Pediatric esophageal motility disorders: studies on (patho)physiology, diagnosis and management

Smits, M.J.

Citation for published version (APA):
CHAPTER 7

Inter- and intrarater reliability of the Chicago Classification in pediatric high-resolution esophageal manometry recordings


*Both authors contributed equally

Neurogastroenterology and Motility. 2015 Feb;27(2):269-76
ABSTRACT

Background
The Chicago Classification (CC) facilitates interpretation of high-resolution manometry (HRM) recordings. Application of this adult based algorithm to the pediatric population is unknown. We therefore assessed intra- and interrater reliability of software-based CC diagnosis in a pediatric cohort.

Methods
Thirty pediatric solid state HRM recordings (13M; mean age 12.1 ± 5.1 years) assessing 10 liquid swallows per patient were analyzed twice by 11 raters (six experts, five non-experts). Software-placed anatomical landmarks required manual adjustment or removal. Integrated relaxation pressure (IRP4s), distal contractile integral (DCI), contractile front velocity (CFV), distal latency (DL) and break size (BS), and an overall CC diagnosis were software-generated. In addition, raters provided their subjective CC diagnosis. Reliability was calculated with Cohen's and Fleiss' kappa (K) and intra-class correlation coefficient (ICC).

Results
Intra- and interrater reliability of software-generated CC diagnosis after manual adjustment of landmarks was substantial (mean K=0.69 and 0.77 respectively) and moderate-substantial for subjective CC diagnosis (mean K=0.70 and 0.58 respectively). Reliability of both software-generated and subjective diagnosis of normal motility was high (K=0.81 and K=0.79). Intra- and interrater reliability were excellent for IRP4s, DCI, and BS. Experts had higher interrater reliability than non-experts for DL (ICC=0.65 vs ICC=0.36 respectively) and the software-generated diagnosis diffuse esophageal spasm (DES, K=0.64 vs K=0.30). Among experts, the reliability for the subjective diagnosis of achalasia and esophageal gastric junction outflow obstruction was moderate-substantial (K=0.45-0.82).

Conclusion
Inter- and intrarater reliability of software-based CC diagnosis of pediatric HRM recordings was high overall. However, experience was a factor influencing the diagnosis of some motility disorders, particularly DES and achalasia.
INTRODUCTION

The recent introduction of high-resolution manometry (HRM) with esophageal pressure topography (EPT) into clinical practice has allowed for better characterization of esophageal motor function and uniform consensus on diagnosis of esophageal motility disorders. The 2012 Chicago Classification (CC) algorithm for esophageal motility, with application through interactive analysis software, facilitates the diagnostic interpretation of pressure recordings. The CC uses five EPT metrics based on ten liquid swallows and characterizes motor dysfunction into four main categories in order of severity, i.e., achalasia (Category 1), esophagogastric junction (EGJ) outflow obstruction (Category 2), disorders never observed in healthy individuals (Category 3; absent peristalsis, diffuse esophageal spasm [DES] or hypercontractile esophagus), and motor patterns outside the normal range (Category 4; weak peristalsis, frequent failed peristalsis, hypertensive peristalsis, or rapid contraction).

In the pediatric population, the spectrum of esophageal motility disorders resembles that seen in adults. However, CC implementation for use in the pediatric population is challenging. There are no established pediatric normative ranges for EPT metrics and some metrics, such as the integrated relaxation pressure (IRP4) and the distal latency (DL), have shown to be significantly influenced by patient age and size. Hence, if adult diagnostic CC criteria are not adjusted to account for these effects, it is likely that esophageal motility disorders, particularly EGJ outflow obstruction and DES, will be overdiagnosed. Pediatric HRM studies are also more challenging to perform and are more likely to have a greater incidence of incomplete studies with fewer than the 10 liquid swallows required. Finally, manometric recordings from children may be harder to interpret due to multiple swallowing and artifacts due to body movement and crying.

Derivation of EPT metrics and a CC diagnosis is reproducible and reliable when applied to the adult population. However, there are no equivalent data available based on the analysis of more challenging pediatric studies. Therefore, the aim of this study was to assess inter- and intrarater reliability of interactive CC analysis software for the diagnosis of esophageal motility disorders 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 Academic Medical Center (Amsterdam, The Netherlands) between December 2008 and September 2013. The typical manometric protocol used a 3.2-mm diameter solid state HRIM catheter incorporating 25 or 36 1-cm-spaced pressure sensors and 12 adjoining impedance segments, each of 2
cm (Unisensor USA Inc, Portsmouth, NH, USA). If the pressure-impedance sensor array was not long enough to accommodate the entire region from upper esophageal sphincter (UES) to EGJ, the catheter was positioned with sensors straddling the distal esophagus from transition zone to stomach. Patients were studied sitting in the supine or semi-supine position with a standard protocol including 10 × 3, 5 or 10 mL swallows (volume based on bolus tolerance and patient’s age) administered via 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 30 de-identified studies (13M; mean age 12.1 ± 5.1 years) was created to assess intra- and interrater reliability. Using the original diagnostic findings as a guide, all four main categories of CC disorders were represented in the database. Furthermore, the distribution of disorders in the study database was designed to be consistent with the overall distribution within the more extensive database. The database consisted of: ten patients (33%) with normal peristalsis, seven (23%) with weak peristalsis (five large breaks), six (20%) with EGJ outflow obstruction, two (7%) with frequent failed peristalsis, two (7%) with distal esophageal spasm and two (7%) with achalasia (type I and type II), and one (3%) with absent peristalsis. Diagnosis was based upon the original expert defined clinical diagnostic analysis (consensus reliability after independent analysis by TIO and MMJS, disagreements were adjudicated by discussion and consensus with a third-party arbiter [SK]).

**Data analysis**

Each rater was provided with reference literature regarding the assessment of esophageal motility based on EPT metrics and the CC. For this study, the commonly used adult cut-off criteria for the EPT metrics and CC diagnoses were used. All raters watched an introductory tutorial explaining the correct use of the MMS automated analysis software and completed a practice run of a patient study to confirm they were proficient. Fact sheets detailing the principle steps of software analysis and the CC algorithm were provided for reference purposes at any stage of analysis. Raters with varying levels of experience with esophageal manometry were invited to participate. Raters with experience from >200 HRM analyses were considered ‘experts’. To assess intrarater reliability, each rater analyzed the dataset twice, with at least 7 days between repeat analyses. To avoid the potential for sequence bias, the order of studies was randomized between raters and between repeat analyses.

Patient studies were analyzed using the MMS analysis software, version 8.23 (MMS, Enschede, The Netherlands). 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. If UES pressures were not visible, the onset
of impedance drop in the most proximal impedance segment was used. 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). Following completion of analysis, the standard EPT metrics (per swallow and mean of 10 swallows) were derived by the software. These were, (i) IRP4 (mmHg), (ii) contractile front velocity (CFV, cm/s), (iii) DCI (mmHg cm s), (iv) DL (s), and (v) peristaltic 20 mmHg isocontour break size (BS, cm).4 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 subjective opinion on the CC diagnosis for each patient.

**Statistical analysis**

Data were analyzed using IBM SPSS Statistics 20 (Chicago, Illinois). For categorical data, inter- and intrarater reliability were calculated using Cohen’s K (2 raters, kappa further annotated as $K$) and Fleiss’ $\kappa$ (>2 raters). For ordinal data, the intraclass correlation coefficient (ICC) was used. The first session of analysis was used to determine intrarater reliability. We additionally calculated inter-rater reliability for the second session to compare reliability between the two sessions. Fleiss’ $\kappa$ 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 $\kappa$ and ICC were calculated using the Fisher’s Z-transformation ($Z=\text{arctanh}(\kappa)$). We applied the common scale for $\kappa$ 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.

**RESULTS**

**Intrarater reliability of software-derived EPT metrics**

The mean Cohen’s $\kappa$ statistics for intrarater reliability between the two sessions are shown in Table 1. Overall, excellent intrarater reliability was noted for the metrics IRP4, DCI, DL and BS, and substantial for CFV. The intrarater reliability was generally similar for experienced and inexperienced raters. However, DL was less reliably scored by experts.

**Interrater reliability of EPT metrics**

Among all raters, excellent reliability was reached for the metrics IRP4s, DCI, and BS, and level of reliability appeared to be independent of level of experience (Table 1). Reliability of CFV and DL was moderate and depended strongly on level of experience, with higher levels of reliability among the group of experienced raters compared to the inexperienced raters (ICC=0.80 vs. ICC=0.39 respectively for CFV and ICC=0.65 vs ICC=0.36 respectively for DL). Among inexperienced raters, reliability of DL was higher for the second session of analysis ($\kappa$=0.56). There were no marked differences in reliability between sessions in relation to the other metrics.
Intra- and interrater reliability of software-generated and subjective CC diagnosis

Both intra- and interrater reliability of the software-generated CC diagnosis were substantial among all raters and did not depend on level of experience (Table 2). The initial software-generated diagnosis was changed according to the raters’ personal opinion in 32.1% of all studies. Experienced raters were more likely to change the software-generated diagnosis in comparison to the inexperienced raters (34.4% vs. 29.3%). Both intra- and interrater reliability were lower for the subjective CC diagnosis, but remained fair to substantial (Table 2). Among experienced raters, a higher level of interrater reliability on the subjective CC diagnosis was reached when compared to the inexperienced raters ($\kappa=0.56$ vs. $\kappa=0.48$). We observed an increase in interrater reliability of subjective CC diagnosis in the second session of analysis among inexperienced raters only ($\kappa=0.62$). Reliability of software-generated diagnosis did not differ between sessions.

Interrater reliability of the individual software-generated and subjective CC diagnoses

Results on interrater reliability for all individual software-generated CC diagnoses are displayed in Table 3A. Highest reliability among all raters was reached for the diagnoses of normal motility, EGJ outflow obstruction and absent peristalsis ($\kappa=0.81$, $\kappa=0.79$ and $\kappa=0.84$ respectively), and level of reliability appeared to be independent of experience. Level of reliability for the software-generated diagnosis DES was higher in the group of experienced raters ($\kappa=0.64$ vs. $\kappa=0.30$ for non-experts). Table 3B displays the levels of interrater reliability for the subjective CC diagnoses based on raters’ personal opinions. Although overall reliability of the CC diagnosis of normal motility was substantial, the level of reliability among experienced raters was substantially higher when compared to the group of inexperienced raters ($\kappa=0.79$ vs. $\kappa=0.49$ respectively). Reliability based on subjective diagnosis of the two most severe CC diagnoses, achalasia and EGJ outflow obstruction, ranged from moderate to substantial ($\kappa=0.45$ to $\kappa=0.82$) among expert raters.

Influence of visibility of the UES on reliability

We performed sub-analysis to determine the possible influence of the lack of visibility of the UES during analysis of patient studies (n=12). Visibility of the UES did not influence the level of interrater reliability for all metrics, aside from DL. For DL, interrater reliability appeared to be higher in studies in which the UES was completely in view among inexperienced raters albeit on a background of relatively poor reliability overall ($\kappa=0.19$ vs. $\kappa=0.44$). However, when the level of experience was taken into account, the overall level of reliability for determining DL was far better in experienced raters and visibility of the UES had no effect (among experts $\kappa=0.69$ vs. $\kappa=0.64$ when UES was absent or present, respectively). For the overall CC diagnosis, there were no differences in reliability between studies with and without UES high pressure zone in view suggesting that variability in placement of the CDP, rather than swallow onset, was the main driver of variability in the DL measurement (Figure 1).
DISCUSSION

This study is the first to report inter- and intrarater reliability for both software and personal opinion-based CC diagnosis of esophageal motility disorders by HRM in a pediatric cohort. Using semi-automated software assisted analysis, we demonstrated high reliability and reproducibility of CC-based diagnoses of esophageal motility disorders among both experienced and inexperienced raters. Our findings are consistent with earlier studies in adult cohorts, indicating that the analysis software is easy to learn and can be easily implemented in children as well.16,17 As expected, reliability was generally lower among inexperienced raters showing that training is needed to ensure diagnostic reliability. The findings of our study support the clinical utility of HRM in the objective CC-based diagnosis of esophageal motor disorders in pediatric patients. On the other hand, our findings also highlight diagnostic challenges specific to the pediatric population.

Our study identified areas of diminished reliability of CC-based diagnosis of pediatric manometric esophageal recordings. This was notable with regard to the reduced reliability and reproducibility of derivation of the DL metric, which in turn contributes to a poorer reliability of the DES diagnosis, particularly among inexperience raters. We have recently shown that in pediatric patients, DL varies in an age- and esophageal length-dependent manner.9 Deriving the DL might therefore be particularly challenging in pediatric patients and may explain the low levels of reliability on DL and DL-driven diagnosis of DES among raters naïve to HRM analysis. The DL is calculated from the swallow onset to the CDP. Manually reviewing all studies, we observed consistent placement of the swallow onset among and between raters. In challenging studies, large variations in the placement of the CDP were observed (Figure 1). This illustrates that consistent placement of the CDP is particularly challenging, even among experienced raters.

We additionally found low levels of reliability for software-derived diagnosis of achalasia subtypes, whereas subjective diagnosis was moderate to substantial among expert raters. To further explore this discrepancy, we manually reviewed the study database. Two patients in this cohort were given an initial diagnosis of achalasia based on expert opinion, prior to the study. Marked differences were noted in the diagnosis of achalasia when comparing the software-based and subjective diagnoses for individual raters. Patient studies that were classified with an achalasia disorder based on subjective diagnosis were either allocated the diagnosis EGJ outflow obstruction or absent peristalsis by software-driven analysis. Both achalasia and EGJ outflow obstruction require an IRP4 ≥15 mmHg, but differ in that, for EGJ outflow obstruction, there is some evidence of an esophageal peristaltic wave which demonstrates normal latency. Retrospective analysis of studies that were allocated a software-driven diagnosis of EGJ outflow obstruction showed that studies subjectively classified as having pan-esophageal pressurization patterns were allocated a ‘normal peristaltic pattern’ by the software. The diagnosis of absent peristalsis requires an IRP4s <15 mmHg in combination with 100% failed swallows.4 Retrospective analysis of EPT metrics revealed that the criterion of an abnormal IRP4s was ignored in the subjective diagnosis of achalasia by three of the expert raters. This finding
might indicate that the decision to change the software-derived diagnosis of absent peristalsis toward an achalasia diagnosis was rather based upon pattern recognition and clinical expertise, than on EPT metrics and stresses the importance of careful review of the motility studies before a diagnosis is made. The diagnostic approach of achalasia is changing and it has been suggested to reduce the IRP4s criteria for Type 1 achalasia to 10 mmHg. Applying this criterion might potentially enhance diagnostic accuracy, leading to a more uniform diagnosis of achalasia and achalasia subtypes.

Another diagnostic category with low reliability is weak peristalsis with small or large peristaltic breaks. There are several possible explanations for this finding. Firstly, these diagnoses did not exist with standard manometry, so incorporation of these findings into routine manometry readings lag behind. Secondly, a more likely explanation is that clinical significance of these diagnoses remains uncertain. While adult studies have shown that the length of esophageal breaks correlates with incomplete bolus transit, no comparable data exist in children. Furthermore, early studies with esophageal function testing suggest that there is an imperfect correlation between bolus transit and some peristaltic defects, particularly non-specific motor disorders. Therefore, while the manometric findings may show breaks in peristalsis, a normal bolus clearance pattern combined with the lack of data to show clinical significance of breaks in children may result in underreporting of a diagnosis with unknown significance.

Our study has several limitations. Firstly, part of the included HRM studies did not incorporate the UES due to limited pressure sensor array length (n=12, 40%). In the absence of visual confirmation, swallow onset is more difficult to determine. Direct visualization of the UES relaxation onset, however, did not appear to influence reliability of overall CC diagnosis. It could nevertheless have influenced reliability and reproducibility of IRP4s and DL, as the calculation of these metrics relies on the position of the swallow onset marker. In our experience, IRP4s calculation is very resilient to swallow onset difference of up to ±1 sec and reliability of IRP4s appeared to be high. Therefore, localization of swallow onset was thought to, more likely, potentially influence the calculation of DL. However, sub-analysis that compared studies with the UES high pressure zone present or absent suggested that the main driver of variability in the DL was location of the CDP, rather than the timing of swallow onset (as illustrated in Figure 1). Importantly, our data show that studies with incomplete capture of the UES can still be analyzed accurately.

A second limitation was that raters were instructed to delete metrics from analysis if considered inapplicable to a swallow (i.e., in the absence of a peristaltic contraction pattern). This approach influenced statistical analysis of EPT metrics CFV, DL, and DCI, as patient studies were pair-wise excluded from analysis when metrics were not uniformly obtained. We therefore also assessed whether raters would consistently apply these landmarks to a swallow. The combined finding of low reliability on both the applicability and the values of the CFV and DL points out that these metrics might be particularly challenging. Furthermore, reliability between first and second analysis
was tested after a minimum of 7 days. Although raters were blinded from patient characteristics and studies were presented in a randomized order for both sessions, this short period could have resulted in raters recognizing some of the patient studies from their initial session. Another limitation is that the limited number of achalasia patients included might have resulted in the lower reliability for its diagnosis, so future studies that focus on this entity will be needed to establish if the CC criteria for the diagnosis of achalasia in children need to be modified.

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 and selected a variety of observers with different background and varying experience. Patient studies were selected in such way that the distribution of the study database matched the distribution of the four broad CC categories earlier reported in a large cohort of pediatric patients referred for manometry. Finally, we incorporated software-derived diagnosis and the personal opinion of raters (subjective diagnosis). Differences between software and subjective diagnosis might be more substantial in clinical practice due to awareness of patients’ clinical history.

In conclusion, automated software-based CC diagnosis of pediatric esophageal motility disorders shows high inter- and intrarater reliability among experts and non-experts. However, the application of CC-based semi-automated software is least reliable and most influenced by raters’ expertise for the diagnosis of disorders such as DES and achalasia. Several issues regarding the application and interpretation of the CC specific to the pediatric population need to be addressed in future studies.
Figure 1 - Swallow of patient classified with normal motility according to initial analysis. Swallow onset (SO) marker and contractile deceleration point (CDP) displayed for all raters. Swallow onset marker uniformly placed at the same point by all raters. Large spread in placement of CDP (1.1 sec), and thus distal latency, between all raters. Difficulties placing the contractile deceleration point marker may therefore explain reduced reliability of the distal latency, rather than the determination of the swallow onset. Yellow: expert raters; red: non-expert raters; white lines: impedance tracing.
<table>
<thead>
<tr>
<th></th>
<th>Mean ICC (range)</th>
<th></th>
<th>Mean ICC (range)</th>
<th></th>
<th>Mean ICC (range)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All raters</td>
<td>Experienced raters</td>
<td>Inexperienced raters</td>
<td>All raters</td>
<td>Experienced raters</td>
<td>Inexperienced raters</td>
</tr>
<tr>
<td><strong>IRP4 (mmHg)</strong></td>
<td>0.92 (0.87 - 0.988)</td>
<td>0.98 (0.96 - 0.988)</td>
<td>0.96 (0.87 - 0.97)</td>
<td>0.94 (0.90 - 0.97)</td>
<td>0.96 (0.93 - 0.98)</td>
<td>0.91 (0.86 - 0.95)</td>
</tr>
<tr>
<td><strong>CFV (cm sec⁻¹)</strong></td>
<td>0.63 (-0.28 - 0.93)</td>
<td>0.62 (-0.28 - 0.88)</td>
<td>0.64 (-0.26 - 0.93)</td>
<td>0.54 (0.39 - 0.71)</td>
<td>0.80 (0.68 - 0.89)</td>
<td>0.39 (0.21 - 0.60)</td>
</tr>
<tr>
<td><strong>DCI (mmHg.cm.sec)</strong></td>
<td>0.97 (0.51 - 0.99)</td>
<td>0.99 (0.98 - 0.99)</td>
<td>0.92 (0.51 - 0.98)</td>
<td>0.89 (0.83 - 0.94)</td>
<td>0.97 (0.95 - 0.99)</td>
<td>0.81 (0.70 - 0.90)</td>
</tr>
<tr>
<td><strong>DL (sec)</strong></td>
<td>0.84 (0.36 - 0.98)</td>
<td>0.72 (0.40 - 0.89)</td>
<td>0.91 (0.36 - 0.98)</td>
<td>0.48 (0.33 - 0.66)</td>
<td>0.65 (0.47 - 0.80)</td>
<td>0.36 (0.18 - 0.58)</td>
</tr>
<tr>
<td><strong>BS (cm)</strong></td>
<td>0.98 (0.94 - 1.00)</td>
<td>0.99 (0.98 - 0.99)</td>
<td>0.96 (0.94 - 1.00)</td>
<td>0.92 (0.87 - 0.95)</td>
<td>0.92 (0.88 - 0.96)</td>
<td>0.90 (0.83 - 0.95)</td>
</tr>
</tbody>
</table>

Table 1. Intra- and interrater reliability of objective EPT metrics. ICC, Intraclass correlation coefficient; CI, confidence interval.  
*Mean kappa values calculated after applying Fisher's Z-transformation.
<table>
<thead>
<tr>
<th></th>
<th>Intrarater reliability</th>
<th>Interrater reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean K (range)*</td>
<td>mean K (95% CI)</td>
</tr>
<tr>
<td></td>
<td>Chicago Classification (objective diagnosis)</td>
<td>Chicago Classification (subjective diagnosis)</td>
</tr>
<tr>
<td>All</td>
<td>0.77 (0.49 - 1.00)</td>
<td>0.70 (0.26 - 1.00)</td>
</tr>
<tr>
<td>Experienced raters</td>
<td>0.78 (0.49 - 1.00)</td>
<td>0.70 (0.33 - 0.99)</td>
</tr>
<tr>
<td>Inexperienced raters</td>
<td>0.77 (0.53 - 0.97)</td>
<td>0.71 (0.26 - 1.00)</td>
</tr>
</tbody>
</table>

Table 2. Overall intra- and interrater reliability of objective software derived and subjective software guided Chicago Classification diagnosis. κ, Cohen's kappa estimate for intra-rater reliability and Fleiss' kappa estimate for inter-rater reliability; CI, confidence interval.

* Mean kappa values calculated after applying Fisher's Z-transformation
<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Achalasia Type I</th>
<th>EGJ outflow obstruction</th>
<th>Diffuse esophageal spasm</th>
<th>Absent Peristalsis</th>
<th>Weak peristalsis - LB</th>
<th>Weak peristalsis - SB</th>
<th>Frequent failed peristalsis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K (95% CI)</td>
<td>K (95% CI)</td>
<td>K (95% CI)</td>
<td>K (95% CI)</td>
<td>K (95% CI)</td>
<td>K (95% CI)</td>
<td>K (95% CI)</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>0.81</td>
<td>0.07</td>
<td>0.79</td>
<td>0.43</td>
<td>0.84</td>
<td>0.73</td>
<td>0.45</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>(0.65 - 0.97)</td>
<td>(-0.06 - 0.19)</td>
<td>(0.62 - 0.96)</td>
<td>(0.32 - 0.55)</td>
<td>(0.73 - 0.95)</td>
<td>(0.59 - 0.87)</td>
<td>(0.33 - 0.56)</td>
<td>(0.10 - 0.35)</td>
</tr>
<tr>
<td>Experienced raters</td>
<td>0.87</td>
<td>0.18</td>
<td>0.78</td>
<td>0.64</td>
<td>0.83</td>
<td>0.85</td>
<td>0.35</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>(0.62 - 1.00)</td>
<td>(-0.08 - 0.44)</td>
<td>(0.52 - 1.00)</td>
<td>(0.43 - 0.86)</td>
<td>(0.61 - 1.00)</td>
<td>(0.60 - 1.00)</td>
<td>(0.14 - 0.57)</td>
<td>(-0.16 - 0.40)</td>
</tr>
<tr>
<td>Inexperienced raters</td>
<td>0.72</td>
<td>-0.007</td>
<td>0.81</td>
<td>0.30</td>
<td>0.80</td>
<td>0.65</td>
<td>0.61</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>(0.42 - 1.00)</td>
<td>(-0.54 - 0.52)</td>
<td>(0.49 - 1.00)</td>
<td>(0.04 - 0.57)</td>
<td>(0.52 - 1.00)</td>
<td>(0.38 - 0.92)</td>
<td>(0.34 - 0.88)</td>
<td>(-0.16 - 0.64)</td>
</tr>
</tbody>
</table>

Table 3a. Fleiss' k statistic for interrater reliability of objective Chicago Classification diagnosis. k, Fleiss' kappa estimate for interrater reliability; CI, confidence interval; LB, large breaks in the 20mmHg isobaric contour (> 5 cm); SB, small breaks in the 20mmHg isobaric contour (2 cm > 5).
<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Achalasia Type I</th>
<th>Achalasia Type II</th>
<th>EGJ outflow obstruction</th>
<th>Diffuse esophageal spasm</th>
<th>Absent Peristalsis</th>
<th>Weak peristalsis - LB</th>
<th>Weak peristalsis - SB</th>
<th>Frequent failed peristalsis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>0.62 (0.43 - 0.80)</td>
<td>0.45 (0.33 - 0.55)</td>
<td>0.75 (0.63 - 0.86)</td>
<td>0.77 (0.64 - 0.91)</td>
<td>0.38 (0.27 - 0.50)</td>
<td>0.52 (0.40 - 0.62)</td>
<td>0.62 (0.49 - 0.76)</td>
<td>0.22 (0.11 - 0.33)</td>
<td>0.29 (0.18 - 0.40)</td>
</tr>
<tr>
<td><strong>Experienced raters</strong></td>
<td>0.79 (0.52 - 1.00)</td>
<td>0.45 (0.21 - 0.68)</td>
<td>0.83 (0.61 - 1.00)</td>
<td>0.82 (0.59 - 1.00)</td>
<td>0.49 (0.27 - 0.72)</td>
<td>0.92 (0.70 - 1.00)</td>
<td>0.72 (0.49 - 0.96)</td>
<td>0.31 (0.10 - 0.53)</td>
<td>0.38 (0.14 - 0.62)</td>
</tr>
<tr>
<td><strong>Inexperienced raters</strong></td>
<td>0.49 (0.14 - 0.84)</td>
<td>0.55 (0.28 - 0.81)</td>
<td>0.62 (0.35 - 0.90)</td>
<td>0.75 (0.47 - 1.00)</td>
<td>0.24 (-0.16 - 0.64)</td>
<td>0.15 (-0.20 - 0.50)</td>
<td>0.51 (0.24 - 0.78)</td>
<td>0.22 (-0.07 - 0.51)</td>
<td>0.17 (-0.10 - 0.45)</td>
</tr>
</tbody>
</table>

Table 3b. Fleiss’ K statistic for interrater reliability of subjective Chicago Classification diagnosis. K, Fleiss’ kappa estimate for interrater reliability; CI, confidence interval; LB, large breaks in the 20mmHg isobaric contour (> 5 cm); SB, small breaks in the 20mmHg isobaric contour (2 ≤ cm  >— 5).
REFERENCES


Chapter 7

Contributors' statement

Marije Smits and Maartje Singendonk: conceptualized and designed the study, built the data base, initialized the study, extracted and analyzed the data, drafted the initial manuscript, and approved the final manuscript as published. MMS and MS contributed equally to the manuscript.

Ilja Heijting: extracted and analyzed the data, drafted parts of the initial manuscript, reviewed and revised the manuscript, and approved the final manuscript as submitted.

Michiel van Wijk, Marc Benninga, Taher Omari, Rachel Rosen and Samuel Nurko: reviewed and revised the manuscript, and approved the final manuscript as submitted.

Maartje Singendonk, Ilja Heijting, Samuel Nurko, Rachel Rosen, Pim Wijenborg, Rammy Abu-Assi, Daniel Hoekman, Sophie Kuizenga-Wessel, Grace Seiboth, Taher Omari, Stamatiki Kritas: analyzed the selected patients studies, reviewed and approved the final manuscript as submitted.

Stamatiki Kritas, as senior author, reviewed and revised the manuscript, and approved the final manuscript as submitted.