Intra- and interrater reliability of the Chicago Classification of achalasia subtypes in pediatric High Resolution Esophageal Manometry (HRM) recordings

Singendonk MMJ, MD[1]; Rosen R, MD, MPH[2]; Oors J[1]; Rommel N, PhD[3,4]; van Wijk MP, MD, PhD[5,1]; Benninga MA, MD, PhD[1]; Nurko S, MD, MPH[2]; Omari TI, PhD[4,6,7].

1. Pediatric Gastroenterology and Nutrition, Emma Children's Hospital/AMC, Amsterdam, The Netherlands. 2. Division of Gastroenterology, Center for Motility and Functional Gastrointestinal Disorders, Boston, MA, United States. 3. Translational Research Center for Gastrointestinal Diseases, University of Leuven, Leuven, Belgium. 4. Department of Neurosciences, ExpORL, University of Leuven, Leuven, Belgium. 5. Department of Pediatric Gastroenterology, VU University Medical Center, Amsterdam, The Netherlands. 6. Gastroenterology Unit, Women's and Children's Health Network, Adelaide, SA, Australia. 7. School of Medicine, Flinders University, Bedford Park, SA, Australia.

Address of correspondence and reprint requests to: M.M.J. Singendonk, MD, Emma Children's Hospital AMC, C2-312, PO Box 22700, 1100 DD Amsterdam, The Netherlands. Telephone: +31 20 5662906; e-mail address: m.m.j.singendonk@amc.uva.nl

Short title: Reliability of achalasia subtypes in high-resolution manometry

Key points:

- Subtyping achalasia by high resolution manometry (HRM) is clinically relevant as response to therapy and prognosis have shown to vary accordingly. The reliability of diagnosing achalasia and its subtypes in children remains undefined.
- In this study, we found very good to excellent intra- and interrater reliability for diagnosing achalasia by HRM and the Chicago Classification (CC) when results of automated analysis software were interpreted by experienced observers.
- However, more variability was seen when relying on the software-driven diagnosis and for subtyping achalasia, indicating a need of improved HRM criteria for achalasia subtyping in children.

Abstract

Background: Subtyping achalasia by high resolution manometry (HRM) is clinically relevant as response to therapy and prognosis have shown to vary accordingly. The aim of this study was to assess inter- and intrarater reliability of diagnosing achalasia and achalasia subtyping in children using the Chicago Classification (CC) V3.0.

Methods: Six observers analyzed 40 pediatric HRM recordings (22 achalasia and 18 non-achalasia) twice by using dedicated analysis software (ManoView 3.0). Integrated relaxation pressure (IRP4s), distal contractile integral (DCI), intrabolus pressurization pattern (IBP) and distal latency (DL) were extracted and analyzed hierarchically. Cohen's κ (2 raters) and Fleiss' κ (>2 raters) and the intraclass correlation coefficient (ICC) were used for categorical and ordinal data respectively.

Results: Based upon results of dedicated analysis software only, intra- and interrater reliability were *excellent* and *moderate* (κ=0.89 and κ=0.52 respectively) for differentiating achalasia from non-achalasia. For subtyping achalasia, reliability decreased to *substantial* and *fair* (κ=0.72 and κ=0.28 respectively). When observers were allowed to change the software-driven diagnosis according to their own interpretation of the manometric patterns, intra- and interrater reliability increased for diagnosing achalasia (κ=0.98 and κ=0.92 respectively) and for subtyping achalasia (κ=0.79 and κ=0.58 respectively).

Conclusion: Intra- and interrater agreement for diagnosing achalasia when using HRM and the CC was very good to excellent when results of automated analysis software were interpreted by experienced observers. More variability was seen when relying solely on the software-driven diagnosis and for subtyping achalasia. Therefore, diagnosing and subtyping achalasia should be performed in pediatric motility centers with significant expertise.

Key words: Manometry – Achalasia – Pediatrics – Reliability

List of abbreviations

BS, Break size; CC, Chicago Classification; CDP, Contractile deceleration point; DCI, Distal contractile integral; DL, Distal latency; EGJ, Esophageal gastric junction; EGJOO, Esophageal gastric junction outflow obstruction; EPT, Esophageal pressure topography; IBP, Intrabolus pressure; HRIM, High-resolution impedance manometry; HRM, High-resolution manometry; ICC, Intraclass correlation coefficient; LES, Lower esophageal sphincter; IRP4s, Integrated relaxation pressure; UES, Upper esophageal sphincter.

Background

Esophageal achalasia is characterized by failure of the lower esophageal sphincter (LES) to completely relax with swallowing in combination with aperistalsis in the smooth muscle esophagus.(1) Presenting symptoms may include dysphagia, regurgitation and/or vomiting, malnutrition and failure to thrive.(2-4)

The introduction of high-resolution manometry (HRM) has allowed for better characterization of esophageal motor function and uniform consensus of esophageal motility disorders.(5-7) Although both clinical and radiological findings can suggest achalasia, HRM is currently considered the gold standard for diagnosis and subtyping of achalasia.(8, 9) Based on HRM, the Chicago Classification V3.0 (CC) of esophageal motility disorders has defined three subtypes of achalasia, differentiated by the patterns of non-peristaltic esophageal pressurization accompanying abnormal relaxation pressure at the LES: type I (classic achalasia), type II (panesophageal pressurization), and type III (spastic achalasia).(9)

Since the CC was not validated in or created for the pediatric population, its implementation in pediatric HRM studies has been challenging.(10) In addition, manometric recordings from children may be harder to interpret due to a higher likelihood of multiple swallowing and artifacts due to body movement and crying.(11-13) A recent study on the inter- and intra-rater agreement (reliability) of the CC diagnosis of pediatric HRM recordings found high levels of agreement overall, whilst the diagnosis of achalasia subtypes appeared to be particularly challenging, even amongst raters considered to be experts on HRM analysis. However, that study focused on the broad application of the CC in children and only included a limited number of achalasia patients.(14) A study in adults showed excellent reliability of differentiating achalasia from non-achalasia, whilst reliability of subtyping achalasia appeared to show more variability, specifically regarding types I and II.(15)

Subtyping is not merely academic. Accumulated evidence in adult patients, uniformly showed the highest treatment success rates in patients with achalasia type II and worst response in patients with type III achalasia.(16-20)

Whether this subtyping has the same prognostic value in children is not known but the first step to understanding this is to identify if CC categorization is accurate and reproducible in children.(16-20) Therefore, the primary aim of this study was to assess the inter- and intraobserver reliability of interactive CC analysis software applied to HRM studies amongst specialists for the differentiation of non-achalasia from achalasia and its subtypes in a pediatric cohort.

Methods

Study database

Combined high-resolution impedance and manometry measurements (HRIM) of pediatric patients were extracted from a database of studies conducted at the Gastroenterology units of the Women's and Children's Hospital (Adelaide, Australia), Boston Children's Hospital (Boston, MA, USA), and the Emma Children's Hospital / Academic Medical Center (Amsterdam, The Netherlands) between September 2010 and June 2015. The typical manometric protocol used a 3.2 or 4.2-mm diameter solid state HRIM catheter incorporating 25 or 36 1-cm-spaced pressure sensors and 12 or 18 adjoining impedance segments, each of 2 cm (Given Imaging, Los Angeles, CA), depending on patient' s age and height. Patients were studied sitting in the supine or semi-supine position with a standard protocol including 3, 5 or 10 mL swallows (volume based on bolus tolerance and patient's age) administered into the mouth via a syringe at ≥30 s intervals. Studies were considered for inclusion if they met the following criteria: (i) 10 liquid swallows performed, (ii) adequate catheter position to resolve EGJ pressures, and (iii) no technical errors, e.g., pressure or impedance channel failure.

From these potential studies, a database of 40 de-identified studies (18 male; median age 14.6 [IQR 12.7 – 16.6] years) was created to assess intra- and interrater reliability. The final database included 22 randomly selected achalasia cases type I (n = 4, 18%), type II (n= 15, 68%), type III (n= 3, 14%) and 18 non-achalasia cases. In all cases, primary achalasia was clinically diagnosed by Authors SN and RR who are highly experienced GI consultants working in a tertiary Center. Per routine clinical practice, the clinical suspicion of achalasia was confirmed based on a combination of clinical findings,

barium-esophagram and HRM. A Chicago classification (i.e. achalasia subtype in case of achalasia patients and primary CC diagnosis in case of non-achalasia patients) was then determined by consensus among two experienced investigators during independent clinical diagnostic analysis by using the ManoView analysis software (TIO and MMJS; ESO 3.0, Given Imaging, Los Angeles, CA). Disagreements were adjudicated by discussion and consensus with a third-party arbiter (SN). This classification was used as a reference standard to compare the results of the software-generated and subjective CC diagnosis amongst observers. It was aimed to obtain a proportion of each subtype similar to that previously reported by Pandolfino et al.(21) Non-achalasia cases were included to assure that the raters could differentiate achalasia from non-achalasia cases and to eliminate the expectation bias that all cases were achalasia. Non-achalasia cases consisted of a distribution of normal motility (n = 11, 61 %) as well as the other primary major motor disorders that may have achalasia-like features. These included esophagogastric junction outflow obstruction (EGJOO; n = 3, 17%), frequent failed peristalsis (n = 2, 11%) and absent peristalsis (n = 2, 11%). At the time of initial investigation, all patients were enrolled in study protocols that were approved by the local Research Ethics Committees.

Data analysis

Each rater was provided with reference literature regarding the assessment of esophageal motility based on the CC V3.0.(7, 9) For this study, the adult cut-off criteria of the CC were used. All raters viewed an introductory PowerPoint tutorial explaining the correct use of the ManoView automated analysis software (Version 3.0, Given Imaging, Los Angeles, CA) and completed a practice run of a patient study in order to confirm they understood the requirements. Fact sheets detailing the principle steps of software analysis and the CC algorithm were provided for reference purposes at any stage of analysis. Raters were selected from different centers, based on their experience with HRM analysis of pediatric recordings (i.e. minimum 500 HRM analyses performed).

To assess intra-rater reliability, each rater analyzed the dataset twice, with at least 7 days between repeat analyses. All raters were blinded to the diagnosis of the patients, and all studies were de-identified. Also to avoid the potential for sequence bias, the order of studies was randomized between

raters and between repeat analyses. Raters were instructed to manually place or adjust the automatically populated landmarks. These included gastric position, EGJ proximal and distal margin, UES margins, transition zone, swallow onset, distal contractile integral (DCI) box, and contractile deceleration point (CDP). Swallow onset was defined by the relaxation of the UES. Raters were instructed to delete analysis landmarks if they considered them to be not applicable to the swallow (e.g., CDP and DCI box in circumstances of failed peristalsis). CC metrics driving the achalasia diagnosis and achalasia subtyping (integrated relaxation pressure (IRP4s), distal contractile integral (DCI), intrabolus pressurization pattern (IBP) and distal latency (DL)) were extracted and analyzed in a hierarchical order according to the CC algorithm. An overall CC diagnosis per study was automatically generated by the software based on these metrics. In addition to the software-based CC diagnosis, raters were asked to provide their own interpretation of the manometric patterns and change the software driven diagnosis accordingly.

Statistical analysis

Data were analyzed using IBM SPSS Statistics 20 (Chicago, Illinois). For categorical data, inter- and intra-rater reliability were calculated using Cohen's κ (2 raters, kappa further annotated as κ) and Fleiss' κ (> 2 raters). For ordinal data, the intraclass correlation coefficient (ICC) was used. The first session of analysis was used to determine interrater reliability. We additionally calculated interrater reliability for the second session to compare reliability between the two sessions. Fleiss' κ was calculated by using a premade syntax for SPSS (available from corresponding author). Statistical analysis on EPT metrics was performed based on mean values. In circumstances where landmarks were removed, preventing an EPT metric average being based on all 10 swallows, data were excluded from reliability analysis. Mean values for κ and ICC were calculated using the Fisher's Z-transformation ($Z = \operatorname{arctanh}(\kappa)$). We applied the commonly used, but arbitrary, scale for κ and ICC values: 0.00 = no agreement, 0.01 to 0.20 = slight agreement, 0.21 to 0.40 = fair agreement, 0.41 to 0.60 = moderate agreement, 0.61 to 0.80 = substantial agreement, 0.81 to 0.99 = excellent agreement, and 1.00 = perfect agreement.

Sample size estimate

The primary goal of the current study was to evaluate the levels of reliability between and within observers using the kappa statistic. The sample size was constrained by the fact that we chose to select a set of studies from three combined databases without replacements. Based on these constraints, we estimated the sample size based on the assumption that the null hypothesis for kappa would be no better than 0.50 (*moderate agreement*) compared with the alternative hypotheses that kappa would be >0.50 given that the true value of kappa was ~0.80 (*substantial – good agreement)*. Based on these assumptions, we determined that if two observers classified 40 cases each, the test would provide 80% power at $\alpha = 0.05$. These calculations were based on the methods of Cantor and have been used previously in studies with similar design.(15, 22, 23)

Results

Five raters completed the analysis of the study database twice. The mean time that elapsed between the first and second analysis was 65.2 ± 59.7 days (range 7 – 146 days). Of the 18 *non*-achalasia cases in the database, one patient (initial diagnosis absent contractility) was allocated an achalasia diagnosis (type II) by four of the six observers. Of the 22 achalasia cases in the database, six cases were diagnosed as *non*-achalasia based upon the results of the dedicated analysis software only by at least one of the observers. When observers were allowed to change the software-driven diagnosis according to their interpretation of the manometric pattern, *all* 22 achalasia cases were recognized as achalasia by the observers.

Intra- and interrater reliability of software-generated and subjective CC diagnosis

Based upon results of the dedicated analysis software only, intra- and interrater reliability were excellent and moderate (κ=0.89 and κ=0.52 respectively) for differentiating achalasia from non-achalasia cases (Table 1). For subtyping achalasia, reliability decreased to substantial and fair (κ=0.72 and κ=0.28 respectively).

The software-generated diagnosis was changed according to the observers' own interpretation of the manometric pattern in 15.4% of the total number of analyzed studies. Overall, change of the software-generated diagnosis did not differ between achalasia (13.4%) or non-achalasia cases (16.7%). In 17 patients, at least one observer decided to change the software-generated diagnosis according to his or her own interpretation (Table 2). Based on the observers' interpretation, intra- and interrater reliability increased for both diagnosing achalasia (κ=0.98 and κ=0.92 respectively) and for subtyping achalasia (κ=0.79 and κ=0.58 respectively).

Intra- and interrater reliability of software-derived EPT metrics

The mean Cohen's *k* statistics for intra- and interrater reliability of the software-derived EPT metrics between two sessions are shown in Table 3. When evaluating the hierarchical application of the CC algorithm, substantial to excellent intra- and interrater reliability was found for parameters driving the achalasia diagnosis (IRP4s [κ=0.79 and κ=0.78 respectively] and DCI [κ=0.77 and κ=0.81 respectively]). The parameters involved in subtyping achalasia cases (IBP and DL to respectively determine panesophageal pressurization and spasm) showed more variability (Tables 3 and 4).

Discussion

The current study is the first on the reliability of diagnosing and subtyping achalasia in pediatric patients based on HRM criteria. Based on software derived diagnosis, we found moderate to excellent reliability for differentiating achalasia from non-achalasia cases, whilst for subtyping achalasia, reliability decreased. In addition to the initial software-generated diagnosis, we incorporated results on the observers' own interpretation of the manometric patterns and found reliability of both diagnosing and subtyping achalasia to be higher for the observers' interpretations when compared to the software-generated diagnoses. This suggests that experienced observers may be more likely to rely on pattern recognition, rather than on the results of automated analysis only. The findings of our study support the clinical utility of HRM in the objective CC-based diagnosis of achalasia in pediatric patients.

However, as achalasia is a chronic disease without cure, it also stresses the importance of careful review of the motility studies by an expert before a final diagnosis of achalasia and most importantly before a subtype classification is made. Differences between software and subjective diagnosis might even be more substantial in clinical practice due to awareness of patients' clinical history.

We retrospectively analyzed those studies that were allocated a different diagnosis when based upon the observers' interpretation of the manometric pattern, rather than on the software-generated results only. The cases that observers decided to change the initial software-driven diagnosis of non-achalasia to achalasia, were either EGJ outflow obstruction, absent peristalsis or normal motility. In line with findings of our earlier study, this shows that observers tended to ignore the software-generated IRP4s value below the cut-off of 15mmHg to draw a final conclusion of achalasia (Figure 1A; patient 2 in Table 2).(14) This additionally indicates that while the software picked up some instance of intact peristalsis with normal latency, thereby hierarchically shifting the diagnosis from achalasia to EGJ outflow obstruction, observers interpreted the manometric pattern as panesophageal or spastic, resulting in a diagnosis of either achalasia type II or III (Figure 1B; patient 40 in Table 2). Regarding changes in achalasia subtyping, observers only changed the subtype of those studies that were classified by the software as having pan-esophageal pressurization patterns (i.e. type II achalasia) to either type I or III. Disagreements amongst observers mainly concerned this same issue, with some raters interpreting tracings with distal esophageal pressurizations limited to only a small segment of the esophagus, or the occurrence of multiple peaks corresponding to contractions along the spatial pressure variation plot as panesophageal. Discrepancies in achalasia type III diagnoses may well be explained by the high level of variability in DL, which is in line with our previous study, showing that the determination of the DL to be particularly challenging even amongst expert observers (figure 1B).(14)

Earlier adult studies reported issues regarding an optimal panesophageal pressurization cut-off value to define esophageal compression resulting in EGJ outflow obstruction in type II achalasia.(9, 24) As the IRP4s is influenced by patterns of distal esophageal contractility, instances of clinically evident achalasia with IRP4s < 15mmHg have shown to exist especially in type I achalasia patients with low intraesophageal pressures and type II achalasia patients with short periods of panesophageal

pressurizations.(24, 25) In an attempt to overcome the limitations of the CC in diagnosing and subtyping achalasia, a new approach has been developed whereby the duration of trans-EGJ-flow can be accurately estimated based on integrated pressure-impedance criteria using high resolution impedance-manometry.(26-28) In adult achalasia patients, trans-EGJ-bolus flow time (BFT) was significantly lower in patients with achalasia types I and II when compared to type III.(27) However, further studies are needed to explore the potential role of these novel parameters in diagnosing and subtyping achalasia in both adults and children.

One of the strengths of our study is that we tested reproducibility of CC-based diagnosis of pediatric HRM recordings in a large cohort of patients by experienced observers from four large academic referral centers worldwide. Patient studies were selected in such a way that distribution of the studies in the database matched the proportion of achalasia subtype as reported, and all patients were very well characterized clinically.(21) Additionally, we included non-achalasia cases to eliminate the expectation bias that all cases were achalasia, as well as the bias related to raters attempted to guess between classifications.

This study also has some limitations. Intra-rater reliability was assessed after a minimum of seven days, which could be considered short and may have resulted in observers recognizing some of the tracing from the initial session, although the mean time when the repeat measurements were done was more than two months and ranged from seven days to almost five months). A second limitation may be that observers were instructed to delete metrics from analysis if considered inapplicable to a swallow, which is inherent to the use of automated analysis software for the evaluation of esophageal motor disorders characterized by the absence of a (normal) peristaltic contraction pattern). This approach influenced statistical analysis of these particular EPT metrics, as patient studies were pair wise excluded from analysis when metrics were not uniformly obtained. Additionally, one of the observers did not provide a final or uniform diagnosis on the scoring sheet in some cases, which also resulted in pair wise exclusion of these studies.

In conclusion, applying the CC to children that have undergone HRM is reliable to distinguish achalasia from non-achalasia patients. We found high intra- and interrater agreement for differentiating achalasia from non-achalasia patients using HRM and the CC when results of

automated analysis software were interpreted by experienced observers. More variability was seen when relying on the software driven diagnosis and for subtyping achalasia patients, indicating the importance of expert evaluation of HRM tracings as well as need for improved HR(I)M criteria for achalasia subtyping. Subtyping achalasia may ultimately predict treatment outcomes and prognosis, and lead to better-targeted treatment options depending on the manometric subtype.

Financial disclosure: The authors have no financial relationships relevant to this article to disclose.
Funding source: No external funding for this manuscript.
Potential conflicts of interest: The authors have no conflicts of interest relevant to this article to disclose.
Author Contributions: *Study concept and design:* Singendonk, Rosen, Nurko, Omari, Benninga, van Wijk. *Acquisition, analysis, or interpretation of data:* Singendonk, Rosen, Oors, Rommel, Omari, Nurko. *Drafting of the manuscript:* Singendonk, Van Wijk. *Critical revision of the manuscript for important intellectual content:* All authors. *Administrative, technical, or material support:* Singendonk, Rosen, Nurko, Omari, Oors. *Study supervision:* Omari, Benninga, Nurko.

References

1 Franklin AL, Petrosyan M, Kane TD Childhood achalasia: A comprehensive review of disease, diagnosis and therapeutic management. World J Gastrointest Endosc 2014;6(4):105-11.

2 Hallal C, Kieling CO, Nunes DL, et al. Diagnosis, misdiagnosis, and associated diseases of achalasia in children and adolescents: a twelve-year single center experience. Pediatr Surg Int 2012;28(12):1211-7.

3 Hussain SZ, Thomas R, Tolia V A review of achalasia in 33 children. Dig Dis Sci 2002;47(11):2538-43.

4 Smits M, van Lennep M, Vrijlandt R, et al. Pediatric Achalasia in the Netherlands: Incidence, Clinical Course, and Quality of Life. J Pediatr 2016;169(110-5.e3.

5 Srinivas M, Balakumaran TA, Palaniappan S, et al. High resolution esophageal manometry--the switch from "intuitive" visual interpretation to Chicago classification. Indian J Gastroenterol 2014;33(2):157-60.

6 Kessing BF, Smout AJ, Bredenoord AJ Clinical applications of esophageal impedance monitoring and high-resolution manometry. Curr Gastroenterol Rep 2012;14(3):197-205.

7 Pandolfino JE, Fox MR, Bredenoord AJ, et al. High-resolution manometry in clinical practice: utilizing pressure topography to classify oesophageal motility abnormalities. Neurogastroenterol Motil 2009;21(8):796-806.

8 Kahrilas PJ Esophageal motor disorders in terms of high-resolution esophageal pressure topography: what has changed? Am J Gastroenterol 2010;105(5):981-7.

9 Kahrilas PJ, Bredenoord AJ, Fox M, et al. The Chicago Classification of esophageal motility disorders, v3.0. Neurogastroenterol Motil 2015;27(2):160-74.

10 Singendonk MM, Kritas S, Cock C, et al. Applying the Chicago Classification criteria of esophageal motility to a pediatric cohort: effects of patient age and size. Neurogastroenterol Motil 2014;26(9):1333-41.

11 Chumpitazi B, Nurko S Pediatric gastrointestinal motility disorders: challenges and a clinical update. Gastroenterol Hepatol (N Y) 2008;4(2):140-8.

12 Roman S, Damon H, Pellissier PE, et al. Does body position modify the results of oesophageal high resolution manometry? Neurogastroenterol Motil 2010;22(3):271-5.

13 Xiang X, Tu L, Zhang X, et al. Influence of the catheter diameter on the investigation of the esophageal motility through solid-state high-resolution manometry. Dis Esophagus 2013;26(7):661-7.

14 Singendonk MM, Smits MJ, Heijting IE, et al. Inter- and intrarater reliability of the Chicago Classification in pediatric high-resolution esophageal manometry recordings. Neurogastroenterol Motil 2015;27(2):269-76.

15 Hernandez JC, Ratuapli SK, Burdick GE, et al. Interrater and intrarater agreement of the chicago classification of achalasia subtypes using high-resolution esophageal manometry. Am J Gastroenterol 2012;107(2):207-14.

16 Pandolfino JE, Ghosh SK, Rice J, et al. Classifying esophageal motility by pressure topography characteristics: a study of 400 patients and 75 controls. Am J Gastroenterol 2008;103(1):27-37.

17 Ju H, Ma Y, Liang K, et al. Function of high-resolution manometry in the analysis of peroral endoscopic myotomy for achalasia. Surg Endosc 2016;30(3):1094-9.

18 Lee JY, Kim N, Kim SE, et al. Clinical characteristics and treatment outcomes of 3 subtypes of achalasia according to the chicago classification in a tertiary institute in Korea. J Neurogastroenterol Motil 2013;19(4):485-94.

19 Rohof WO, Salvador R, Annese V, et al. Outcomes of treatment for achalasia depend on manometric subtype. Gastroenterology 2013;144(4):718-25; quiz e13-4.

20 Salvador R, Costantini M, Zaninotto G, et al. The preoperative manometric pattern predicts the outcome of surgical treatment for esophageal achalasia. J Gastrointest Surg 2010;14(11):1635-45.

21 Pandolfino JE, Kwiatek MA, Nealis T, et al. Achalasia: a new clinically relevant classification by high-resolution manometry. Gastroenterology 2008;135(5):1526-33.
22 Cantor A Sample-size calculations for Cohen's kappa. Psychological Methods 1996:150-53.
23 Singendonk MM, Pullens B, van Heteren JA, et al. Reliability of the reflux finding score for infants in flexible versus rigid laryngoscopy. Int J Pediatr Otorhinolaryngol 2016;86(37-42.
24 Lin Z, Kahrilas PJ, Roman S, et al. Refining the criterion for an abnormal Integrated Relaxation Pressure in esophageal pressure topography based on the pattern of esophageal contractility using a classification and regression tree model. Neurogastroenterol Motil 2012;24(8):e356-63.
25 Kahrilas PJ, Bredenoord AJ, Fox M, et al. The Chicago Classification of esophageal motility disorders, v3.0. Neurogastroenterol Motil 2015;27(2):160-74.
26 Lin Z, Imam H, Nicodeme F, et al. Flow time through esophagogastric junction derived during high-resolution impedance-manometry studies: a novel parameter for assessing esophageal bolus transit. Am J Physiol Gastrointest Liver Physiol 2014;307(2):G158-63.
27 Lin Z, Carlson DA, Dykstra K, et al. High-resolution impedance manometry measurement of bolus flow time in achalasia and its correlation with dysphagia. Neurogastroenterol Motil 2015;27(9):1232-8.
28 Hong SJ, Bhargava V, Jiang Y, et al. A unique esophageal motor pattern that involves longitudinal muscles is responsible for emptying in achalasia esophagus. Gastroenterology 2010;139(1):102-11.

Table 1 – Intra- and interrater reliability for diagnosing and subtyping achalasia

	Manoview diagnosis (κ, 95%CI)					Clinical diagnosis (κ, 95%CI)				
	Achalasia vs no-achalasia (n=40 studies)	Subtyping achalasia (n=22 studies)	Type I	Type II	Type III	Achalasia vs no-achalasia (n=40 studies)	Subtyping achalasia (n=22 studies)	Type I	Type II	Type III
INTRA-RATER RELIABILITY										
Rater 1	0.90 (0.69 – 1.01)	0.49 (0.11 – 0.86)	1.00 (1.00 – 1.00)	0.58 (0.16 – 1.00)	-0.05 (-0.11 – 0.02)	0.90 (0.76 – 1.03)	0.48 (0.11 – 0.84)	0.39 (-0.10 – 0.88)	0.54 (0.15 – 0.94)	0.65 (0.01 – 1.28)
Rater 2	0.95 (0.85 – 1.05)	0.87 (0.62 – 1.12)	1.00 (1.00 – 1.00)	0.86 (0.60 – 1.13)	0.65 (0.01 – 1.28)	1.00 (1.00 – 1.00)	0.66 (0.31 – 1.01)	0.69 (0.30 – 1.09)	0.64 (0.26 – 1.01)	0.65 (0.01 – 1.28)
Rater 3	0.95 (0.85 – 1.05)	0.61 (0.31 – 0.91)	0.59 (0.20 – 0.99)	0.68 (0.36 – 1.00)	0.65 (0.01 – 1.28)	0.90 (0.76 – 1.04)	0.52 (0.21 – 0.83)	0.40 (-0.07 – 0.87)	0.64 (0.30 – 0.98)	0.65 (0.01 – 1.28)
Rater 4	0.75 (0.50 – 0.90)	0.41 (-0.08 – 0.90)	1.00 (1.00 – 1.00)	0.39 (-0.10 – 0.88)	NA[2]	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)
Rater 5	0.88 (0.72 – 1.04)	0.77 (0.34 – 1.20)	NA[2]	0.77 (0.34 – 1.20)	NA[2]	0.84 (0.66 – 1.01)	0.36 (-0.19 – 0.91)	NA[2]	0.33 (-0.25 – 0.91)	1.00 (1.00 – 1.00)
Rater 6	0.85 (0.70 – 1.01)	0.27 (-0.24 – 0.77)	0.65 (0.01 – 1.28)	0.24 (-0.28 – 0.75)	NA[2]	0.85 (0.62 – 0.96)	0.35 (-0.09 – 0.80)	0.33 (-0.25 – 0.91)	0.30 (-0.17 – 0.78)	0.65 (0.01 – 1.28)
MEAN[1]	0.89	0.72	0.96	0.63	NA[3]	0.97	0.79	0.73	0.72	0.89
INTERRATER RELIABILITY										
	0.52 (0.43 – 0.60)[3]	0.27 (0.19 – 0.45)[3]	0.16 (0.09 – 0.42)[3]	0.36 (-0.46 – 1.00)[3]	0.28 (-0.01 – 0.56)[3]	0.92 (0.84 – 1.00)	0.52 (0.43 – 0.60)	0.53 (0.27 – 0.79)	0.46 (-0.31 – 1.00)	0.73 (0.47 – 0.98)

κ, Cohen's kappa estimate for intra-rater reliability and Fleiss' kappa estimate for inter-rater reliability; CI, confidence interval. [1]Mean kappa values calculated after applying Fisher's Z-transformation.
[2] Not calculable as at least one of the variables was a constant, [3]One study was pairwise excluded due to missing data for one rater. [3]Fishers Z-transformation not possible due to negative κ.

Table 2 – Overview of changes of the software-generated diagnosis based upon the observers' interpretation of the manometric pattern

Initial classification (reference standard)[#] *(number in database; n, %)**		**Change of broad CC diagnosis** *(number of observers)*		**Change of achalasia subtype** *(number of observers)*	
Normal (n=5/11; 45%)	*Patient 11*	Normal to achalasia Type II	N = 1		
	Patient 14	Normal to IEM	N = 1		
	Patient 18	DES to normal	N = 1		
	Patient 25	EGJOO to normal	N = 1		
		EGJOO to achalasia Type II	N = 1		
		Achalasia Type II to absent contractility	N = 1		
	Patient 36	DES to IEM	N = 1		
Achalasia (n=8/22; 36%)					
Type 1 (n=3/4; 75%)	*Patient 7*			Achalasia Type II to Type I	N = 2
	Patient 27	Normal to achalasia Type I	N = 1	Achalasia Type II to Type I	N = 1
	Patient 34			Achalasia Type II to Type I	N = 2
Type II (n=3/15; 20%)	*Patient 2*	Absent contractility to achalasia Type III	N = 3		
	Patient 4	EGJOO to achalasia Type II	N = 1		
	Patient 12	Achalasia Type II to normal	N = 1		
Type III (n=1/3; 33%)	*Patient 1*	EGJOO to achalasia Type II	N = 2	Achalasia Type II to Type III	N = 1
EGJOO (n=3/3; 100%)	*Patient 40*	EGJOO to achalasia Type III	N = 2	Achalasia Type II to Type III	N = 1
	Patient 13	Normal to EGJOO	N = 1		
		DES to EGJOO	N = 1		
	Patient 37	Normal to EGJOO	N = 1		
		Normal to DES	N = 1		
		DES to IEM	N = 1		
Absent contractility (n=1/2; 50%)	*Patient 23*	Achalasia Type II to EGJOO	N = 1		
		EGJOO to achalasia Type II	N = 1		
		Achalasia Type II to absent contractility	N = 1		
Frequent failed peristalsis (n=1/2; 50%)	*Patient 30*	IEM to EGJOO	N = 1		

CC = Chicago Classification; IEM = ineffective esophageal motility; DES = distal esophageal spasm; EGJOO = Esophagogastric junction outflow obstruction

**Patients included in table if at least one observer decided to change the software-generated diagnosis according to his or her interpretation of the manometric pattern*

[#] *Reference standard consisted of a combination of clinical findings, radiography and independent HRIM analysis by two experienced analysts.*

Table 3. Reliability of CC metrics involved in achalasia diagnosing and subtyping

	All patient studies (n = 40 studies)		**Achalasia patient studies only** (n = 22 studies)
	Mean IRP4s (mmHg) (ICC, 95%CI)	Mean DCI (mmHg s^{-1} cm^{-1}) (ICC, 95%CI)	Mean DL (s) (ICC, 95%CI)
INTRA-RATER RELIABILITY			
Rater 1	0.98 (0.96 – 0.99)	0.73 (0.54 – 0.85)	NA[2]
Rater 2	0.95 (0.91 – 0.98)	0.77 (0.61 – 0.87)	NA[2]
Rater 3	0.95 (0.90 – 0.97)	0.88 (0.73 – 0.94)	0.79 (0.20 – 0.96)[3]
Rater 4	0.49 (0.21 – 0.70)	0.84 (0.73 – 0.91)	-1.98 (-2.78 – 0.66)[4]
Rater 5	0.93 (0.88 – 0.96)	0.88 (0.79 – 0.94)	0.34 (-1.92 – 0.98)[4]
Rater 6	0.85 (0.74 – 0.93)	0.125 (-0.19 – 0.42)	NA[2]
MEAN[1]	0.92	0.77	NA[5]
INTERRATER RELIABILITY			
	0.90 (0.86 – 0.94)	0.78 (0.69 – 0.86)	NA[2]

κ, Cohen's kappa estimate for intra-rater reliability and Fleiss' kappa estimate for inter-rater reliability; CI, confidence interval; IRP4s, integrated relaxation pressure; DCI, distal contractile integral; IBP, intrabolus pressurization pattern; DL, distal latency; NA, not applicable. [1]Mean kappa values calculated after applying Fisher's *Z*-transformation; [2]n=22 studies pairwise excluded, no studies to perform analysis; [3]n=15 studies pairwise excluded, analyses based on n=7 studies; [4]n=19 studies pairwise excluded, analyses based on n=7 studies; [5]Fishers Z-transformation not possible due to negative ICC.

Table 4. Reliability of the hierarchical application of the CC algorithm in the diagnosis of achalasia and its subtypes

	Intra-rater reliability, κ (Mean)[1]	Interrater reliability, κ
Diagnosing Achalasia (all cases; n = 40)		
1. Mean IRP4s > 15mmHg	0.79	0.78
2. 100% of swallows failed peristalsis	0.77	0.81
Subtyping Achalasia (achalasia cases only; n = 22)		
3. Panesophageal pressurization ≥ 20% of swallows	0.93	0.58
4. Spasm ≥ 20% of swallows	0.35[2]	NA[3]

κ, Cohen's kappa estimate for intra-rater reliability and Fleiss' kappa estimate for inter-rater reliability; IRP4s, integrated relaxation pressure, NA; not applicable.
[1]Mean kappa values calculated after applying Fisher's *Z*-transformation; [2]based on results of 3observers as for the other 3 observers n=22 studies were pairwise excluded (distal latency (DL) not uniformly obtained); [3]n=22 studies pairwise excluded (DL not uniformly obtained), no studies to perform interrater reliability analysis.

Legends:

Figure 1 – Differences in software-generated diagnosis based upon EPT metrics. A) Example of different placement of IRP4s box in patient with borderline IRP4s values (amongst observers: mean IRP4s = 15mmHg (range 13.9 – 16.4mmHg)). The IRP4s box should be spanning a 10 second time-frame, however this time-frame was adjusted by some observers as illustrated in this swallow. This patient was allocated a software-generated diagnosis of either absent contractility (n = 3 observers) or Achalasia Type II (n = 3 observers). The manometric pattern was interpreted as either Achalasia Type I or Achalasia Type II (n = 3 observers each). In other patients, differences in IRP4s values were related to placement of the swallow onset and/or the gastric pressure marker. **B)** Example of different placement of CDP by two independent observers. Observer 1: DL = 4.4s (swallow classified as spastic). Diagnosis of Achalasia Type III (mean IRP4s of 48mmHg with ≥ 20% spastic contractions). Observer 2: DL = 6.8s (DL classified as normal). Diagnosis of EGJ Outflow obstruction (mean IRP4s = 52.5mmHg with instances of intact peristalsis). Both observers interpreted the manometric pattern as consistent with Achalasia Type III.