Revised 5/10/18

Inter and Intrarater Reliability of a Grading System for Congenital Diaphragmatic Hernia Defect Size

Reliability of CDH Grading System

Hunter, Chelsea E[a]; Saenz, Zoe M[a]; Nunez, Daisy[a]; Timsina, Lava[b]; Gray, Brian W[a]

a. Indiana University School of Medicine
Division of Pediatric Surgery, Department of Surgery
RI 2500, 705 Riley Hospital Drive
Indianapolis, IN 46202
USA
hunterce@iupui.edu, zoesaenz@iupui.edu, dnunez1989@gmail.com,
graybw@iupui.edu

b. Indiana University School of Medicine
Center for Outcomes Research in Surgery, Department of Surgery
545 Barnhill Drive, Emerson Hall
Indianapolis, IN
USA
ltimsina@iu.edu

Corresponding author

Brian W. Gray, MD
Assistant Professor of Pediatric Surgery
Riley Hospital for Children
Indiana University School of Medicine
705 Riley Hospital Drive, RI2500
Indianapolis, IN 46202
+1 (317)-439-1949
graybw@iupui.edu

Author contributions

CH: study design, data collection, data analysis, and manuscript drafting
ZS, DN: clinical data collection
LT: data analysis and manuscript drafting
BG: study design, data collection, data analysis, and manuscript editing

Disclosure

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest: none

This is the author's manuscript of the article published in final edited form as:
Hunter, C. E., Saenz, Z. M., Nunez, D., Timsina, L., & Gray, B. W. (2019). Inter- and Intra-rater Reliability of A Grading System for Congenital Diaphragmatic Hernia Defect Size. Journal of Surgical Research, 233, 82–87. https://doi.org/10.1016/j.jss.2018.07.045

Abstract

Background:
The Congenital Diaphragmatic Hernia Study Group (CDHSG) registry is a multi-institutional tool to track outcomes of CDH patients. The CDHSG asks surgeons to categorize diaphragmatic defect sizes as type A-D based on published guidelines. The reported size of the defect has been correlated with patient outcomes, but the reliability of this system has never been studied. Our goal was to evaluate the inter-rater and intra-rater reliability of the CDHSG grading system.

Materials and Methods:
Forty-six operative notes from CDH patients that underwent surgical repair at a single institution were collected and cropped to include only the information necessary to grade the hernia defect based on the CDHSG guidelines. The defects were graded by 9 pediatric surgeons on two separate occasions (18 weeks apart). Inter-rater reliability was calculated using a Cohen's kappa (κ). Intra-rater reliability was calculated using an intraclass correlation coefficient (ICC).

Results:
Inter-rater reliability was minimal to weak (κ_{round1} = 0.395, κ_{round2} = 0.424). Agreement ranged from 19.57% (κ= -0.0745) to 82.61% (κ= 0.7543). Inter-rater agreement was similar despite operative findings and outcomes: survival yes/no (κ=0.3690, κ=0.3518), need for ECMO yes/no (κ=0.3323, κ=0.3362), patch repair yes/no (κ=0.2050, κ=0.1916), and liver up/down (κ=0.2941, κ=0.4404).

Intra-rater reliability was good to excellent (ICC = 0.88, 95% CI [0.83-0.92]). Agreement with oneself ranged from 71.74% to 93.48%.

Conclusion:
The demonstrated weak inter-rater reliability of the current CDHSG grading system shows the need for improvement in how the grading system is applied by surgeons when reporting CDH defect size.

Keywords:
Congenital Diaphragmatic Hernia; CDH; grading system; reliability

Introduction:

Congenital diaphragmatic hernia (CDH) is a condition that is associated with an approximate 80% overall survival, and size of the defect has been found to be one of many important prognostic indicators.[1] Due to the low incidence and wide-ranging severity of this disease, a unified database is vital to study overall disease burden, outcomes, and new treatment modalities. The Congenital Diaphragmatic Hernia Study Group (CDHSG) registry is a multi-institutional tool that collects data on CDH patients from participating centers worldwide.

After showing that patients with larger defects had worse survival in a 2007 report from the registry, the CDHSG updated the data collection form and started gathering more specific information regarding hernia defect size.[2] A grading system was created that divides the size of the hernia defect into four types (A, B, C, and D) based on intraoperative findings. When institutions enter their CDH patients into the registry, hernia defect size is reported for those patients who underwent surgical repair. This is accomplished by interpreting the information that the surgeon provides in the operative note, which can include naming an exact defect type (A-D), reporting the diaphragmatic defect as a percentage, or purely qualitative descriptions.

Since this data has been added to the registry, studies have shown that defect size correlates with outcomes, such as morbidity, mortality, and hospital length of stay.[3,4] The ability to conduct meaningful clinical research regarding the implications of the size of a CDH depends heavily on the reliability of the registry data, particularly the reported grade of the hernia defect. However, the reliability of this grading system has not been evaluated. The goal of this study was to determine the inter-rater and intra-rater reliability of the current CDHSG grading system.

Materials and Methods:

Study Design

This study was approved by the Indiana University Institutional Review Board (#1705606018). Operative notes were collected from all patients that underwent surgical repair of a unilateral CDH at Riley Hospital for Children at Indiana University Health between 2010 and 2016. These 46 operative notes were cropped to include only the information necessary to grade the hernia defect based on the CDHSG grading system guidelines. Descriptions of the size and location of the defect, description of the repair, and size of the patch, when applicable, was retained. Any information that may bias the surgeon when grading the defect size (such as ECMO status and viscera found in the chest) was removed from the operative notes. Each note was cut and copied to fit on a single page and then placed in a random order. In addition to the cropped operative notes, the surgeons were provided with the standard CDHSG diagram that depicts the four grades of diaphragmatic hernias. The diagram included a description of each defect type, using percentages (Table 1). The surgeons were provided with no formal training in the use or application of the grading system. Each hernia was initially graded by 9 pediatric surgeons. The exercise was then repeated 18 weeks later with the order of the operative notes rearranged. The packets were otherwise not changed in any way. All surgeons were part of the same group, but they have varied training pedigrees. Each surgeon reported their number of years in practice and estimated the number of CDH repairs that they have performed. Additionally, intraoperative findings (liver up vs. down, ECMO status, repair type) and patient outcomes (length of stay, survival to discharge or transfer) were recorded.

Statistical Analysis

Inter-rater reliability between multiple raters was assessed using Cohen's kappa generalized for more than two rating outcomes (A, B, C, and D) by nine raters.[5,6] Kappa analysis for inter-rater reliability adjusts for the percentage of concordant and discordant agreements that arise from chance alone between the raters. We used –kap– in Stata/SE 14.2[7] to compute unweighted Kappa's statistics[8].

Unweighted kappa scores were used to more accurately evaluate the true reliability of the grading system. For use in the clinical setting, a stricter interpretation of kappa has been proposed and recently applied in clinical research.[9,10] The following cutoffs were used to interpret ***κ: ≤ 0.20 - no agreement, 0.21-0.39 - minimal agreement, 0.40-0.59 - weak agreement, 0.60-0.79 - moderate agreement, 0.80-0.90 - strong agreement, > 0.90 – almost perfect.***

We also analyzed intraclass correlation coefficient (ICC) using a multi-level model with random effects[11,12]. Because two ratings were taken for each patient and there exists multiple ratings per patient, we fit a three-level mixed-effect model with the following random effects: a random intercept and random slope on follow-up time at the patient level, and a random intercept at the rater's level. This model assumed that the residuals were independent with constant variance (homoskedastic). We then estimated ICC for this three-level nested model: ICC for the patient, the correlation between CDH ratings in the same patient, and ICC for the raters within a patient, the correlation between CDH ratings in the same rater and patient. We also reported 95% CI and p-value at 0.05 level of significance. The following cutoffs were used to interpret ***ICC[11]: <0.50 – poor reliability, 0.50-0.75 – moderate reliability, 0.75-0.90 – good reliability, >0.90 – excellent reliability.***

Results:

Demographics

Patient characteristics and outcomes of the 46 CDH patients in our study group is displayed in Table 2.

Inter-rater reliability

The pairwise Cohen's kappa from the first round of grading the hernia defects ranged from -0.027 to 0.7543, with an average agreement of 0.395 ($p<0.001$). When repeated eighteen weeks later, the pairwise Cohen's kappa ranged from -0.0745 to 0.7143, with an average of 0.424 ($p<0.001$). Figure 1 summarizes the pairwise kappa scores for each of the possible surgeon pairs.

From the first round of grading the 46 operative notes, there was 57.49% agreement across all raters, corresponding to a minimal level of agreement (κ=0.395, $p<0.001$). Between any two raters, agreement ranged from no agreement (21.74% agreement, κ= -0.027) to moderate agreement (82.61% agreement, κ= 0.7543). All 9 of the surgeons agreed on 2 of the 46 patients, both of which were assigned an "A" grade. Four patients received 3 different grades: 3 received grades A, B, and C, and 1 received grades B, C, and D. No patients were given all four grades.

The second review revealed similar inter-rater data. There was 59.6% agreement across all raters, corresponding to a weak level of agreement (κ = 0.424, $p <0.001$). Agreement between the pairs of raters ranged from 19.57% (κ= -0.0745, no agreement) to 80.43% (κ=0.7143, moderate agreement). All 9 surgeons agreed on the grade of two of the patients, both of which were categorized as an "A". Seven patients received 3 different grades: 5 received grades A, B, and C and 2 patients received grades B, C, and D. Again, no patients were given all four grades. There was no significant difference in overall surgeon agreement or κ between the first and second rounds.

When the two rounds of gradings are combined (giving each patient a total of 18 grades), all 9 surgeons agreed on the grade of one patient. Overall, nine patients received 3 different grades: 6 received grades A, B, and C and 3 received the grades B, C, and D. With all the grades combined, no patient received all four of the grades.

Inter-rater agreement was similar despite different operative findings and patient outcomes ($p > .05$): survival yes/no (κ=0.3690, κ=0.3518), need for ECMO yes/no (κ=0.3323, κ=0.3362), patch repair yes/no (κ=0.2050, κ=0.1916), and liver up/down (κ=0.2941, κ=0.4404). Only the first round of grading received this analysis since the two rounds revealed similar inter-rater data.

When the grades from the first round were grouped into small (A+B) versus large (C+D) defects, inter-rater reliability was not significantly increased ($p=0.0944$). The kappa for the small defects was .4707 and for the large defects it was 0.14.

Intra-rater reliability

Agreement with oneself ranged from 71.74% to 93.48%, with an average of 78.50%. The ICC was 0.88 (95% CI [0.83-0.92]), corresponding with good to excellent intra-rater reliability. No surgeon changed the grade of a particular defect by more than one severity (Fig. 2). For example, if a surgeon graded a defect as an "A" in the first round, it was never classified by the same surgeon as a "C" in the second round.

Surgeon Experience

Surgeon participation at our institution was 100% for both rounds of grading. The surgeons had a wide range of experience. Their years in practice (including fellowship) ranged from 2 to 31 years, with an estimated number of CDH repairs ranging from 8 to 115 (Table 3). There was no

significant correlation between intra-rater reliability and surgeon experience. The correlation coefficient between years in practice and intra-rater reliability was 0.2123 (p=0.5833). With a correlation coefficient of 0.1927, there was also no significant correlation between the number of CDH repairs performed and intra-rater reliability (p=0.6193). Four of the surgeons had participated in 20 or less CDH repairs (regardless of years in practice). They had an average intra-rater agreement of 76.09%. Of those surgeons with 70 or more repairs, agreement was slightly better at 80.43%, but not significantly so (p=0.41).

Discussion:

CDH is one of the most difficult and frustrating diagnoses that Pediatric Surgeons and Intensivists deal with on a daily basis. Our ability to objectively study patients with CDH is paramount when looking at retrospective data and in the prospective design of studies. We can use this data to improve risk stratification, family counseling, and care strategies. The data from the CDHSG registry has been used in multiple studies to evaluate the relationship between hernia defect size and outcomes. However, to our knowledge, this is the first study to examine the reliability of the grading system that is used by the CDHSG for reporting hernia defect size to the registry.

When determining defect size for the CDHSG registry, the reviewer may be biased because they know of the patient's treatment course and outcome. For example, knowing that a patient did not survive may impact the reviewer's assessment of the operative note and result in a more severe grading of the hernia defect. In the absence of patient characteristics that may bias a surgeon when grading the size of a hernia defect, this study demonstrates that agreement among surgeons on defect size is weak. The intra-rater reliability of the grading system, however, is good to excellent.

The variability seen in the grading of some defects was striking, with four patients receiving three different grades in the first round and seven receiving three grades in the second round. An "A" defect is clearly different than a "D" defect, but we found a fairly high amount of ambiguity in the middle. It is concerning that the same patient could receive both an "A" and a "C" grade or both a "B" and a "D" grade. We were unable to determine factors that increased the reliability of the grading system. As expected, agreement did not vary between the patient characteristics to which the surgeons were blinded (survival, ECMO status, liver position). However, the type of repair performed (primary vs. patch) also had no effect on overall agreement, even though this information was present in each of the operative notes.

None of our surgeons reported an actual grade in the operative note. One of the operative notes used the term "agenesis". Two surgeons still graded this hernia as a "C". Five of the operative notes reported the defect size as a percentage. Surprisingly, 3 of these patients received 3 different grades (all were A, B, and C). None of the operative notes that used percentages had perfect agreement among the 9 surgeons. One explanation for this may reveal a potential flaw of the grading system. The grading system classifies defects based on the percentage of hemi-diaphragm that is present and on the percent of the circumference of chest

wall that is involved. However, it is not clear what grade to select when these two factors fall under two different grades. For example, one operative note states that "[the hernia defect] encompassed about 30% of the area of the diaphragm", which is a type "B" defect. The note then goes on to state that "there was a good rim circumferentially", which falls under a type "A" defect. This particular note was graded as an "A" 22% of the time and as a "B" 73% of the time.

Intra-rater reliability among the surgeons was similar despite their various levels of experience when looking at either years in practice or number of CDH repairs performed. When we compared those surgeons with less experience (20 or fewer repairs) and those with more (70 and greater repairs), we still found no significant difference in intra-rater agreement, even though the more experienced surgeons did trend towards better self-agreement.

Not surprisingly, the intra-rater reliability was higher than the inter-rater reliability for this grading system. This may reflect differences in how individual surgeons interpret the information in the operative notes, which may be exacerbated when the operative note is written by someone else. The improved intra-rater reliability is likely due to generally fastidious surgeons following a consistent system to determine the defect grade. This system was likely based off their own interpretation of the CDHSG grading system. The issue is that different surgeons from the group in this study obviously have different interpretations of the grading system outlined by the CDHSG.

Extrapolation of our findings from this admittedly small sample to the reliability of the reported hernia grade to the CDHSG registry is difficult for a number of reasons. First, the operative notes used in this study were obtained from patients that were repaired at our own institution. It is not known how the quality of these notes compares to that of the operative notes at other institutions. Second, we used pediatric surgeons to grade the defects. The person determining the defect type at other participating institutions may not be someone with knowledge of diaphragmatic anatomy or with clinical experience in the operating room in general. Third, the process of reporting hernia grade to the registry at other participating institutions may vary drastically from our own. Even if one surgeon or research assistant at each CDHSG institution is reporting defect grade, there is likely to be a wide variety in how those people grade defects across the board. Thus, our data may reflect the true variability of the reported defect size to the registry, since compliance with the guidelines is unlikely to be perfect among the multiple institutions. And lastly, some institutions may instruct their surgeons to grade the defect intraoperatively and report this grade in their operative note. Our study only evaluates the reliability of applying the grading system to operative notes retrospectively. A different study would need to be done in order to determine the reliability of the grading system if it is used intraoperatively by the operating surgeon. A study conducted in this fashion would likely show an increased inter-rater and intra-rater reliability.

We applaud the efforts of the CDHSG to standardize how we report CDH defect size. However, the information provided in the operative notes we examined was highly variable, even from a single institution. Therefore, determining the hernia grade from this information alone was difficult to do with much accuracy or precision amongst reviewers. Going forward, institutions should educate their surgeons on the CDHSG guidelines and emphasize the importance of the

contents of their operative notes. Surgeons should be trained to report the hernia defect size in their operative notes using percentages instead of using descriptor terms that may be ambiguous or difficult to interpret. This could potentially increase the reliability with which defect size is reported to the registry. Alternatively, efforts towards making changes to the current grading system to limit ambiguity could be considered. Using a measured defect to diaphragmatic ratio, as suggested by Rygl, et al, could be considered[13]. Additionally, a completely objective system using intra-operatively obtained photographic imaging of the diaphragm could be developed. This would facilitate exact measurement of defect size by objective viewers. However attractive this option is, it would likely be prohibitively difficult to institute and standardize at all sixty plus CDHSG centers

Conclusions:

The grading system used by the CDHSG registry produces good to excellent reliability within a single rater but there is minimal to weak agreement between different raters. In order to study the relationship between defect size and outcomes, a reliable grading system is necessary. This can be accomplished through changing how defect sizes are reported in operative notes or by improving how the current guidelines are applied.

Acknowledgements

The authors want to thank the pediatric surgeons at Riley Hospital for Children at Indiana University Health for their time and participation in this study. We would also like to thank Kayli Mellencamp for proof reading the manuscript.

Disclosure

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest: none

Tables

Table 1. Description of Hernia Grades

Table 2. Patient characteristics and outcomes

ECMO- Extracorporeal membrane oxygenation, SD- Standard deviation

Table 3. Surgeon experience and intra-rater agreement

Years of practice and self-reported estimation of the number of CDH repairs performed.
Percent agreement represents the percent each surgeon (1-9) agreed with themselves between the two rounds of grading the CDH defects.

Figures

Figure 1. Pairwise Cohen's kappa

Pairwise Cohen's kappa for each possible pair of surgeons (dots) and the mean (black line) from the first and second rounds of grading the hernia defects. The interpretation of the kappa scores is indicated to the right.

Figure 2. Correlation of ratings for each surgeon at different times

Each plot represents the intra-rater reliability for each surgeon. Perfect agreement with oneself would create dots along the diagonal. Increasing size of the dot correlates with increasing agreement for each hernia grade (A-D).

References

1. Barriere F, et al. One-year outcome for congenital diaphragmatic hernia: results from the French national register. *J Pediatr.* 2018;193:204-10.
2. The Congenital Diaphragmatic Hernia Study Group, Lally KP, Lally PA, et al. Defect Size Determines Survival in Infants with Congenital Diaphragmatic Hernia. *Pediatrics.* 2007;120(3):e651-7.
3. The Congenital Diaphragmatic Hernia Study Group. Congenital diaphragmatic hernia: Defect size correlates with developmental defect. *J Pediatric Surg.* 2013;48:1177-82.
4. Putnam LR, Harting MT, Tsao K, et al. Congenital Diaphragmatic Hernia Defect Size and Infant Morbidity at Discharge. *Pediatrics.* 2016;138(5):e20162043
5. Fleiss JL. Measuring Nominal Scale Agreement among Many Raters. *Psychol Bull.* 1971;76:378-82.
6. Fleiss JL, B. Levin, and M. C. Paik. 2003. Statistical Methods for Rates and Proportions. 3rd ed. New York: Wiley
7. StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP
8. https://www.stata.com/manuals13/rkappa.pdf
9. McHugh, ML. Interrater reliability: the kappa statistic. *Biochem Med.* 2012;22(3):276-82
10. Spring AM, Pittman DJ, Aghakhani Y, et al. Interrater reliability of visually evaluated high frequency oscillations. *Clinical Neurophysiology.* 2017;128:433-41.
11. Koo TK, Li MY. A guideline of selecting and reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med.* 2016;15(2):155-63
12. https://www.stata.com/manuals13/me.pdf
13. Rygl M, Kuklova P, Zemkova D, Slaby K, Pycha K, Stranak Z, Melichar J, Snajdauf J. Defect-diaphragmatic ratio: a new parameter for assessment of defect size in neonate with congenital diaphragmatic hernia. *Pediatr Surg Int.* 2012;28(10):971-6.

Table 1

Hernia Grade	% of hemi-diaphragm present	% of chest wall involvement
A	>90%	<10%
B	50-75%	<50%
C	<50%	>50%
D	<10%	>90%

Table 2

Characteristic	Number (%)
Repair Type	
Primary	25 (54.3)
Patch	21 (45.7)
Liver Position	
Abdomen	26 (56.5)
Chest	20 (43.5)
ECMO Requirement	13 (28.3)
Repaired on ECMO	12 (26.1)
Length of Stay (days)	
Mean (SD)	69.4 (94.5)
Median	24.5
Survival	40 (87)

Table 3

	1	2	3	4	5	6	7	8	9
Years in Practice	31	2	22	17	6	4	14	30	31
CDH Repairs	115	8	100	15	20	8	70	80	70
Agreement (%)	80.43	71.74	71.74	71.74	78.26	82.61	82.61	93.48	73.91

Figure 1
Click here to download high resolution image

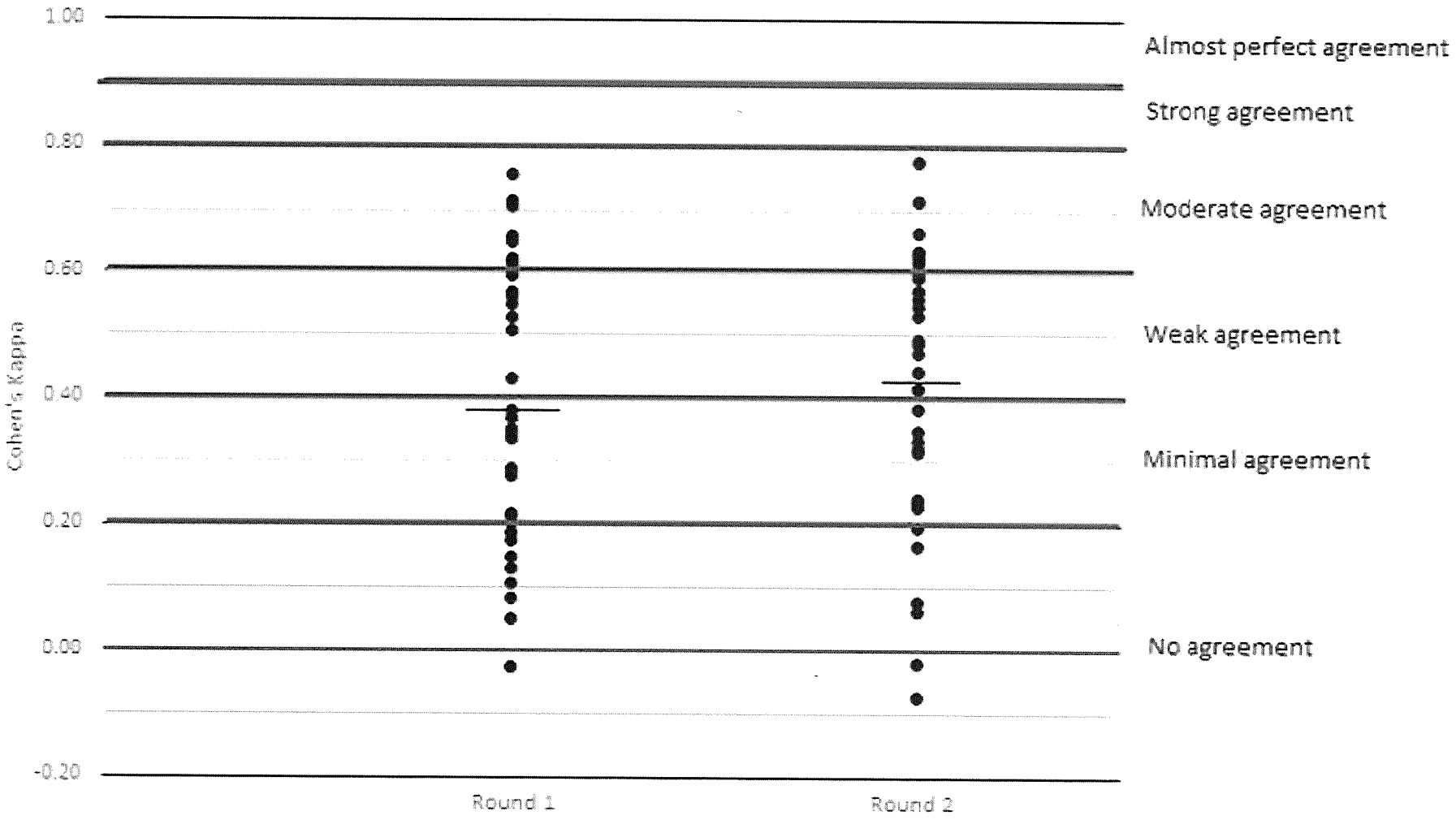

Figure 2
Click here to download high resolution image