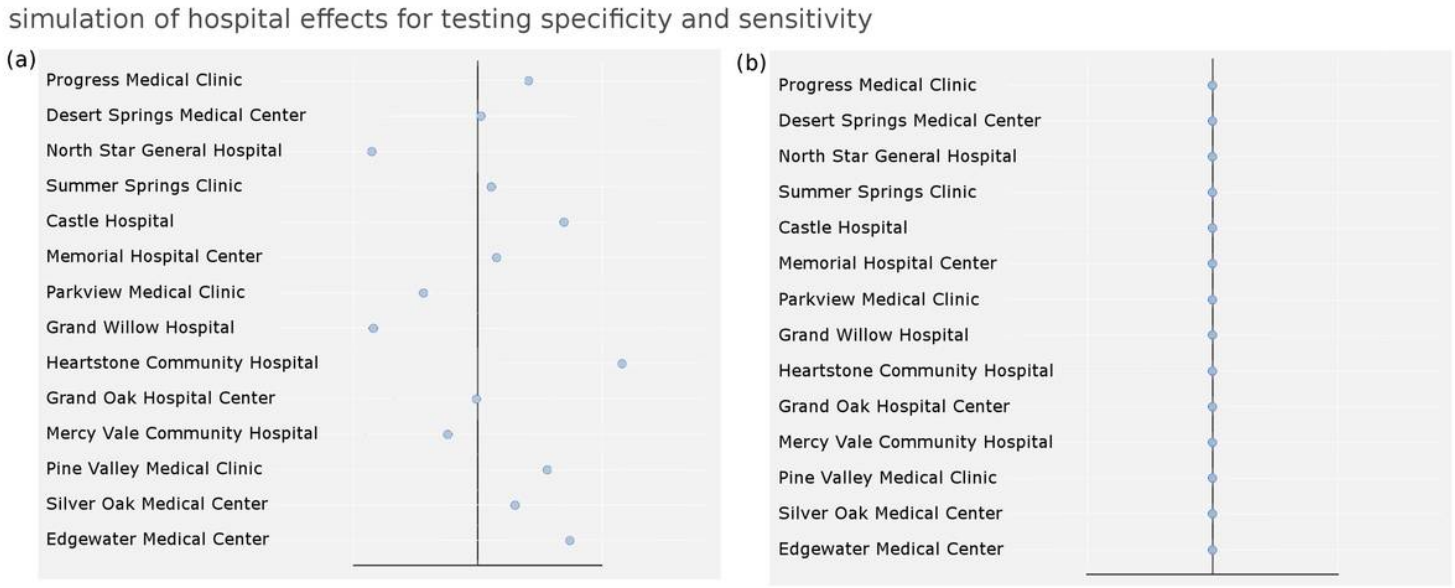

Figure 1

Illustration of the data generation process: (a) when a center effect is added, resulting in every hospital performing, to different degrees, better or worse than the mean (b) without a hospital effect added all hospitals perform the same.

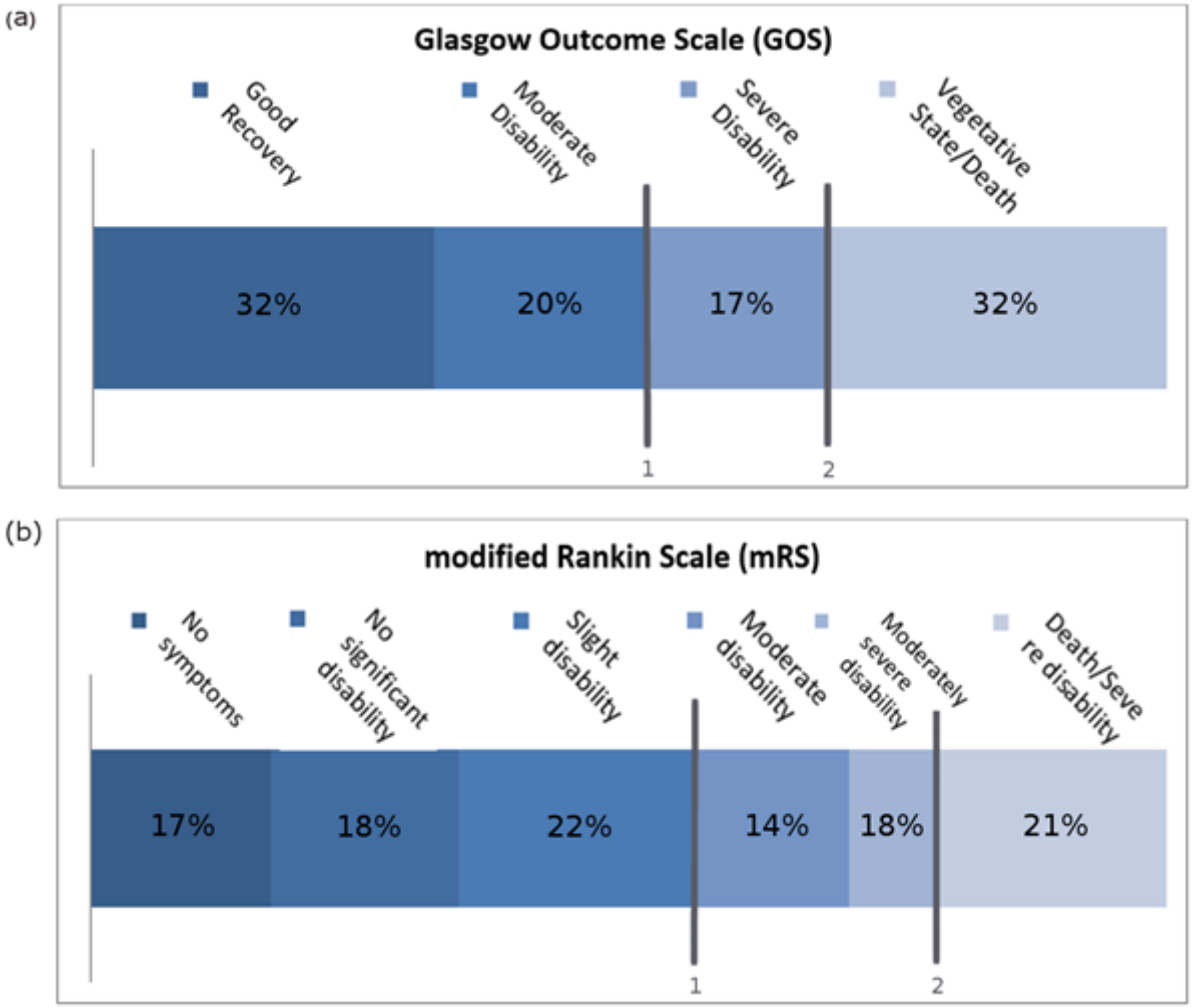

Figure 2

Distributions of the Glasgow Outcome Scale (a) and the modified Rankin scale (b), with the vertical line 1 illustrating the point of dichotomization at the clinically relevant outcome, and line 2 illustrating the point of dichotomization for mortality.

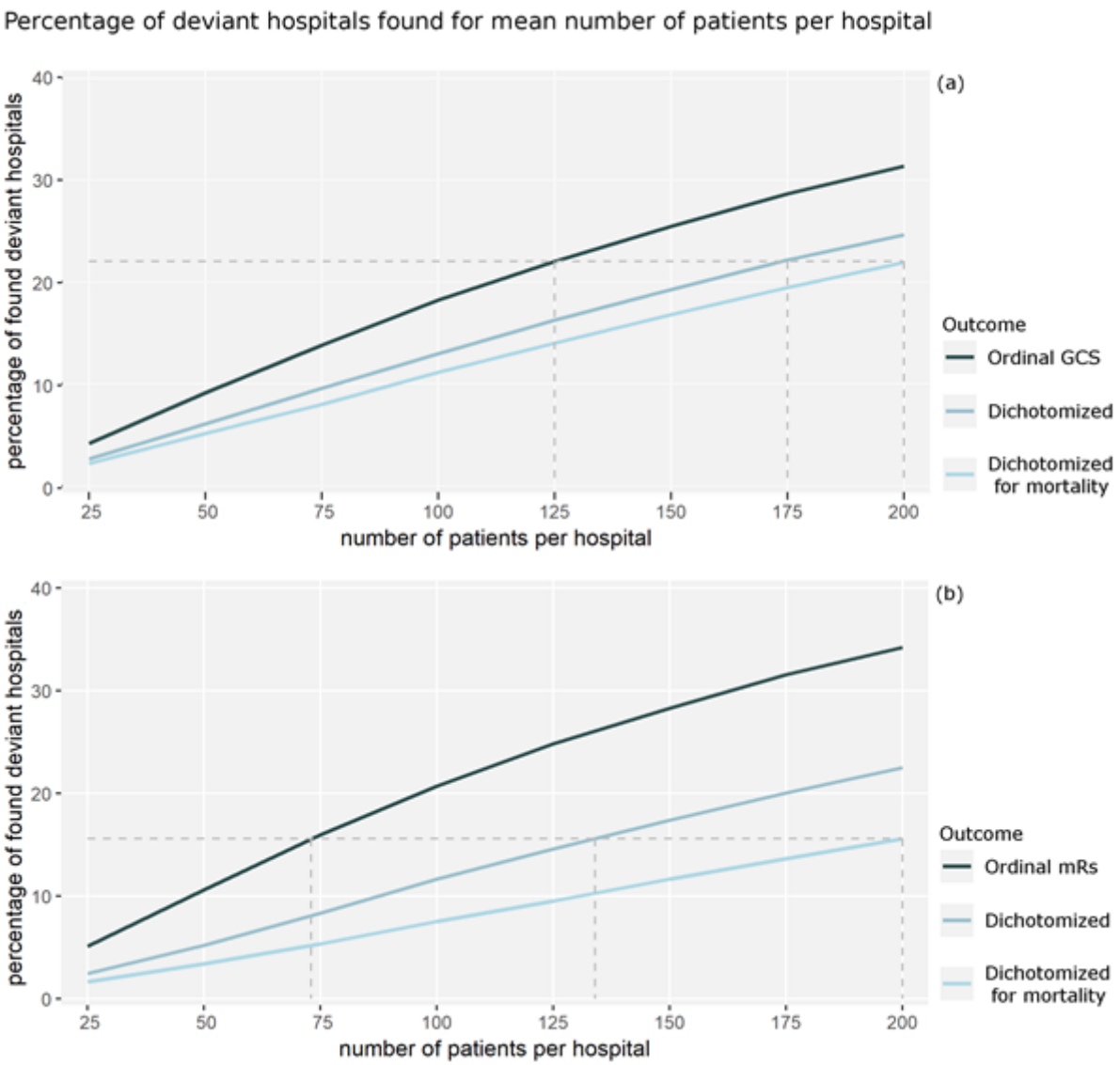

Figure 3

Results of the simulation based on the IMPACT database (a) and results of the simulation based on the PRACTICE trial (b). The graph shows mean number of patients which need to be included per hospital in order to be able to find the number better or worse performing hospitals, set out for data which has been dichotomized, dichotomized for mortality/severe disability, and which was analyzed respectively on the full ordinal GOS scale (a), the modified Rankin scale(b).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Rcode.pdf
- Appendix.pdf