Interobserver and Intraobserver Variability in pH-Impedance Analysis between 10 Experts and Automated Analysis

Clara M. Loots, MSc[1], Michiel P. van Wijk, MD, PhD[1], Kathleen Blondeau, PhD[2], Kasper Dalby, MD, PhD[3], Laura Peeters, MSc[1], Rachel Rosen, MD, PhD[4], Silvia Salvatore, MD[5], Tobias G. Wenzl, MD, PhD[6], Yvan Vandenplas, MD[7], Marc A. Benninga, MD[1], and Taher I. Omari, PhD[8]

Objective To determine interobserver and intraobserver variability in pH-impedance interpretation between experts and accuracy of automated analysis (AA).

Study design Ten pediatric 24-hour pH-impedance tracings were analyzed by 10 observers from 7 world groups and with AA. Detection of gastroesophageal reflux (GER) episodes was compared between observers and AA. Intraobserver agreement was assessed in 3 observers after 3 to 5 months.

Results Overall, 1242 liquid and mixed GER events were detected, 490 (42%) were scored by the majority of observers, yielding moderate agreement (Cohen's kappa [κ] = 0.46). Intraclass co-efficient for numbers of GER per study was 0.84 ($P < .001$). AA has 94% sensitivity rate and 74% specificity rate compared with 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 combined pH-multichannel intraluminal impedance analysis in experts is moderate; only 42% of GER episodes were detected by the majority of observers. Detection of total GER numbers is more consistent. Considering these poor outcomes, AA seems favorable compared with manual analysis because of its reproducibility. However, the lower specificity rate suggests the need for refinement of AA before widespread use can be advocated. *(J Pediatr 2012;160:441-6)*.

Combined pH-multichannel intraluminal impedance (pH-MII) has been used increasingly to assess gastroesophageal reflux (GER) in infants, children, and adults, and this technique is now recommended by the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition 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 also.[2-4] In infants and children particularly, pH-MII has been shown to detect more GER than pH-metry alone.[5,6] Detecting GER with 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 interobserver and intraobserver variability. Several papers assessing interobserver and intraobserver 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 in investigators from different groups and between AA and the consensus between observers is unknown.

The aims of this study were to determine interobserver variability in pH-MII interpretation in 10 experts in pH-MII analysis and pediatric GER, determine the accuracy of AA compared with majority observer consensus, and assess intraobserver variability in 3 investigators.

Methods

Patients and Tracings

Tracings from patients with clinical indication for pH-impedance or who participated in research were selected. All primary research protocol were approved by

AA	Automated analysis
GER	Gastroesophageal reflux
κ	Cohen's kappa
MMS	Medical Measurement Systems (Enschede, The Netherlands)
MII	Multichannel intraluminal impedance
pH-MII	Combined pH-multichannel intraluminal impedance

From the [1]Department of Pediatric Gastroenterology and Nutrition, Emma Kinderziekenhuis, Academic Medical Centre, Amsterdam, The Netherlands; [2]Translational Research Center for Gastrointestinal Disorders, University of Leuven, Leuven, Belgium; [3]H.C. Andersen Children's Hospital, Odensee, Denmark; [4]Children's Hospital Boston, Boston, MA; [5]Università dell'Insubria, Varese, Italy; [6]Klinik für Kinder- und Jugendmedizin, Universitätsklinikum der RWTH Aachen, Aachen, Germany; and [7]Universitair Ziekenhuis Brussel, Brussels, Belgium; [8]Women's and Children's Hospital, University of Adelaide, Adelaide, South Australia, Australia

The authors declare no conflicts of interest.

0022-3476/$ - see front matter. 10.1016/j.jpeds.2011.08.017

the ethical board from the Academic Medical Centre, Amsterdam, The Netherlands. Because the 10 tracings were anonymized before distribution and reanalysis, ethical approval was not obtained for the inter and intra observer study. All pH-MII studies were performed in children and infants (median age, 4.5 years; age 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 on the basis of thorax radiography or videofluoroscopy.[1]

Ten 24-hour pH-MII tracings with different characteristics were selected from a research database. Five tracings were considered to be "easy" to analyze because of clear GER patterns with clear retrograde patterns and GER extending high in the esophagus. Five tracings were considered to be more challenging because the GER patterns were less obvious because of low baselines, retrograde patterns during swallowing, and moving/crying artifacts. The 24-hour pH-MII tracings were recorded with the Omega ambulatory system (Medical Measurement Systems [MMS], Enschede, The Netherlands). All tracings had >20 hours of recorded pH-MII measurements. The tracings were randomized and distributed without markers from AA to the 10 observers. Observers had different levels of experience in the analysis of pH-MII tracings, ranging from 6 months to >15 years, having analyzed 100 to >2000 pH-MII tracings.

Interobserver Analysis

Tracings were analyzed by 10 experts in pediatric GER and analysis of pH-MII tracings from 7 groups around the world and with AA (MMS Omega ambulatory Autoscan version 8.17, with standard settings and the option to reduce overdetection selected). Observers were asked to analyze the 10 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. When observers recognized GER in the same timeframe, including the point of nadir impedance, that episode was scored positive for both observers. When 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 for 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 cutoff value of 73 GER episodes in 24 hours[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. Because the value of this cutoff point is arbitrary in the pediatric population, we assessed the impact of different cutoff values ranging from 40 to 101 GER episodes per 24 hours, the cutoff value in preterm infants.[15]

Furthermore, AA was compared with the observers' consensus. The majority consensus was defined as GER episodes scored as positive by any majority (≥6) of observers.

Intraobserver Analysis

Three to 5 months after the first analysis, 3 observers reanalyzed the same tracings for assessment of intraobserver variability. For the purpose of intraobserver analysis, all GER episodes marked in the first and in the second analysis by one observer were compared. When the observer recognized one long GER episode in one analysis and two short GER episodes in the second analysis, one GER episode was scored as positive twice and one GER episode was scored once. To calculate agreement, the "truly negative" number in the kappa table, timeframes judged as negative in both analyses, was calculated from the GER events that were identified by any of the other observers in the interobserver analysis.

Statistical Analysis

Data are presented as median (range) unless otherwise stated. Analysis was performed per tracing and for all tracings combined. Interobserver and intraobserver agreement was assessed by using Cohen's kappa (κ). The arbitrary but commonly used scale for κ values is: 0.0 = 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 0.8 = substantial agreement, 0.81 to 0.99 = excellent agreement, 1.00 = perfect agreement.The overall κ is calculated as the mean of all κs combined. Agreement of continuous data was compared by using the intraclass co-efficient. Significance was defined as a P value <.05. SPSS software version 17.0 (SPSS Inc, Chicago, Illinois) was used for statistical analysis.

Results

Interobserver Analysis

A total of 1242 liquid or mixed GER episodes were scored by one or more observers in 10 24-hour 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 of observers, and 377(31%) were scored by one observer only.

Mean agreement for recognition of GER episodes in the observers for all tracings combined was moderate (κ = 0.46); agreement between observers is shown in **Table I**. The level of experience of the observers did not influence the agreement. Five "easy" pH-MII tracings and

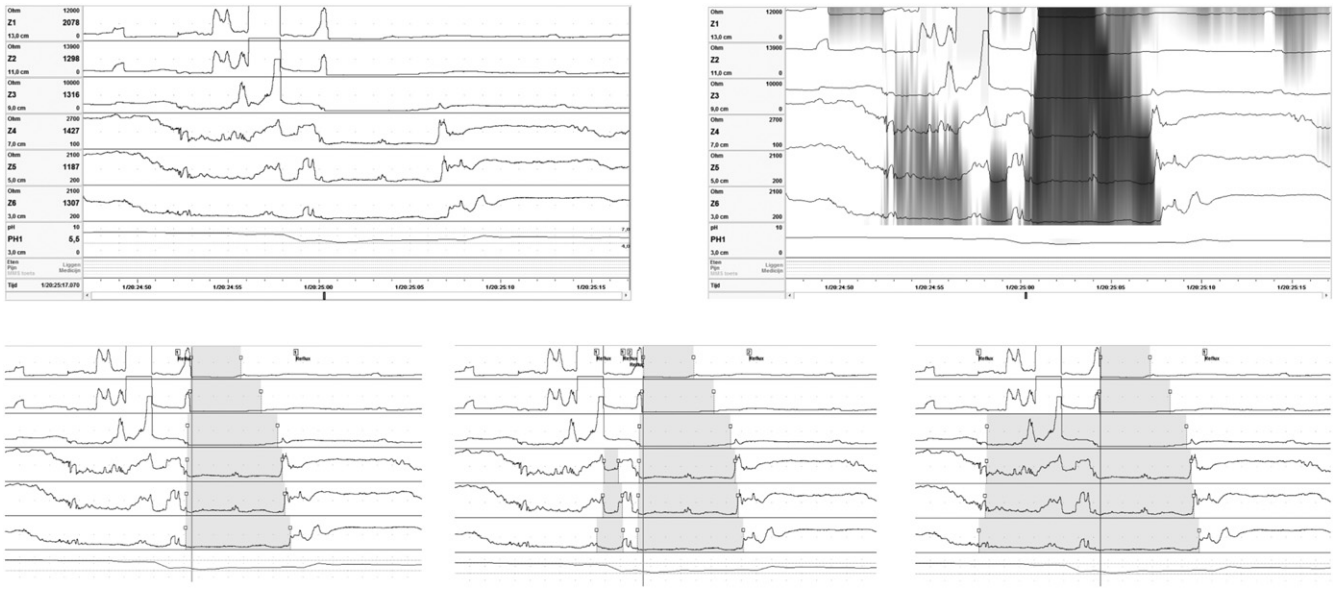

Figure 1. GER episode analysis example for interobserver agreement. **Top panel**, GER episode in line plot and in color contour plot (in grey scale for print publication). **Bottom panel**, 3 different ways of marking this GER episode, all calculated as one GER episode scored as positive.

five "challenging" pH-MII tracings were selected. Agreement in all observer pairs for the 5 "easy" tracings (total GER episodes n = 472) is comparable with the agreement for all tracings, $\kappa = 0.50$ (moderate agreement).

Numbers of GER episodes detected per 24-hour study by all observers are presented in Figure 2. The range of number of GER episodes detected per study varies from 4 to 19 in one study (Figure 2, study 1) to 30 to 240 in another study (Figure 2, study 9). Intraclass coefficient for total numbers of GER recognized per study was 0.84 ($P < .001$). The range of number of GER episodes scored varies less in the 5 "easy" studies (Figure 2, studies 1-5) represented in better intraclass co-efficient values of 0.95 ($P < .001$), compared with 0.8 ($P < .001$) in the "challenging" tracings.

In Figure 2, the vertical dotted line represents the 73 GER episodes per 24-hour cutoff value 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 5 observers and pathological by 6 observers (Figure 2, study 7). 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 on the basis of the 73 GER episodes cutoff value is substantial (mean $\kappa = 0.70$). For comparison, other cutoff values between 40 and 101 GER episodes per 24 hours are presented in Table II (available at www.jpeds.com). Agreement in observers is moderate for cutoff values <60 GER episodes per 24 hours and substantial for cutoff values between 73 and 101 GER episodes per 24 hours, with the exception of a cutoff value of 80 GER episodes per 24 hours, which shows an excellent agreement in the assessment of a normal or pathological study on the basis of number of GER episodes.

Table I. Kappa values in all observer pairs

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

OBS, observer.
Number of GER episodes to calculate kappa = 1242 (Liquid and mixed GER).
Mean kappa between all observers is 0.46 (moderate agreement).

Gas GER

Several observers did not mark gas episodes because they considered gas GER to be of little importance and more challenging to recognize. In the 4 observers who analyzed gas GER and AA (n = 5 observers), agreement was poor (mean $\kappa = 0.11$; 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 gas GER episodes (range, 53-216) were identified by the observers.

Automated Analysis

A total of 490 GER episodes were detected by the observer consensus (most observers), and with AA a total of 653 GER events were detected. AA missed 32 GER events (6.5%) that were scored by ≥6 observers and detected 195

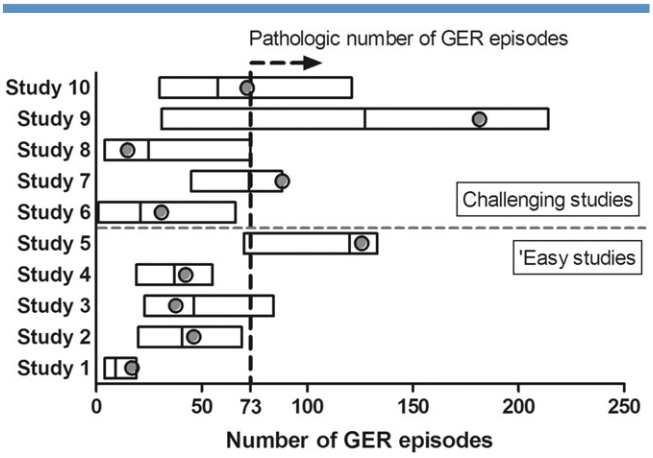

Figure 2. Range number of GER episodes scored per study by all observers. Studies 1 to 5 represent the "easy" studies; studies 6 to 10 were more challenging to analyze. *Box* represents range; *bars* represent median number of GER episodes scored. The *oval* represents the number of GER episodes detected with AA. The *vertical dotted line* represents the 73 GER episodes/24-hour cutoff value for a normal or pathological number of GER episodes.

events (30%) not scored by the majority of observers. Majority observer consensus and AA showed a substantial agreement, $\kappa = 0.65$. For comparison, the overall agreement between different observers and the majority consensus was also substantial (median $\kappa = 0.73$; range, 0.34-0.81). On the basis of majority consensus, the AA sensitivity rate is 94%, and the specificity rate is 74%.

With AA, more gas GER episodes were identified than detected by the observers (216 versus median number of 106 gas GER episodes). Of those 216 episodes, most, 124 (57%), were only detected with AA.

Intraobserver Analysis

Intraobserver agreement in GER episodes marked in the first and second analysis is moderate to excellent (Table III). Intraclass co-efficient for numbers of GER per study was high in all observers, 0.90 to 0.99 ($P < .001$). Observers re-analyzed the tracings after 3 to 5 months. A longer time between the two analyses corresponded to a lower intraobserver agreement.

Personal Guidelines

Of the 10 observers, 5 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. Most observers take a 50% drop in the most distal channel (although not always in the two most distal channels) and the raw impedance values into account. There is no consensus in 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 infants, tracings with low baseline values, and children with co-morbidities (eg, esophageal atresia or achalasia).

Table III. Intraobserver analysis

	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

ICC, intraclass coefficient; *OBS*, observer.
Total number of total GER episodes marked in all tracings combined in the first and second analysis and agreement in the two analyses.

Discussion

We have demonstrated that the interobserver and intraobserver agreement for the identification of liquid and mixed GER in experts in pediatric GER from 7 different groups in the world is moderate on the basis of the kappa statistics. However, 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 is substantial. Interobserver agreement for the detection of gas GER is poor, and only 4 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 is smaller. The range in numbers of GER episodes detected in "challenging" pH-MII tracings (eg, because of low baselines) is larger compared to "easy" pH-MII tracings (intraclass coefficient 0.80 versus 0.95). When applying a cutoff value of <73 GER episodes per 24 hours for normal numbers of GER episodes, agreement between all observers was substantial (mean $\kappa = 0.70$). These numbers show that the total number of GER episodes detected and their clinical impact is more consistent in 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 when being used to guide clinical decision making.

In this study we analyzed the interobserver agreement on micro level, for detection of specific GER episodes, and on macro level, for a positive or negative study, and observed a lower agreement in observers on micro level than on macro level (mean $\kappa = 0.46$ and 0.70, respectively). Other studies have reported on interobserver and intraobserver agreement in the analysis of pH-MII tracings.[10,12,13,16] Two studies report substantial to excellent agreement in observers for 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 with the first study.[13] There are two explanations for the discordance in the higher agreement observed by Peter et al[16] compared with the moderate agreement in this study. The κ calculation requires a value for "true-negative" counts. Because 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 κ. 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 κ. Furthermore, all previous interobserver and intraobserver variability studies were performed within one group. It is likely that members within one group analyze pH-MII tracings similarly, resulting in higher interobserver agreement.

AA accuracy was analyzed on the basis of majority consensus and showed substantial agreement. AA missed 6.5% of events scored by observer consensus, represented by a high sensitivity rate of 94%. However, 30% of the GER episodes detected with AA were not detected with majority consensus, yielding a lower specificity rate of 74%. This indicates overdetection of liquid and mixed GER episodes with the current AA, as has been shown 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, the true impact of AA on clinical outcomes in infants and children remains undefined. Furthermore, the low specificity rate suggests that AA only may not be refined enough yet for the detection of GER in individual patients. We used MMS software in this study. AA is provided by all software companies, and the accuracy of AA may differ in software packages.

Our data show great variability in the detection of gas GER in observers; moreover, 6 of 10 observers did not consider gas GER of importance for the analysis of pH-impedance tracings. Gas GER is substantially overdetected with AA compared with majority consensus. It has been shown that the inclusion of gas GER improves the yield of symptom associations[5,8]; however, the poor agreement in observers compromises the comparability in studies carried out by different groups. Acknowledging this 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 whether 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- to 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 interobserver variability on symptom association indices. In a recent paper in adults, Hemmink et al[12] showed that 83% of the studies had a concordant symptom association probability despite substantial underdetection of GER episodes with AA (after removal of overdetected GER). The other 17% of patients were judged to have a positive symptom association probability as assessed with manual analysis and not with AA. The authors suggested running AA and using that result when 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, because 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 infants, in tracings with low baseline values and in children with co-morbidities. This is presumably the greatest factor driving the moderate interobserver 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.

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 with observer consensus. In theory, AA is favored over manual analysis because of its reproducibility, time effectiveness, and accessibility to the wider public. The moderate interobserver agreement, moderate to excellent intraobserver agreement, and the high AA sensitivity rate 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. ■

Submitted for publication Apr 13, 2011; last revision received Jun 14, 2011; accepted Aug 4, 2011.

Reprint requests: Clara M. Loots, Academisch Medisch Centrum, Kinder Motiliteitscentrum, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. E-mail: c.m.loots@amc.nl

References

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;4:167-72.
3. Blondeau K, Tack J. Pro: Impedance testing is useful in the management of GERD. Am J Gastroenterol 2009;104:2664-6.
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;157:949-54.
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;101:453-9.
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;96:647-55.
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;85:11-4.
12. Hemmink GJ, Bredenoord AJ, Aanen MC, Weusten BL, Timmer R, Smout AJ. Computer analysis of 24-h esophageal impedance signals. Scand J Gastroenterol 2011;46:271-6.
13. Ravi K, DeVault KR, Murray JA, Bouras EP, Francis DL. Inter-observer agreement for multichannel intraluminal impedance-pH testing. Dis Esophagus 2010;23:540-4.
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;99:1037-43.
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;85:11-4.
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;18:978-86.

Table II. Median and mean kappa values applying different cutoff values for number of GER episodes per 24-hour study

Cutoff 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

In the last row, we used the cutoff value of 101 for two infants (<1 year of age at time of study) and 73 GER episodes per 24 hours for children >1 year of age.