Pediatric esophageal motility disorders: studies on (patho)physiology, diagnosis and management
Smits, M.J.

Citation for published version (APA):
An expert panel-based study on recognition of gastroesophageal reflux in difficult esophageal pH-impedance tracings

Marije Smits, Clara Loots, Michiel van Wijk, Albert Bredenoord, Marc Benninga and André Smout

Neurogastroenterol Motil. 2015 Mar 10. [Epub ahead of print]
ABSTRACT

Introduction
Despite existing criteria for scoring gastroesophageal reflux (GER) in esophageal multichannel pH-impedance measurement (pH-I) tracings, inter- and intrarater variability is large and agreement with automated analysis is poor.

Objective
To identify parameters of difficult to analyze pH-I patterns and combine these into a statistical model that can identify GER episodes with an international consensus as gold standard.

Methods
Twenty-one experts from 10 countries were asked to mark GER-presence for adult and pediatric pH-I patterns in an online pre-assessment. During a consensus meeting, experts voted on patterns not reaching majority consensus (>70% agreement). Agreement was calculated between raters, between consensus and individual raters and between consensus and software-generated automated analysis. With 8 selected parameters, multiple logistic regression analysis was performed to describe an algorithm sensitive and specific for detection of GER.

Key results
Majority consensus was reached for 35/79 episodes in the online pre-assessment (intrarater K=0.332). Mean agreement between pre-assessment scores and final consensus was moderate (K=0.466). Combining 8 pH-I parameters did not result in a statistically significant model able to identify presence of GER. Recognizing a pattern as retrograde is the best indicator of GER, with 100% sensitivity and 81% specificity with expert consensus as gold standard.

Conclusion & inferences
Agreement between experts scoring difficult impedance patterns for presence or absence of GER is poor. Combining several characteristics into a statistical model did not improve diagnostic accuracy. Only the parameter ‘retrograde propagation pattern’ is an indicator of GER in difficult pH-I patterns.
INTRODUCTION

Gastroesophageal reflux (GER) is the passive movement of gastric contents into the esophagus and is referred to as GER disease (GERD) when causing troublesome symptoms or complications. When symptoms are refractory to medical therapy or complications are suspected, additional diagnostic tests are warranted. Combined esophageal multichannel pH-impedance (pH-I) measurement has replaced conventional pH-metry for research purposes and has become a well-accepted diagnostic tool in clinical practice.

Impedance is defined as the resistance that a circuit presents to an alternating electrical current. An impedance catheter allows measurement of bolus flow (antegrade [swallow] and retrograde [GER]) by making use of six circular electrode pairs, or channels, placed longitudinally along the catheter. The impedance baseline is mainly dependent on the conductivity of the esophageal wall.

It decreases with the passing of an ionized liquid bolus (drop in resistance) and increases with gas boluses (rise in resistance), allowing discrimination between liquid, gaseous or mixed boluses. In combination with the pH sensor, liquid bolus episodes can be further categorized in acid (pH<4), weakly acid (4>— pH>7) or non-acid GER (pH >— 7).

It is generally accepted to classify liquid GER as a drop of >50% of baseline impedance signal in the distal >2 channels, moving in retrograde direction. Similarly, gas GER is defined as a retrograde rise of impedance to >3000 Ω/s in ≥2 channels. Mixed GER is a combination of patterns meeting both liquid and gas GER criteria. Software algorithms for automated analysis, based on these criteria, are available but have been shown less sensitive and specific than consensus agreement of >1 rater. Current guidelines for the analysis of a pH-I tracings suggest running automated analysis first, followed by visual screening of the study to add or remove GER episodes. This screening is based on pattern recognition rather than applying strict criteria.

Although the criteria for GER in pH-I tracings seem relatively clear-cut, several studies show there is considerable inter- and intrarater variability and only a minority of GER is detected by all raters. Automated analysis has the advantage of no intrarater variability, but specificity is low (74%). Nevertheless, both manual readers and software packages are well capable of identifying GER episodes that clearly meet impedance criteria mentioned above and are accompanied by a pH drop below 4. In practice, only a minority of GER episodes is represented by pH-I patterns that are so easily recognized. Poor inter- and intrarater agreement figures are further explained because analysis depends heavily on measurement conditions. A previous article discussing various techniques to measure GER only briefly addressed the possibility of false positive GER on impedance and pH monitoring. Despite the consensus statement that pH-I measurement should be able to detect ≥ 90% of all GER episodes, it was not addressed how achieve this and how to overcome difficulties when scoring pH-I patterns.

In this study, we aimed to identify parameters of difficult to analyze pH-I patterns and combine these into a statistical model that can identify GER episodes with an international consensus as gold standard.
METHODS

In this author-initiated study, a consensus meeting was organized in Amsterdam, involving international experts in the field of gastroesophageal reflux disease. Prior to this meeting, one author (CL) selected eighty-three patterns of possible GER episodes selected from pediatric tracings with low agreement levels in a previously conducted interrater variability study.\(^\text{11}\) In addition, 43 patterns from adult clinical studies were selected, all of which originating from tracings that raised discussion during plenary review with several experienced gastroenterologists. All pH-I patterns (N=126) were reviewed by two other authors (AB and MS), who excluded patterns that they considered a duplication of other episodes, ultimately resulting in 88 selected patterns.

All studies were recorded using Ohmega ISFET disposable catheters with 6 impedance channels and 1 distal pH channel and an Ohmega ambulatory recording device (MMS, Enschede, The Netherlands). MMS provided support to combine all 88 episodes into one study file with sufficient time preceding and following the episode for the software to function properly. Experts from eight different centers worldwide (see list of contributing experts at the end of this article) were approached to participate in the online pre-assessment. All experts had extensive experience levels (\(>500\) analyzed pH-I tracings) and were accustomed to working with the software used for this study. After their consent, the experts received one file containing all 88 episodes. They were asked to review the study file using MMS analysis software version 8.17 and mark presence (yes/no) and type (liquid/mixed/gas) of GER events. They were allowed to use the software to its full extent, including the use of enhancement of patterns, pH values and color plot setting. Results of the online pre-assessment were collected and analyzed prior to the consensus meeting. Automated analysis was run on the original, complete study files from which the impedance patterns were selected using MMS analysis software version 8.23.

**Categories of pH-impedance patterns**

Based on the online pre-assessment, the 88 pH-I patterns were grouped into 7 categories: unclear flow direction (antegrade or retrograde), flow following a swallow in uncertain direction (GER or poor clearance), impedance drops during low baseline impedance not reaching 50\% of baseline, short impedance drops (pH-only drops, short-lived, impedance drop in the most distal channel only), mixed impedance patterns and patterns reaching high impedance. Six retrograde patterns easily recognizable as GER ('clear GER patterns') and one antegrade pattern easily recognizable as a swallow ('clear swallow' pattern) were included for reference purposes. For all patterns, impedance as well as pH tracings were included.

**Consensus meeting**

A consensus meeting was organized in Amsterdam, The Netherlands, two months after all experts completed the online pre-assessment. During the consensus meeting, experts voted on the pH-I patterns that did not reach majority consensus in the pre-assessment. Majority consensus was
arbitrarily defined as ≥ 70% outcome agreement. pH-I patterns that reached majority consensus during pre-assessment were briefly presented and all experts were asked if they agreed with these results. pH-I patterns without majority consensus during the pre-assessment were presented using images of the line plot and the color plot. There was a possibility to use the analysis software to its full extent for more detailed information upon request. Experts voted anonymously with a wireless voting system. Outcomes were displayed immediately after voting. The decision that consensus was reached was based on a >50% agreement. The chairmen were decisive in case of an equal vote (50%-50%). Both episodes agreed on in the online pre-assessment and the patterns voted on during the consensus meeting will be discussed in this report.

**Pattern recognition**

Based on qualitative analysis of all impedance patterns, the authors determined parameters that could potentially distinguish between ‘GER present’ patterns and ‘GER absent’ patterns, as determined by the consensus group (see Figure 1). Consequently, two authors scored all impedance patterns for these parameters.

**Statistical analysis**

SPSS 20.0 (IBM statistics) was used to perform all statistical procedures. For all tests, differences were considered statistically significant when p<0.05. Only liquid-containing impedance patterns were incorporated in statistical analysis, because these are currently incorporated in clinical GER assessment and automatic analysis.

**Overall agreement**

Weighed Cohen's kappa was used to calculate inter- and intrarater agreement between experts in the online pre-assessment, between individual experts and the consensus meeting outcomes, and between the consensus outcome, individual experts and automated analysis. The commonly used scale for kappa values was used with a kappa score of <0 indicating no agreement, 0-0.2 slight, 0.2-0.4 fair, 0.4-0.6 moderate, 0.6-0.8 substantial, 0.8-0.99 excellent and 1.0 perfect agreement.

**Statistic model development**

Results of the analysis of the eight different parameters scored for each impedance pattern were compared based on consensus outcome (‘GER present’ or ‘GER absent’). Data are shown as median values (interquartile range) and compared using Mann-Whitney U tests. Multiple logistic regression analysis was performed to describe an algorithm with optimal sensitivity and specificity to predict GER for difficult impedance patterns. To do so, univariate associations between consensus outcome and all indicated parameters (co-variates) were assessed using single logistic regression analysis. Consequently, multiple logistic regression analyses were performed with the co-variates that were found significant in univariate analysis, using stepwise forward and backward procedures with 0.1 and 0.05 alpha levels of entry and removal respectively. Collinearity was checked for every significant main factor in the multiple logistic models using the variance
inflation factor (VIF). No collinearity was assumed when VIF <3. Model selection was based on -2 log likelihood ratio test statistics. The fit of the final model was assessed using Nagelkerke $R^2$. Based on sample size, a maximum of three co-variates were allowed in the model.

RESULTS

Twenty-one experts from 10 different international centers completed the online pre-assessment and 16 experts (from eight different centers) eventually attended the consensus meeting. CL and AS chaired the meeting and refrained from voting.

Online pre-assessment

Of all 88 patterns, 8 were unanimously scored as gas only and 1 as a swallow pattern and therefore discarded from further analysis (Figure 2). After the online pre-assessment, the remaining 79 pH-I patterns were divided into subgroups as follows (Figure 1): Unclear flow direction patterns (n=27), Flow after swallow patterns (n=11), Low baseline patterns (n=17), patterns with limited impedance change (n=13), mixed liquid-mixed impedance patterns (n=5) and ‘easy’ patterns (n=6). Forty-four patterns with ≥70% agreement were discussed during the consensus meeting.

Agreement during consensus meeting

All experts agreed with the majority consensus of the online pre-assessment prior to voting. In 26 of the 44 pH-I patterns voted on during consensus meeting, majority agreement (≥70%) was reached. In 16 (36%) patterns only threshold consensus (>50%-70 outcome agreement) was reached (Figure 2 and 3). For 2 patterns, the chairmen decided GER was present after an indecisive vote (50-50).

Interrater variability

Majority agreement per pH-I pattern category during online pre-assessment is graphically displayed in Figure 3. Mean overall interrater agreement between the 21 raters regarding analysis of the 79 episodes during online pre-assessment was fair ($\kappa=0.332$) and ranged from no agreement ($\kappa=-0.033$) to substantial agreement ($\kappa=0.631$). Interrater agreement between the mean participant score and final consensus was moderate ($\kappa=0.466$). Agreement between individual participants and consensus ranged from fair ($\kappa=0.243$) to substantial ($\kappa=0.692$).

Agreement with Automatic Analysis

Interrater agreement between the automated analysis and consensus outcome was slight ($\kappa=0.129$). Mean agreement between the automated analysis and individual experts was fair ($\kappa=0.219$), ranging from no agreement ($\kappa=-0.156$) to moderate agreement ($\kappa=0.571$).
Modeling pattern recognition

We identified eight parameters that could potentially differentiate the 'GER present' patterns from 'GER absent' patterns. These parameters are listed and graphically shown in Figure 4. Four of eight parameters (duration of impedance drop, Δ impedance from baseline, duration of impedance fall and impedance nadir) were quantified for both the most distal (sixth) and second distal (fifth) channel. The mean values of all but two of the continuous parameters differed significantly between 'GER present' and 'GER absent' patterns (Table 1). However, the ranges of the continuous parameters overlapped in such a way that no cut-off values discriminating 'GER present' from 'GER absent' could be determined per parameter.

Combining different parameters with multiple regression analysis did not result in a satisfactory model to describe difficult GER impedance patterns. Including the variable 'antegrade or retrograde' in combination with the covariate 'nadir impedance in channel 6' resulted in a significant model with the highest explained variance of all models tested (Nagelkerke R² 0.786). No significant collinearity was found between the independent variables, although for 'nadir impedance in channel 6,' VIF values only just remained under the threshold of 3 (VIF=2.998). However, when adding the co-variante 'antegrade or retrograde' to the model, the additional values of other co-variates diminished. We judged this model as invalid to predict the occurrence of GER in a difficult impedance pattern because of the overriding influence of the co-variante 'antegrade or retrograde' and possible multi-collinearity of the co-variante 'nadir impedance in channel 6.'

The variable 'antegrade or retrograde' alone proved highly sensitive. All 48 episodes judged as 'GER present' in the consensus outcome were classified with a retrograde propagation pattern (sensitivity: 100% (95% CI 92.5 - 100.0%), while from the 31 episodes scored as 'GER absent', 6 were found to have a retrograde pattern (specificity: 81% [95%CI 62.5 – 92.5%]). Positive and negative predictive values were 89% and 100% respectively. When judging a difficult impedance pattern, retrograde flow in (at least) the two distal channels thus should be leading. Additional factors associated with GER presence in this study (long duration impedance drop in channels 5 and 6, Δ impedance compared to baseline in channels 5 and 6, Δ pH from baseline, long duration pH<4, impedance nadir in channels 5 and 6 (Table 1) can help identify these patterns as GER or not. Retrograde propagation patterns that were nonetheless judged as 'GER absent' by experts are characterized by (Figure 5): 1. A retrograde pattern not reaching above the three most distal channels 2. Fluctuating and low baseline impedance patterns preceding the pattern 3. Concomitant swallow patterns 4. An absent or limited pH drop less than 1s in duration, or a baseline pH<4.
DISCUSSION

In this article, the process of finding criteria for better manual and automatic analysis of difficult impedance patterns is described, including a consensus meeting with experts from centers all over the world. For the pH-I patterns included, the level of agreement between experts in an online pre-assessment and with the automated analysis is poor and no satisfactory statistical model could be identified based on several characteristics of the impedance patterns. Of all parameters, only the presence of a retrograde propagation pattern on pH-I is an indicator of GER, according to consensus outcome.

pH-I measurements are favored above pH-metry in pediatric as well as adult practice, mainly because adding impedance allows for the detection of weakly acid and non-acid GER as well as for a better association between GER and symptoms. However, many studies have shown great inter- and intrarater variability, possibly caused by the many difficult to classify pH-I episodes in a 24h tracing. Currently, there is no examination available to reliably determine whether these subtle ‘GER-like’ impedance patterns indeed represent a retrograde flow of gastric contents into the esophagus. Therefore, our aim was to reach consensus on these difficult patterns to be “GER” or “no GER”, using international experts’ opinion as gold standard.

It is well known that the presence of acid and proximal GER can cause symptoms, but also non-acid and distal GER are recognized as associated with symptoms in all age groups. Sifrim et al. demonstrated that subtle but significant differences exist in GER impedance patterns between GERD patients and healthy controls and that the severity of GERD is reflected in impedance patterns. These subtle differences between GERD and no GERD might be represented in difficult impedance patterns as described in this study. Criteria on how to judge these patterns are lacking and therefore these patterns are often overlooked, narrowing the diagnostic yield of pH-I measurements. There is also a possibility that overestimation of GER occurs, when patterns which are not GER are scored too eagerly. Nevertheless, firm conclusions are commonly drawn from pH-I measurements both in clinical practice and for research purposes. Therefore, international consensus on how to judge difficult impedance patterns is indispensable, especially for the group of patients for whom the outcome of the pH-I could lead to anti-reflux surgery or other (invasive) treatment.

Consensus outcome matched poorly with automated analysis. This raises the question whether automated analysis should be used to evaluate complex tracings (e.g. complicated by body movement or low baselines). Ideally, a visual-manual approach by several experienced raters with consensus agreement on each GER episode should be used for such tracings. In daily practice, this is not feasible and clinical results are generally heavily based upon automated analysis. In an attempt to more accurately detect GER episodes on impedance tracings, we identified eight parameters (Figure 4) based on consensus outcome. Some of these were closely related to known parameters currently assessed in pH-I tracings. Four of the eight parameters (duration of impedance drop, Δ impedance...
from baseline, duration of impedance fall and impedance nadir) were quantified for both the most distal (sixth) and second distal (fifth) channel. This distinction was made because impedance drops in the most distal channel alone might lead to over-detection of GER when the catheter is dipping in the gastric cavity due to esophageal shortening during peristalsis. Using these parameters, we showed that no satisfying model could be constructed, confirming that experts rely on visual pattern recognition, which is not easily described by one or more parameters. The only exception is the presence of retrograde flow. Thus, to ensure optimal assessment of pH-I tracings, we suggest that the automated analysis is refined in such a way that it detects all retrograde patterns and asks for human confirmation when it does not fulfill all classic criteria. Ideally, these patterns are then visually assessed by ≥1 rater, before drawing conclusions on diagnosis and treatment options. Especially in research, such a uniform, partially visual-manual analysis of all retrograde patterns could improve reliability and reproducibility of study outcomes.

The fact that more pediatric than adult impedance tracings were selected, could be considered a limitation of the study. It is unknown how physiological differences between children and adults influence impedance patterns. However, it is unlikely that this had a direct effect on the results, as differences in catheter size between children and adults have been shown not to influence pH-I analysis. Moreover, purpose of this study was to evaluate difficult patterns. It is known that pH-I tracings are influenced by body movement and position and pediatric measurements are indeed often complicated by body movement or crying.

Consistent with earlier data, this study shows that the interrater variability in interpretation of pH-I measurements is large and consensus outcome matches poorly with automated analysis. Using consensus of a large group of international experts, several characteristics of pH-I tracings can be identified that correlate with the presence of GER episodes. However, combining these additional parameters into a statistical model did not improve diagnostic accuracy. Of all parameters tested, the best indicator of GER, according to consensus outcome, is presence of a retrograde propagation pattern on pH impedance tracings.
ACKNOWLEDGEMENTS

We would like to thank all the experts on pediatric and adult gastroesophageal reflux research that participated in the expert panel for the consensus meeting:

T.I. Omari; Flinders Medical Science and Technology, Flinders University, South Australia. K. Blondeau; Centre for Gastroenterological Research, Catholic University of Leuven, Belgium. K. Dalby; H.C. Andersen Children's Hospital, Odensee, Denmark. N. Rommel; Departments of Ear-Nose-Throat, Head and Neck Surgery, and Pediatric Gastroenterology & Nutrition, University Hospitals, Leuven, Belgium. R. Rosen; Center for Motility and Functional Gastrointestinal Disorders, Children's Hospital Boston, Harvard Medical School, Boston, United States of America. S. Salvatore; Universita dell’Insubria, Varese, Italy. D. Sifrim; The Wingate Institute of Neurogastroenterology, London, United Kingdom. R. Tutuian; Division of Gastroenterology/Hepatology, Medical University of South Carolina, United States of America. Y. Vandenplas; Universitair Kinderziekenhuis Brussel, Brussels, Belgium. T. Wenzl; Klinik für Kinder- und Jugendmedizin, Universitätsklinikum der RWTH Aachen, Aachen, Germany. R.J. Van der Pol; Department of Pediatric Gastroenterology and Nutrition, Emma Children's Hospital AMC, Amsterdam, the Netherlands. B. Kessing, W. Rohof, P. Weijenborg and B. Van Rhijn; Department of Gastroenterology and Hepatology AMC, Amsterdam, the Netherlands.
Figure 1. Examples of pH-I patterns of each category. For each category, an example of a pattern with the consensus outcome ‘GER present’ and ‘GER absent’ is shown. Please note that for the patterns in the categories ‘mixed GER patterns’ and ‘easy GER patterns’ all were classified as ‘GER present’, thus no ‘GER absent’ examples can be provided.
Figure 2. Consensus meeting design and outcome.
Figure 3. Agreement for the 79 impedance patterns during online pre-assessment and during consensus meeting, prior to the consensus meeting scored by the 21 raters. Per category, the number of episodes for which majority consensus was reached (i.e. >=70% of raters scored either GER or no GER, the number of episodes where threshold consensus (>50-70% outcome agreement) was reached) and those for which majority consensus was not reached are shown.
Figure 4. Example of an easy pH-I pattern to illustrate criteria for the judgment of more difficult impedance patterns.

Figure 5. Example of a pH-I pattern, scored as ‘GER absent’ despite having a retrograde propagation pattern (arrow).
<table>
<thead>
<tr>
<th>Variable</th>
<th>GER present</th>
<th>GER absent</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrograde (n)</td>
<td>48</td>
<td>6</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Antegrade (n)</td>
<td>0</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Impedance channels involved (n)</td>
<td>4 (4-6)</td>
<td>6 (4-6)</td>
<td>0.115</td>
</tr>
<tr>
<td>Duration impedance drop channel 6 (s)</td>
<td>18.5 (11-39)</td>
<td>11 (5-15)</td>
<td>0.001*</td>
</tr>
<tr>
<td>Duration impedance drop channel 5 (s)</td>
<td>17 (9-36.5)</td>
<td>10 (6-14)</td>
<td>0.003*</td>
</tr>
<tr>
<td>Δ impedance compared to baseline Ch6 (%)</td>
<td>80 (65-85)</td>
<td>60 (40-80)</td>
<td>0.018*</td>
</tr>
<tr>
<td>Δ impedance compared to baseline Ch5 (%)</td>
<td>70 (50-80)</td>
<td>50 (30-70)</td>
<td>0.004*</td>
</tr>
<tr>
<td>Duration impedance fall channel 6 (s)</td>
<td>1 (1-2.75)</td>
<td>1 (0.5-1)</td>
<td>0.347</td>
</tr>
<tr>
<td>Duration impedance fall channel 5 (s)</td>
<td>1 (0.89-2.0)</td>
<td>1 (0.5-1)</td>
<td>0.190</td>
</tr>
<tr>
<td>Δ pH from baseline</td>
<td>1.42 (2.6-4)</td>
<td>0.7 (0.1-3.1)</td>
<td>0.001*</td>
</tr>
<tr>
<td>Duration pH&lt;4 (s)</td>
<td>31.5 (0.8925)</td>
<td>0 (0.1-3)</td>
<td>0.001*</td>
</tr>
<tr>
<td>Nadir impedance channel 6 (Ω)</td>
<td>377 (231-567)</td>
<td>625 (442-921)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Nadir impedance channel 5 (Ω)</td>
<td>502 (305-865)</td>
<td>761 (543-1149)</td>
<td>0.005*</td>
</tr>
</tbody>
</table>

Table 1. Median (IQR) and p-values for scored parameters in every impedance pattern. Significant p-values are flagged (*).
REFERENCES


Chapter 6

Contributors’ statement

Marije Smits: Took part in organizing and executing the study, reviewed and selected relevant impedance patterns for the study, collected and analyzed study data, drafted the initial manuscript, and approved the final manuscript as submitted.

Clara Loots: conceptualized, designed and organized the study, collected study data, co-chaired the consensus meeting, reviewed and revised the manuscript and approved the final manuscript as submitted.

Andre Smout: co-chaired the consensus meeting, reviewed and revised the manuscript and approved the final manuscript as submitted.

Arjan Bredenoord: reviewed and selected relevant impedance patterns for the study, reviewed and revised the manuscript and approved the final manuscript as submitted.

Michiel van Wijk and Marc Benninga: reviewed and revised the manuscript, and approved the final manuscript as submitted.