Gastroesophageal reflux in children: the use of pH-impedance measurements and new insights in treatment

Loots, C.M.

Publication date
2011

Inter- and intra observer variability in pH-impedance analysis between ten experts and automated analysis

5

Clara Loots
Michiel van Wijk
Kathleen Blondeau
Kasper Dalby
Laura Peeters
Rachel Rosen
Silvia Salvatore
Tobias Wenzl
Yvan Vandenplas
Marc Benninga
Taher Omari

J Pediatr. 2011 Sep 15. [Epub ahead of print]

ABSTRACT

Gastroesophageal reflux (GER) detection on pH-impedance tracings is based on pattern recognition. Although automated analysis (AA) exists, most investigators prefer manual analysis, possibly introducing inter- and intraobserver variability.

Objective

To determine inter- and intraobserver variability in pH-impedance interpretation between experts and accuracy of AA.

Study design

Ten pediatric 24-hr pH-impedance tracings were analyzed by ten observers from seven world groups and with AA. Detection of GER episodes was compared between observers and AA. Intraobserver agreement was assessed in three observers after three to five months.

Results

Overall 1242 liquid and mixed GER events were detected, 490 (42%) were scored by the majority, yielding moderate agreement (κ = 0.46). Intraclass coefficient for numbers of GER per study was 0.84 ($p<0.001$). AA has 94% sensitivity and 74% specificity compared to majority consensus (≥6 observers). Agreement for gas GER was poor (κ = 0.11). Intraobserver agreement was κ = 0.49, κ = 0.71, and κ = 0.85 in 3 observers.

Conclusion

Interobserver agreement in pH-MII analysis amongst experts is moderate, only 42% of GER episodes was detected by the majority of observers. Detection of total GER numbers is more consistent. Considering these poor outcomes, AA seems favourable over manual analysis due to its reproducibility. However, the lower specificity suggest the need for refinement of AA before widespread use can be advocated.

INTRODUCTION

Combined pH-multichannel intraluminal impedance (pH-MII) has been used increasingly to assess gastroesophageal reflux (GER) in infants, children and adults. This technique is now recommended by ESPGHAN for the detection of GER in pediatric patients.[1] Esophageal pH-MII detects bolus movement in the esophagus, allowing assessment of not only acid GER but non acid GER as well.[2-4] In infants and children particularly pH-MII has been shown to detect more GER than pH-metry alone.[5,6] Detecting GER by pH-metry alone underestimates the amount of GER.[2,7] Adding MII significantly improves the yield of assessing GER – symptom associations.[5,8]

Detection of GER on pH-MII tracings is based on pattern recognition. Criteria for detection of bolus GER have been defined.[9] All available software packages use these criteria as a basis on which the automated analysis (AA) is built. However, AA is not validated and most investigators prefer manual analysis of pH-MII tracings to ensure confidence in marking of GER episodes. This introduces the potential for inter- and intra observer variability. Several papers assessing inter- and intra observer variability for the analysis of pH-MII tracings have been published,[10-13] however the observers in these papers were all from one group. Agreement between investigators from different groups and between AA and the consensus between observers is unknown.

The aim of this study is to 1) determine inter observer variability in pH-MII interpretation between ten experts in pH-MII analysis and pediatric GER, 2) determine the accuracy of AA compared to majority observer consensus and 3) assess intra observer variability in three investigators.

METHODS

Patients and tracings

All pH-MII studies were performed in children and infants (median age 4.5 years range 4 months – 14 years) who were referred for evaluation of GER symptoms. Catheters were transnasally positioned in the esophagus and the position was confirmed, based on X-ray thorax or videofluoroscopy.[1]

Ten 24hr pH-MII tracings with different characteristics were selected from a research database. Five tracings were considered 'easy' to analyze due to clear GER patterns with clear retrograde patterns and GER extending high into the esophagus. Five tracings were considered more challenging as the GER patterns were less obvious due to low baselines, retrograde patterns during swallowing and moving/crying artifacts. The 24hr pH-MII tracings were recorded with the Omega ambulatory system (Medical Measurement Systems, Enschede, The Netherlands). All tracings had >20 hrs of recorded pH-MII measurements. The tracings were randomized and distributed without markers from AA

to the observers. Observers had different levels of experience in the analysis of pH-MII tracings, ranging from six months to >15 years, having analyzed 100 to >2000 pH-MII tracings.

Inter observer analysis

Tracings were analyzed by ten experts in pediatric GER and analysis of pH-MII tracings from seven groups around the world and with AA (MMS Omega ambulatory Autoscan version 8.17, with standard settings and the option to reduce over-detection selected). Observers were asked to analyze the ten tracings in the same manner as they would analyze a pH-MII tracing in their hospital, including liquid, mixed and gas GER episodes.
Observers were also asked to provide their 'personal guidelines' for pH-MII analysis. Observers commented on the use of AA, color contour plot and whether they follow the current impedance analysis guidelines.
Reports from the tracings were created, meal time was excluded from analysis. Liquid, mixed and gas GER episodes were analyzed. Liquid and mixed GER were grouped together for analysis.
Tracings were compared for the recognition of GER episodes. For the assessment of the detection of GER episodes, all GER episodes scored by ≥1 observer, the exposure time per GER episode and the point of nadir impedance were recorded. If observers recognized GER in the same timeframe, including the point of nadir impedance, that episode was scored positive for both observers. If one observer recognized one long reflux episode and another observer recognized two shorter GER episodes, the longest timeframe was chosen and both observers scored positive for the longer timeframe (Figure 1).
We sought to assess the clinical impact of the number of GER episodes detected in terms of defining a study pathological. Normative data in pediatrics do not exist and although adult normative data are not transferable to pediatric patients in clinical setting, the 95th percentile cut off value of 73 GER episodes in 24hrs [14] was used in this research setting to assess the impact of the number of detected GER episodes on a positive or negative study outcome. As the value of this cut off in arbitrary in the pediatric population, we assessed the impact of different cut off values ranging from 40 to 101 GER episodes per 24hrs, the cut off value in preterm infants.[15]
Furthermore, AA was compared to the observers' consensus. The majority consensus was defined as GER episodes scored positive by any majority (≥6) of observers.

Intra observer analysis

Three to five months after the first analysis, three observers re-analyzed the same tracings for assessment of intra observer variability. For the purpose of intra observer analysis, all GER episodes marked in the first and in the second analysis by one observer were compared. If the observer recognized one long GER episode in one analysis and two short GER episodes in the second analysis, one GER episode was scored positive twice and one GER episode was scored once. To calculate agreement, the 'truly negative' number in the

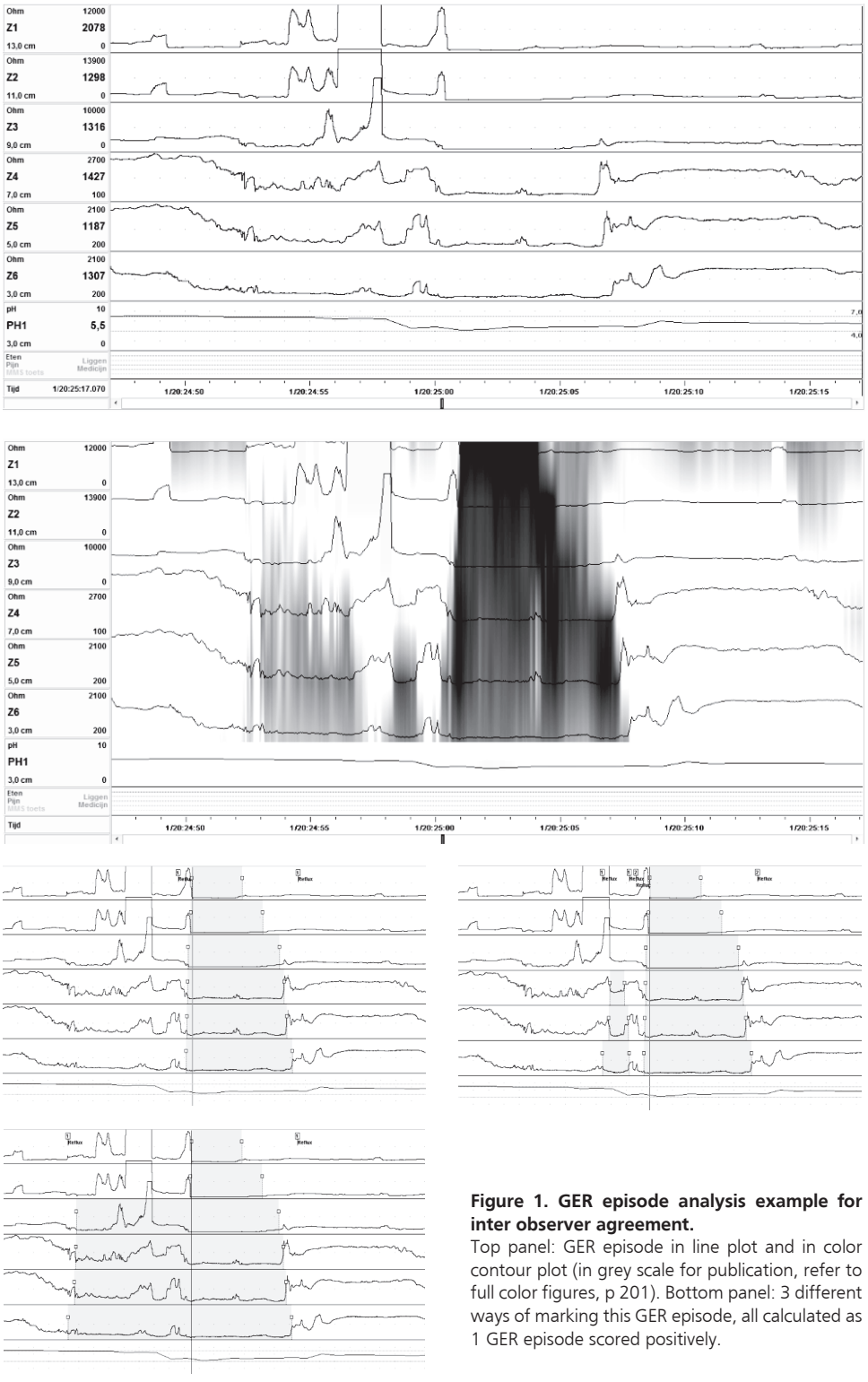

Figure 1. GER episode analysis example for inter observer agreement.
Top panel: GER episode in line plot and in color contour plot (in grey scale for publication, refer to full color figures, p 201). Bottom panel: 3 different ways of marking this GER episode, all calculated as 1 GER episode scored positively.

kappa table, timeframes judged negative in both analyses, was calculated from the GER events that were identified by any of the other observers in the inter observer analysis.

Statistical analysis

Data are presented as median (range) unless otherwise stated. Analysis was performed per tracing and for all tracings combined. Inter and intra observer agreement was assessed using Cohen's kappa (κ). The arbitrary but commonly used scale for kappa values is: 0.0 = no agreement, 0.01 – 0.20 = slight, 0.21-0.40 = fair 0.41-0.60 = moderate, 0.61-0.8 = substantial, 0.81-0.99 = excellent, 1.00 = perfect agreement.The overall κ is calculated as the mean of all κ's combined. Agreement of continuous data was compared using the intra class coefficient. Significance was defined as $p<0.05$. SPSS 17.0 was used for statistical analysis.

RESULTS

Inter observer analysis

A total of 1242 liquid or mixed GER episodes were scored by ≥1 observer in ten 24hr pH-MII tracings. The median number of GER events scored in all tracings per observer was 518 (range 249-922). Of the 1242 GER episodes, 490 (42%) were scored by the majority and 377 (31%) were scored by one observer only.

Mean agreement for recognition of GER episodes between the observers for all tracings combined was moderate (κ = 0.46), agreement between observers is shown in Table 1. The level of experience of the observers did not influence the agreement. Five 'easy' pH-MII tracings and five 'challenging' pH-MII tracings were selected. Agreement between

OBS2	0.62									
OBS3	0.44	0.48								
OBS4	0.69	0.69	0.50							
AA	0.47	0.52	0.31	0.54						
OBS6	0.25	0.30	0.14	0.28	0.17					
OBS7	0.56	0.64	0.39	0.62	0.71	0.29				
OBS8	0.52	0.62	0.37	0.56	0.58	0.24	0.70			
OBS9	0.51	0.50	0.27	0.52	0.36	0.27	0.49	0.46		
OBS10	0.65	0.69	0.47	0.74	0.52	0.25	0.59	0.54	0.48	
OBS11	0.42	0.51	0.50	0.46	0.31	0.12	0.45	0.45	0.30	0.43
	OBS1	**OBS2**	**OBS3**	**OBS4**	**AA**	**OBS6**	**OBS7**	**OBS8**	**OBS9**	**OBS10**

Table 1. Kappa values between all observer pairs.
Number of GER episodes to calculate kappa N=1242 (Liquid and mixed GER). Mean kappa between all observers is 0.46 (moderate agreement).

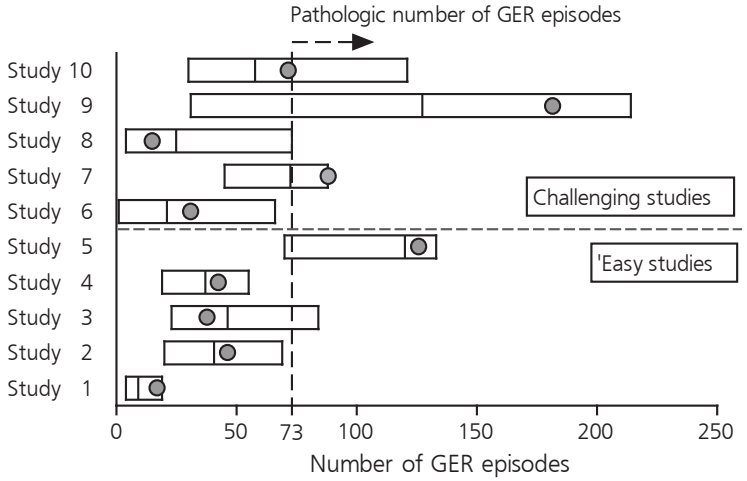

Figure 2. Range number of GER episodes scored per study by all observers.
Study 1-5 represent the 'easy' studies, study 6-10 were more challenging to analyze. Box represents range, bars represent median number of GER episodes scored. The oval represents the number of GER episodes detected by AA. The vertical dotted line represents the 73 GER episodes /24hr cut off for a normal or pathological number of GER episodes.

all observer pairs for the five 'easy' tracings (total GER episodes N=472) is comparable to the agreement for all tracings, κ = 0.50 (moderate agreement).

Numbers of GER episodes detected per 24 hr study by all observers are presented in Figure 2. The range of number of GER episodes detected per study varies from 4-19 in one study (Figure 2, study 1), to 30-240 in another study (Figure 2, study 9). Intra class coefficient for total numbers of GER recognized per study was 0.84 ($p<0.001$). The range of number of GER episodes scored varied less in the five 'easy' studies (Figure 2 study 1-5) represented in better intra class coefficient values of 0.95 ($p<0.001$) compared to 0.8 ($p<0.001$) in the 'challenging' tracings.

In Figure 2 the vertical dotted line represents the 73 GER episodes per 24hrs cut off used in adults to assess a pathological number of GER episodes. Four studies were judged normal by all observers. Six studies cross the line of pathological number of GER episodes (vertical dotted line in Figure 2), however only one study was judged normal by five observers and pathological by six observers (study 7 in Figure 2). In the other studies all observers agreed on either a normal or abnormal study except for one observer. Agreement for judging a study normal or pathological based on the 73 GER episodes cut off is substantial (mean κ = 0.70). For comparison, other cut off values between 40 and 101 GER episodes per 24hrs are presented in Table 2. Agreement between observers is moderate for cut off values below 60 GER episodes per 24hrs and substantial for cut of values between 73 and 101 GER episodes per 24hrs, with the exception of a cut off of 80 GER episodes per 24hrs which shows an excellent agreement in the assessment of a normal or pathological study based on number of GER episodes.

Cut off value	Median kappa	Mean kappa	Agreement
40	0.41	0.45	Moderate
50	0.58	0.56	Moderate
60	0.62	0.62	Substantial
73	0.74	0.7	Substantial
80	0.74	0.74	Substantial
90	1	0.87	Excellent
101	0.74	0.77	Substantial
101 and 73	0.74	0.72	Substantial

Table 2. Median and mean kappa values applying different cut off values for number of GER episodes per 24hr study.
In the last row we used the cut off value of 101 for 2 infants (<1 year of age at time of study) and 73 GER episodes per 24hr for children >1 year.

Gas GER

Several observers did not mark gas episodes because they considered gas GER to be of little importance and more challenging to recognize. Between the four observers that analyzed gas GER and AA (N=5 observers) agreement was poor (mean κ = 0.11 (range -0.24 – 0.22)). In total 394 gas GER episodes were detected in all tracings, however only 63 GER episodes (16%) were identified by the majority of observers. A median number of 106 (range 53-216) gas GER episodes were identified by the observers.

Automated analysis

A total of 490 GER episodes were detected by the observer consensus (majority of observers) and with AA a total of 653 GER events were detected. AA missed 32 (6.5%) GER events that were scored by ≥6 observers and detected 195 (30%) events not scored by the majority of observers. Majority observer consensus and AA showed a substantial agreement, κ = 0.65. For comparison, the overall agreement between different observers and the majority consensus was also substantial (median κ = 0.73 range 0.34 to 0.81). Based on majority consensus, AA sensitivity is 94% and specificity is 74%.
AA identified more gas GER episodes than detected by the observers (216 vs median number of 106 gas GER episodes). Of those 216 episodes, the majority, 124 (57%) were only detected with AA.

Intra observer analysis

Intra observer agreement between GER episodes marked in the first and second analysis is moderate to excellent (Table 3). Intra class coefficient for numbers of GER per study was high in all observers, 0.90 to 0.99 ($p<0.001$). Observers re-analyzed the tracings after three to five months. A longer time between the two analyses corresponded to a lower intra observer agreement.

	First analysis	Second analysis	Kappa	ICC
OBS1	521	595	0.71	0.90
OBS2	469	513	0.85	0.99
OBS3	781	733	0.49	0.90

Table 3. Intra observer analysis.
Number of total GER episodes marked in all tracings combined in the first and second analysis and agreement between 2 analyses.

Personal guidelines

Of the ten observers five normally run AA before they start their manual analysis (however for the analysis in this study the AA was not used). Six observers use the color contour plot regularly. All observers state that a retrograde pattern is the most important factor in the recognition of GER. The majority of observers take a 50% drop in the most distal channel, however not always is the two most distal channels and the raw impedance values into account. There is no consensus between the observers on the accuracy of the current guidelines. Most observers state that they mark GER episodes that do not fulfill the guidelines. The observers felt that the guidelines were inadequate particularly in young infants, tracings with low baseline values and in children with co morbidities (eg esophageal atresia or achalasia).

DISCUSSION

We have demonstrated that the inter- and intra observer agreement for the identification of liquid and mixed GER amongst experts in pediatric GER from seven different groups in the world is moderate based on the kappa statistics. Only 42% of all GER episodes were scored by the majority of observers (≥6 observers scored GER), this is a poor outcome considering the relative experience of the observers. Agreement between observer majority consensus and AA was substantial. Inter observer agreement for the detection of gas GER was poor and it is noteworthy that only four observers considered gas GER an important entity that should be included in the analysis. The variability in terms of total number of liquid and mixed GER episodes detected per study was smaller. The range in numbers of GER episodes detected in 'challenging' pH-MII tracings (e.g. due to low baselines) was larger compared to 'easy' pH-MII tracings (intraclass coefficient 0.80 vs 0.95). When applying a cut off value of <73 GER episodes per 24hrs for normal numbers of GER episodes, agreement between all observers was substantial (mean κ = 0.70). These numbers show that the total number of GER episodes detected and their clinical impact is more consistent between observers than agreement on the level of the detection of individual GER episodes. However a mean κ of 0.70 (substantial agreement) for the determination of a normal or pathological study by experts can be regarded as a poor result if being used to guide clinical decision making.

In this study we analyzed the inter observer agreement on micro level, in terms of detection of specific GER episodes and on macro level, in terms of a positive or negative study and observed a lower agreement between observers on micro level than on macro level (mean $\kappa = 0.46$ and $\kappa = 0.70$ respectively). Others have previously reported on inter and intra observer agreement in the analysis of pH-MII tracings.[10,12,13,16] Two studies report substantial to excellent agreement between observers in terms of a positive or negative study ($\kappa = 0.72$ in one and 0.79 - 0.83 in the other).[13,16] The agreement we observed on macro level is comparable to the first study.[13] There are two explanations for the discordance between the higher agreement observed by Peter et al[16] compared to the moderate agreement in this study. The Cohen's kappa calculation requires a value for 'true negative' counts. Since no gold standard exists in pediatric pH-MII testing, the other studies have chosen to take the number of time windows with no GER events as true negative counts. This results in a high number of true negative counts and therefore positively influences the Cohen's kappa. In our study we calculated all GER episodes scored by one or more observers. When another observer pair did not recognize that episode as GER it was calculated as a 'true negative', allowing more accurate calculation of Cohen's kappa. Furthermore, all previous inter and intra observer variability studies were performed within one group. It is likely that members within one group analyze pH-MII tracings similarly, resulting in higher inter observer agreement.

AA accuracy was analyzed based on majority consensus and showed substantial agreement. AA missed 6.5% of events scored by observer consensus, represented by a high sensitivity of 94%. However, 30% of the GER episodes detected with AA were not detected by majority consensus yielding a lower specificity of 74%. This indicates over detection of liquid and mixed GER episodes by the current AA, as has been shown previously by other authors.[12,17]

For research and clinical purposes, reproducibility of GER detection is highly important. The substantial agreement between AA and majority consensus suggests that the use of AA only instead of manual analysis can be advocated. However, as yet the true impact of AA on clinical outcomes in infants and children remains undefined. Furthermore, the low specificity suggests that AA only may not be refined enough yet for the detection of GER in individual patients. It is important to note that we used MMS software in this study, AA is provided by all software companies and the accuracy of AA may differ between software packages.

Our data show great variability in the detection of gas GER between observers, moreover 6 of ten observers did not consider gas GER of importance for the analysis of pH-impedance tracings. Gas GER is substantially overdetected with AA compared to majority consensus. It has been shown that the inclusion of gas GER improves the yield of symptom associations,[5,8] however the poor agreement between observers compromises the comparability between studies carried out by different groups. Acknowledging the additional yield of gas GER in symptom associations, the poor agreement between majority consensus and AA indicates that a consensus should be reached to define the

criteria for the detection of gas GER and if gas GER should be included for analysis. This consensus should then be implemented in the AA.

Intraobserver agreement was moderate to excellent, and the total number of GER episodes detected was very similar between the first and the second analysis. In our study the observers analyzed the tracings with a 3 - 5 month period in between, with a longer time between the two analyses correlating to a lower kappa value.

A shortcoming of this study was the inability to assess the impact of the inter observer variability on symptom association indices. In a recent paper in adults Hemmink et al showed that 83% of the studies had a concordant symptom association probability (SAP) despite substantial under detection of GER episodes with AA (after removal of overdetected GER).[12] The other 17% of patients were judged to have a positive SAP judged by manual analysis and not with AA. The authors suggested running AA and using that result if the symptom association was positive. If symptom association was negative, they suggested manual analysis of the tracings.

Although guidelines for the analysis of pH-MII tracings exist, the visual interpretation of pH-MII tracings is self-taught and based on what an observer considers pathophysiologically plausible. All observers state that a retrograde pattern recognition and a marked decrease in impedance in the most distal channel are the most important factors in determining retrograde bolus flow. However pattern recognition appears to be highly subjective as only 42% of all GER events were recognized by the majority of observers. Furthermore, most observers state that they mark GER episodes that do not fulfill the guidelines, especially in young infants, in tracings with low baseline values and in children with co morbidities. This is presumably the greatest factor driving the moderate inter observer agreement. The high variability in personal guidelines calls for refining GER detection to ensure accuracy for GER disease detection in the individual patient and reproducibility of research performed by different groups around the world.

In conclusion, we have shown that pH-MII analysis by different observers shows moderate agreement in terms of specific GER episode and that the total numbers of GER episodes per 24hr study are more consistent. The finding that only 42% of all GER episodes was detected by the majority of observers is a poor outcome considering the experience of the observers. These poor results need to be addressed as we conclude from this study that pH-impedance analysis is not uniform enough to compare between centers. AA showed a high sensitivity and a lower specificity compared to observers consensus. In theory AA is favored over manual analysis due to its reproducibility, time effectiveness and accessibility to the wider public. The moderate inter observer agreement, moderate to excellent intra observer agreement, the high AA sensitivity suggests a substantial role for AA. However, AA does not seem specific enough to ensure correct marking of GER episodes in individual infants and children yet. Therefore automated GER detection needs to be refined and tested before it can be advocated for the analysis of pH-MII studies in both a clinical and research setting. A consensus to refine AA needs to be reached in due course to retain confidence to the use of impedance in this setting.

REFERENCE LIST

1. Vandenplas Y, Rudolph CD, Di LC, Hassall E, Liptak G, Mazur L, et al. Pediatric gastroesophageal reflux clinical practice guidelines: joint recommendations of the North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition (NASPGHAN) and the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition (ESPGHAN). *J Pediatr Gastroenterol Nutr.* 2009;49:498-547.
2. Rosen R, Lord C, Nurko S. The sensitivity of multichannel intraluminal impedance and the pH probe in the evaluation of gastroesophageal reflux in children. *Clin Gastroenterol Hepatol.* 2006;4:167-172.
3. Blondeau K, Tack J. Pro: Impedance testing is useful in the management of GERD. *Am J Gastroenterol.* 2009;104:2664-2666.
4. van Wijk MP, Benninga MA, Omari TI. Role of the multichannel intraluminal impedance technique in infants and children. *J Pediatr Gastroenterol Nutr.* 2009;48:2-12.
5. Loots CM, Benninga MA, Davidson GP, Omari TI. Addition of pH-impedance monitoring to standard pH monitoring increases the yield of symptom association analysis in infants and children with gastroesophageal reflux. *J Pediatr.* 2009;154:248-52.
6. Salvatore S, Arrigo S, Luini C, Vandenplas Y. Esophageal impedance in children: symptom-based results. *J Pediatr.* 2010;157:949-954.
7. Condino AA, Sondheimer J, Pan Z, Gralla J, Perry D, O'Connor JA. Evaluation of infantile acid and nonacid gastroesophageal reflux using combined pH monitoring and impedance measurement. *J Pediatr Gastroenterol Nutr.* 2006;42:16-21.
8. Bredenoord AJ, Weusten BL, Timmer R, Conchillo JM, Smout AJ. Addition of esophageal impedance monitoring to pH monitoring increases the yield of symptom association analysis in patients off PPI therapy. *Am J Gastroenterol.* 2006;101:453-459.
9. Sifrim D, Holloway R, Silny J, Tack J, Lerut A, Janssens J. Composition of the postprandial refluxate in patients with gastroesophageal reflux disease. *Am J Gastroenterol.* 2001;96:647-655.
10. Dalby K, Nielsen RG, Markoew S, Kruse-Andersen S, Husby S. Reproducibility of 24-Hour Combined Multiple Intraluminal Impedance (MII) and pH Measurements in Infants and Children. Evaluation of a Diagnostic Procedure for Gastroesophageal Reflux Disease. *Dig Dis Sci.* 2007;52:2159-65.
11. Peter CS, Sprodowski N, Ahlborn V, Wiechers C, Schlaud M, Silny J, et al. Inter- and intraobserver agreement for gastroesophageal reflux detection in infants using multiple intraluminal impedance. *Biol Neonate.* 2004;85:11-14.
12. Hemmink GJ, Bredenoord AJ, Aanen MC, Weusten BL, Timmer R, Smout AJ. Computer analysis of 24-h esophageal impedance signals. *Scand J Gastroenterol.* 2010.
13. Ravi K, DeVault KR, Murray JA, Bouras EP, Francis DL. Inter-observer agreement for multichannel intraluminal impedance-pH testing. *Dis Esophagus.* 2010;23:540-544.
14. Shay S, Tutuian R, Sifrim D, Vela M, Wise J, Balaji N, et al. Twenty-four hour ambulatory simultaneous impedance and pH monitoring: a multicenter report of normal values from 60 healthy volunteers. *Am J Gastroenterol.* 2004;99:1037-1043.
15. Lopez-Alonso M, Moya MJ, Cabo JA, Ribas J, Macias MD, Silny J, et al. Twenty-Four-Hour Esophageal Impedance-pH Monitoring in Healthy Preterm Neonates: Rate and Characteristics of Acid, Weakly Acidic, and Weakly Alkaline Gastroesophageal Reflux. *Pediatrics.* 2006;118:e299-308.
16. Peter CS, Sprodowski N, Ahlborn V, Wiechers C, Schlaud M, Silny J, et al. Inter- and intraobserver agreement for gastroesophageal reflux detection in infants using multiple intraluminal impedance. *Biol Neonate.* 2004;85:11-14.
17. Roman S, Bruley Des Varannes S, Pouderoux P, Chaput U, Mion F, Galmiche JP, et al. Ambulatory 24-h oesophageal impedance-pH recordings: reliability of automatic analysis for gastro-oesophageal reflux assessment. *Neurogastroenterol Motil.* 2006;18:978-986.