Published in final edited form as:

Alzheimer Dis Assoc Disord. 2010 ; 24(3): 264–268. doi:10.1097/WAD.0b013e3181d489c6.

Sample size requirements for training to a kappa agreement criterion on Clinical Dementia Ratings

Rochelle E. Tractenberg, Ph. D., M. P. H.¹, Futoshi Yumoto², Shelia Jin, M.D., M.P.H.³, and John C. Morris, M. D.⁴

¹ Departments of Neurology, Biostatistics, Bioinformatics & Biomathematics, and Psychiatry, Georgetown University School of Medicine, Washington, D.C

² Department of Measurement and Statistics, University of Maryland, College Park MD

³ Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine University of California, San Diego

⁴ Department of Neurology, Washington University School of Medicine, St. Louis, MO

Abstract

The Clinical Dementia Rating (CDR) is a valid and reliable global measure of dementia severity. Diagnosis and transition across stages hinge on its consistent administration. Reports of CDR ratings reliability have been based on one or two test cases at each severity level; agreement (kappa) statistics based on so few rated cases have large error, and confidence intervals are incorrect. Simulations varied the number of test cases, and their distribution across CDR stage, to derive the sample size yielding a 95% confidence that estimated kappa is at least .60. We found that testing raters on five or more patients per CDR level (total N=25) will yield the desired confidence in estimated kappa, and if the test involves greater representation of CDR stages that are harder to evaluate, at least 42 ratings are needed. Testing newly-trained raters with at least five patients per CDR stage will provide valid estimation of rater consistency given the point estimate for kappa is roughly .80; fewer test cases increases the standard error and unequal distribution of test cases across CDR stages will lower kappa and increase error.

Keywords

Agreement; kappa; CDR; training

Introduction

The Clinical Dementia Rating (CDR; [1]) is a global measure of dementia representing five stages, ranging from “no impairment” (CDR=0) to “severe dementia” (CDR = 3) [2]. The CDR is frequently used as an entry criterion and/or a primary outcome measure in Alzheimer’s disease clinical trials [3], and it is often used as the basis for a consensus decision, i.e., CDR results are reviewed by a committee of practitioners and transition from one severity stage to another is determined based on consensus. In this context, it is critical that raters are consistent not only with other raters given the same patients and information, but also within themselves, consistently identifying the same dementia severity level across patients. More particularly, it

is essential for the validity of a study that the highest level of rater accuracy be achieved *prior* to the start of enrollment. This accuracy is a typical endpoint of training for personnel across multiple sites before a multi-site clinical study begins.

Reports of the validity and reliability of the CDR have been based on 3 – 80 raters who applied the CDR to between 3 and 15 cases each [4–6]. Whether the consistency of only a few raters (e.g., a clinical practice) or a large group of personnel (e.g., a multi-center clinical study) is the outcome of interest, estimates of agreement based on small test samples (<10) will have large associated standard errors (see e.g., [7]). Additionally, when estimating confidence intervals around the kappa estimate, large-sample methods assume no fewer than 20 [8–9] and preferably at least 25–50 rated cases [10]. Thus, it is important to test each rater on a larger sample set than has been reported to date. Simulations have shown that the empirically-derived lower 95% confidence interval bound for estimated kappa was found to be 0.6 when 20 ratings of dichotomous items yielded an estimated kappa of .80 [9]. This is not a function of the agreement statistic that is employed, but rather depends on how the confidence interval for the statistic is calculated. Cohen's kappa statistic has many detractors (e.g., see [11] pp. 35–37; [12], p. 31) but is widely used. It is interpreted to represent the level to which independent people will agree after taking into account the fact that they would agree by chance [13–14]. However, in any set of ratings, kappa can only be estimated; as such, the variability of this agreement characterization must be acknowledged.

Landis and Koch [13] provided guidance for classifying/interpreting kappa estimates ($\hat{\kappa}$) as reflective of “poor” ($\hat{\kappa} < 0.0$), “slight” ($0.0 \leq \hat{\kappa} \leq 0.2$), “fair” ($0.21 \leq \hat{\kappa} \leq 0.4$), “moderate” ($0.41 \leq \hat{\kappa} \leq 0.6$), “substantial” ($0.61 \leq \hat{\kappa} \leq 0.8$), and “almost perfect” ($0.81 \leq \hat{\kappa} \leq 1.0$) agreement among raters. The motivation for the present work is that, due to variability in estimation, it is the lowest probable value for this estimate of agreement (i.e., the lower bound of the confidence or credibility interval) that should be classified along this continuum, and not the estimate itself. For example, a one-sided 95% confidence interval constructed around any calculated value of $\hat{\kappa}$ should have a lower bound no smaller than 0.61, particularly when training multiple users of an instrument, such as the CDR, in a clinical study. With this criterion, the true level of agreement in the sample is likely to be captured in the interval .61–1.0. When too few ratings are performed, the confidence interval can be mis-estimated, resulting in a much larger interval than .61 – 1.0.

In the context of a multi-center clinical trial, the ‘test’ for rater certification might provide estimates of agreement of individuals with the ‘correct’ rating, and the precision of this estimate must be high. The ideal situation is both a high level of agreement and a high value representing the lower bound of the confidence interval around the estimate. Thus, training programs should strive to achieve a calculated kappa value in the “excellent range” (i.e., at least .80 [13]); we also suggest that the lower bound of the confidence interval should be no lower than “substantial” (i.e., at least .61 [13]). This study describes the estimation of optimal numbers of test case ratings, and their distribution across the possible CDR levels, for certifying this combination (point estimate and lower bound) criterion for concordance and consistency among multiple CDR raters. Simulations were carried out to establish the relative sizes of confidence intervals around kappa statistics (agreement with the gold standard rating) based on testing sample size, distribution of testing sample CDR stages, and previously reported levels of agreement.

Methods

Materials

Estimates of kappa for a global CDR rating have been at least .80 whether derived from physicians [4], nurses [5], clinical study personnel [15], or clinical study coordinators [6].

These kappas were estimated by constructing contingency tables with columns reflecting “true” dementia severity (global CDR) according to a gold standard rater, and rows representing the dementia severity ratings of the individuals being ‘tested’. Raters (individuals being tested) can rate one or more individuals in a column, but no ratee (individual being rated) falls into more than one column. The formula for Cohen’s kappa is: $[(\text{proportion observed agreement}) - (\text{proportion expected agreement})] / [1 - (\text{proportion expected agreement})]$ [9].

The simulations were based on such a table, created from ratings given by 82 individuals who had just been trained to use the CDR for a clinical trial. Not all individuals were tested on a case at every CDR severity level, or within each column, but this table, derived from published data [15], is the largest published sample whose CDR decisions were tested.

Importantly, the data in Table 1 represents a large sample of *raters*, not *ratees*, and asymptotic standard error (ASE) for kappa requires large samples of cases (things being rated) and is not dependent on how many raters were involved. Our simulations were set up to determine the estimated kappa and its standard error when increasingly large numbers of *ratees* were ‘evaluated’. Therefore, using response probabilities determined by the distributions of raters’ decisions (rows) relative to the ‘truth’ of the gold standard ratings (columns) (see Table 1), 10,000 5×5 tables were generated under each of two conditions: equal numbers of test cases per CDR level (column marginals equal) and unequal numbers of test cases (column marginals unequal), where CDR levels that have been reported to be more difficult have *greater* numbers of cases and *fewer* test cases are given at less difficult CDR levels.

Procedure

The matrix shown in Table 1 was used to generate the probabilities that a ‘rater’ in the simulation would give a rating of that CDR severity level given that the gold standard (column) value was true. Ten thousand kappa estimates were generated for several different test sizes (numbers of cases rated). In the first condition, the cases in the ‘test’ at each size were distributed evenly (‘unweighted’) over the CDR severity levels (totals of 5, 10, 20, 25, 30, 40, and 50 cases rated); so that for a test of size five, one case was rated per CDR level; for a test of size 10, two cases per level were rated, and 10 cases per level were rated for test size 50. In the weighted conditions, the test sizes were not evenly distributed over the CDR severity levels but instead, harder-to-evaluate severity levels were weighted, that is, twice as many test cases were rated from those levels than from the others (totals of 7, 14, 21, 35, 42, and 49 cases rated). In one of the weighted condition sets of 5×5 matrices, greater numbers of test cases were ‘rated’ for CDR = 0 and 0.5 (in Table 1 these are shown to be the hardest levels to rate, resulting in the highest errors; [15]). We also repeated the differential weighting simulation for CDR levels 0.5 and 1, which represents a testing situation where trainee, or training, attention would be focused at distinguishing cases or transitions from 0.5 to 1. It is important to note that these conditions do not represent weighted and unweighted kappa estimates but instead represent the distribution of test cases over the CDR levels in our simulated matrices. In all simulations, the simple unweighted kappa was calculated (simulation details in Appendix).

For each rated sample size, kappa statistics were computed reflecting agreement of a single rater with a second rater [14] where the second rater (columns) is the gold standard. Each kappa statistic was calculated based on 10,000 generated 5×5 matrices in order to generate an empirical distribution of estimates. Using the output for each matrix under each agreement estimate, we were able to determine the interval wherein 95% of the kappa estimates fell. This approach obviates problems with large-sample standard error estimates in the construction of confidence intervals, as well as providing empirically-derived interval boundaries [16], since we generated a 10,000-item distribution of kappa estimates and reported the median and 2.5th and 97.5th percentile values of these distributions (yielding intervals closer to credibility

than confidence (see [17])). Asymptotic standard errors (ASEs) were also computed for the estimated kappas, in each simulation under each condition, and the mean over 10,000 ASE values were combined with the mean of the kappa estimate sampling distribution ($\hat{\kappa} \pm 1.96 * ASE$) to obtain confidence intervals that correspond to those that might be obtained in any study using kappa. That is, we used a single ASE estimate with a single kappa estimate (each estimate, however, was derived from 10,000 replications, which is not typical). Simulations were run in Microsoft Excel using Visual Basic for Applications (Microsoft, Inc. 2007). Details of the simulation appear in the Appendix.

Results

The results from the simulations are shown in Table 2.

The kappa estimates were lower in the weighted conditions than in the unweighted condition (as expected given the sensitivity of kappa to marginal values, see [12]). The intervals for the estimated kappas in the unweighted condition were narrower than for those in the weighted conditions when fewer than 25 (unweighted) or 35 (weighted, 0.5 and 1) ratees were rated; beyond this point, the intervals are more similar although for the weighted (0 and 0.5) condition with 14 ratees, the interval was narrower for weighted than unweighted estimates. The differences and relations of kappa estimates and their associated intervals to number of ratees are shown in Figures 1A–1C.

Endpoints representing the 2.5th and 97.5th percentiles of the simulated distributions for the kappa estimates at each test size are shown in Figure 1 for the unweighted (panel A) and weighted (panel B: 0 and 0.5 are differentially weighted; panel C: 0.5 and 1 are differentially weighted) conditions. In the unweighted condition (panel A), the precision of the estimate reaches the criterion (above 0.60) when the total sample size surpassed 25. When more cases are rated at the CDR levels that are harder ([15] CDR=0 and CDR=0.5), the lower bound never passes 0.60; while when CDR 0.5 and 1 levels are differentially weighted (e.g., training and testing focus on distinguishing these two levels), more than 42 cases overall must be rated for the lower bound of the confidence interval to pass the 0.60 criterion. Thus, the sample size required to achieve some target level of the lower bound of a kappa estimate will vary depending on differential weighting of CDR levels in a testing situation; in an unweighted CDR distribution, the lower-bound criterion is met with five cases rated per CDR level.

Discussion

These simulations suggest that, in training raters to use the CDR, each rater should be tested on at least five different cases at each of the five levels of CDR severity. A typical goal in a training session is for all raters to achieve a kappa criterion of .80 (i.e., that the estimate of kappa will be at least 0.80 for the sample). Our simulations, which assumed a ‘true’ kappa of 0.77, suggested that in a ‘test’, raters must rate a total of 25 cases for the lower bound of the 95% CI of similar kappa estimates to be greater than 0.60. Thus, we recommend that training paradigms be based on attaining not only the target point estimate for kappa (excellent agreement, .80) but also ensuring the lower bound of the CI for their estimate be greater than the minimum acceptable kappa value (.60); to achieve this combination criterion, individuals should be required to rate no fewer than five cases with CDR severities at each of the five levels. This context is contrasted with inter-rater reliability studies which may focus on the instrument’s performance rather than the raters, or both (e.g., [18]).

We have shown empirically that, if a kappa estimate is based on fewer than 10 *ratees* (cases rated, 1–2 per CDR severity level), it is likely that the lower bound of the 95% CI for that kappa will fall in the .4–.6 (“moderate”) range, even if the kappa estimate itself is substantial or

excellent. The commonly reported situation is that one case per severity level is rated; this results in the least precision for the kappa estimate. Further, in cases where marginal distributions are not equal (i.e., differential weighting of some CDR levels) the CI based on asymptotic standard error (ASE) will give an incorrectly high value for the lower CI bound (suggesting greater precision). This shows the inappropriateness of using ASE with small numbers of rated cases; with more rated cases, the ASE-based interval endpoints are close to the empirical endpoints.

Although many researchers feel that kappa should not be used (see, e.g., [12]), kappa is a common summary of chance-adjusted agreement that can be useful as a criterion for training and testing provided that the lower bound of the estimate's confidence interval is sufficiently high, and is not obtained with an inappropriate application of the asymptotic standard error. We do not address the criticisms of kappa or the use of kappa statistics in inference drawing (e.g., [19]); our results are simply intended to highlight the potential for imprecision in studies where large sample-specific (asymptotic) standard errors for kappa are inappropriately utilized or where the "large sample" is incorrectly interpreted to reflect raters, rather than cases/items rated. In most cases, the standard errors used to generate confidence intervals for point estimates of kappa require far larger samples of items/cases rated than is feasible. Blackman & Koval [9] provide an excellent review of more appropriate methods for standard error estimation; our results agree with theirs in that small samples will yield low precision (although see [20]).

We have provided a minimum rated-case sample size for testing with respect to assessing training on the CDR – these sample sizes are specific to our simulated data (Table 1). Our results also show that weighting the testing so that harder-to-rate CDR levels are tested with more cases (unequal column marginals) will *decrease* kappa. This is not to suggest that the training and assessment of trained raters in their use of the CDR should not focus on that distribution of CDR levels which will be the most meaningful. On the contrary, our results serve to emphasize the importance of the assessment design in the evaluation of CDR training: it should be appropriate to the planned use of the CDR as well as the implications of the evaluation's summary (kappa, in our simulations).

Any plan for assessing CDR training should include a training-specific simulation to estimate the distribution of cases to be rated by all in the test. Our results concord with previous research that kappa confidence intervals are unacceptably liberal when too few cases are rated, and to determine specifically how many cases should be rated for some target kappa estimate to be obtained, a simulation should be carried out with a CDR distribution that matches the to-be-trained distribution of CDR levels. It should also be noted that, with sufficiently high numbers, the appropriately-computed confidence interval could fall close enough to the kappa estimate (e.g., 0.77) such that the upper bound falls below the target of 0.80.

A combination of point estimate and lower-bound value as the criterion for 'success' in a training situation can be applied to other instruments and contexts in order to improve efficiency and protect power where multiple raters are collaborating over time on inclusion and endpoint criteria. Our results support the use of sample sizes appropriate (i.e., large enough, and in the correct variable, the *ratees*, not *raters*) for a reliable estimate with a lower precision bound that is above the desired criterion. We also describe a method for estimating the lower precision bound that is appropriate to the distribution of items rated as well as the context of the summary (kappa here) and other concordance-oriented training assessment paradigms. These results have implications for the design of the training, as well as of the assessment of that training. Our approach is adaptable to the use of Cohen's kappa as an agreement criterion in other settings and instruments.

Acknowledgments

This work was supported by the ADCS NIA grant AG10483. Work by RET & FY was supported by K01AG027172 from the National Institute on Aging.

References

1. Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. *Br J Psychiatry* 1982;140:566–572. [PubMed: 7104545]
2. Morris JC. The Clinical Dementia Rating: Current version and scoring rules. *Neurology* 1993;43:2412–2414. [PubMed: 8232972]
3. Leber, PD. FDA draft guidelines November 1990. 1990. Guidelines for the clinical evaluation of antidementia drugs.
4. Burke WJ, Miller PJ, Rubin EH, et al. Reliability of the Washington University Clinical Dementia Rating. *Arch Neurol* 1988;45:31–32. [PubMed: 3337672]
5. McCulla MM, Coates M, Van Fleet N, Ducheck J, Grant E, Morris JC. Reliability of Clinical Nurse Specialists in the Staging of Dementia. *Arch Neurol* 1989;46:1210–1211. [PubMed: 2818255]
6. Schafer K, Tractenberg RE, Sano M, Mackell JA, Thomas RG, Thal LJ, Morris JC. Reliability of monitoring the Clinical Dementia Rating in multicenter clinical trials. *Alzheimers Disease and Associated Disorders* 2004;18(4):219–22.
7. Schouten HJ. Nominal scale agreement among observers. *Psychometrika* 1986;51:453–466.
8. Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine* 1998;17(10):1157–68. [PubMed: 9618775]
9. Blackman NJ, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine* 2000;19(5):723–74. [PubMed: 10700742]
10. Donner A, Eliasziw. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Stat Med* 1992;11(11):1511–9. [PubMed: 1410963]
11. Shoukri, MM. Measures of Interobserver Agreement. Boca Raton, FL: Chapman and Hall; 2003.
12. Von Eye, A.; Mun, EY. Analyzing Rater Agreement: Manifest Variable Methods. Lawrence Erlbaum Associates; 2005.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:671–679.
14. Fleiss, JL. Statistical Methods for Rates and Proportions. 2. New York: John Wiley and Sons; 1981.
15. Tractenberg RE, Schafer K, Morris JC. Interobserver disagreements on Clinical Dementia Rating assessment: Interpretation and implications for training. *Alzheimer Disease and Associated Disorders* 2001;15(3):155–161. [PubMed: 11522933]
16. Mooney, CZ.; Duval, RD. Bootstrapping: A nonparametric approach to statistical inference. Newbury Park, CA: Sage Publications; 1993.
17. Winkler, RL. An introduction to Bayesian inference and decision (2E). Gainesville FL: Probabilistic Publishing; 2003.
18. Saito Y, Sozu T, Hamada C, Yoshimura I. Effective number of subjects and number of raters for inter-rater reliability studies. *Stat Med* 2006;25(9):1547–60.S. [PubMed: 16143966]
19. Altaye M, Donner A, Klar N. Inference procedures for assessing interobserver agreement among multiple raters. *Biometrics* 2001;57(2):584–8. [PubMed: 11414588]
20. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med* 1998;17(1):101–10. [PubMed: 9463853]

Appendix: Details of simulation

Data generation

Using the agreement (and disagreement) rates and tendencies shown in Table 1, and aiming for a 'true' kappa of 0.77 (which was achieved in the 2001 study on which our simulation was

modeled), 10,000 replications were carried out using Excel (2007, Microsoft, Inc.) at each of the following number-of-ratings/design-of-testing-scenario combinations:

- Non-weighted simulation: 1 to 10 items (excluding 3, 7 and 9) rated at each of the five CDR levels (total ratings ranging from 5–50, omitting totals of 15, 35, and 45). This corresponds to a testing situation where all CDR levels are tested equally.
- Weighted simulations: 1 to 7 items rated (excluding 4) at each of the three CDR levels 0, 2 and 3, and twice as many items rated at CDR levels 0 and 0.5 (total ratings ranging from 7–49, omitting totals of 28) in one weighted condition; in the other weighted condition CDR levels 0.5 and 1 were differentially weighted. These correspond to testing situations where two CDR levels are tested more heavily (requiring 2–14 ratings of cases with CDRs of 0 and 0.5, which have been reported to be the most challenging; or of cases with CDRs of 0.5 and 1, which might be of greatest interest in a study tracking MCI conversions to AD), while the other three levels are tested less heavily (requiring 1–7 ratings at each CDR level).

For each replication, Cohen's kappa was computed and contributed to create a sampling distribution of 10,000 estimates of kappa corresponding to the number-of-ratings/design-of-testing-scenario combinations listed above. Since 50 cases in a test of newly trained CDR users seemed an upper limit to the number that was reasonable, we did not seek combinations beyond this level (49 in the weighted testing scenario).

Data Analysis

For each sampling distribution of kappa estimates, the following descriptive statistics were generated:

- Mean, median and standard deviation of sampling distribution of kappa –all shown in figures; medians only are reported since the sampling distribution is not symmetric.
- 95% CI based on empirical distribution percentile values (i.e. 2.5 and 97.5 percentiles from the 10,000 kappa estimates)
- 95% CI based on the asymptotic standard error.

Figure 1. Empirical and ASE confidence intervals/bands, for kappa estimates by condition

Notes: 2.5% and 97.5% represent the empirical percentile values from the 10,000 value kappa sampling distribution; ASE 2.5% and ASE 97.5% represent the percentile values obtained using the asymptotic standard error for the kappa estimates, given the condition and empirical kappa estimate. The median and mean kappa values are derived from the empirical distribution of 10,000 estimates under the given conditions and sample (# items rated) sizes. A reference line is included for kappa = 0.60, representing the target (criterion) lower bound of 'acceptable' kappa.

Table 1

The agreement between novice (n=16) and experienced (n=25) raters and a Gold Standard on one or two cases with Gold Standard CDR.

RATERS SAID:	GOLD STANDARD SAID:					
		0	0.5	1	2	3
	0	.732 *	.049			
	0.5	.268	.512	.04 [‡]		
	1		.439	.951	.025	
	2				.926	.033
3				.049	.967	

Notes. These data are adapted from Tractenberg et al. (2001), Table 2 (p. 158)

* of the 41 responses given by raters (16N + 25E) on one case with Gold Standard CDR of 0, 30 agreed with the Gold Standard and 11 did not.

[‡] only 25 ratings were considered in this cell (1/25) since one rating was missing.

Table 2

Results based on 10,000 simulation runs for two conditions, by test size (# rates total).

Unweighted condition (equal column marginals)					
# Rates	Median $\hat{\kappa}$	Empirical 2.5%	Empirical 97.5%	ASE 2.5%	ASE 97.5%
5	0.75	0.50	1.00	0.42	1.00
10	0.75	0.50	1.00	0.52	1.00
20	0.75	0.56	0.94	0.59	0.95
25	0.80	0.60	0.90	0.61	0.94
30	0.79	0.63	0.92	0.62	0.92
40	0.78	0.63	0.91	0.64	0.90
50	0.78	0.65	0.88	0.66	0.89

Weighted condition (unequal column marginals, differential weighting of 0 & 0.5)					
# Rates	Median $\hat{\kappa}$	Empirical 2.5%	Empirical 97.5%	ASE 2.5%	ASE 97.5%
W7	0.65	0.30	1.00	0.40	1.00
W14	0.73	0.46	0.91	0.48	0.92
W21	0.70	0.47	0.88	0.52	0.88
W35	0.71	0.54	0.85	0.56	0.84
W42	0.70	0.55	0.85	0.57	0.83
W49	0.69	0.56	0.82	0.58	0.82

Weighted condition (unequal column marginals, differential weighting of 0.5 & 1)					
# Rates	Median $\hat{\kappa}$	Empirical 2.5%	Empirical 97.5%	ASE 2.5%	ASE 97.5%
W7	0.81	0.43	1.00	0.42	1.00
W14	0.73	0.43	1.00	0.50	0.96
W21	0.75	0.51	0.94	0.54	0.92
W35	0.74	0.59	0.89	0.59	0.88
W42	0.72	0.59	0.88	0.60	0.87
W49	0.73	0.60	0.84	0.61	0.85

Note: “#rates” refers to the number of cases that are rated by each trainee (rate_r). $\hat{\kappa}$ is the estimate for kappa. “Empirical” refers to the percentile (2.5, 97.5) of the distribution of $\hat{\kappa}$ values (the sampling distribution of $\hat{\kappa}$) generated in the simulation. ASE refers to the large sample asymptotic standard error of the estimate — that is, the value that, with sufficient sample size (number of cases rated), maps onto the empirical

percentiles of the sampling distribution for $\hat{\kappa}$. # Rates column: “W7” refers to the situation where a total of 7 ‘cases’ were distributed in the differential (weighted) way described in the text and table titles, either more cases at 0 and 0.5 or more at 0.5 and 1).