Gastroesophageal reflux scintigraphy: interpretation methods and inter-reader agreement

Murat Tuncel, Pınar Özgen Kıratlı, Tamer Aksoy, Murat Fani Bozkurt

Ankara, Turkey

Background: Scintigraphic imaging is a useful screening tool for patients with suspected gastroesophageal reflux. New scintigraphic interpretation methods have recently been introduced. This study was undertaken to evaluate the efficiency of various scintigraphic interpretation methods in the detection of gastroesophageal reflux and to measure their influence on inter-reader agreement.

Methods: Scintigraphic images of 49 children with suspected gastroesophageal reflux were interpreted by three different methods: visual interpretation, time activity curves, and condensed images. The readings were performed by three specialists and a resident. The discordant results were resolved by a consensus reading done together by all interpreters based on the three different methods. The gastroesophageal refluxes were grouped according to their number, location and intensity.

Results: Gastroesophageal reflux scintigraphy revealed 22 patients with negative results and 27 with positive results. The sensitivity, positive predictive value and specificity for each of the three specialists *vs.* the resident were 96%, 96% and 81% *vs.* 96%; 93%, 90% and 96% *vs.* 81%; and 90%, 86%, and 95% *vs.* 73%, respectively. The mean inter-observer reproducibility (κ value) was 0.910 for visual interpretation, 0.652 for time activity curves and 0.789 for condensed images. Twenty-seven percent of the results were discordant and most of these refluxes were of low grade (92%), low intensity (77%) and localization in the distal esophagus (54%).

Conclusion: Gastroesophageal scintigraphy is a useful tool for detecting patients with suspected reflux, and visual interpretation is better than the other two methods in terms of accuracy and inter-observer reproducibility.

World J Pediatr 2011;7(3):245-249

Key words: gastroesophageal reflux;
inter-observer variability;
scintigraphy;
time activity curve

Introduction

Gastroesophageal reflux (GER) is a common, self-limited process in infants and children that usually resolves at 12 to 18 months of age. Clinical management of GER includes conservative treatment, thickened feedings, positional therapy and parental reassurance.[1] On the other hand, GER disease (GERD) is a less common, more serious pathological process that is manifested by poor weight gain, signs of esophagitis and persistent respiratory symptoms that usually warrant medical management and diagnostic evaluation.[2] Diagnostic studies are indicated only in cases of doubtful diagnosis or manifestations outside the digestive system. Esophageal 24-hour pH probe monitoring, radionuclide scintigraphy, multichannel intraluminal impedance and ultrasonography have gained wide-spread acceptance.[3-5] GER scintigraphy (GERS) is a validated diagnostic modality with a sensitivity of 75%-100%.[6] Several factors affect the accuracy of GERS, especially the technique (anterior *vs.* posterior imaging, acquisition time per frame and special manoeuvres), and the experience of the interpreter.[7-11] Several interpretation methods such as time activity curve (TAC) and condensed image have been introduced to aid visual image analysis.[10]

This study was undertaken to evaluate the efficiency of various scintigraphic interpretation methods in the detection of GER and to measure their influence on inter-reader agreement.

Methods

Forty-nine patients (29 girls, 20 boys, mean age:

Author Affiliations: Department of Nuclear Medicine, Hacettepe University Faculty of Medicine, Sihhiye, Ankara, Turkey (Tuncel M, Kıratlı PÖ, Aksoy T, Bozkurt MF)

Corresponding Author: Murat Tuncel, MD, Department of Nuclear Medicine, Hacettepe University Faculty of Medicine, Sihhiye, Ankara 06110, Turkey (Tel: +90 536 213 03 41; Fax: +90 312 309 35 08; Email: muratmtx@yahoo.com)

doi: 10.1007/s12519-011-0322-4

8±3 years) who underwent GERS were included in this retrospective study. The patients were selected randomly from the database with no information about their clinical history or test results.

All patients were fasted for 2 to 4 hours prior to their imaging examination. After oral administration of 300 μCi (111 MBq) 99mTc-labelled colloid mixed in milk or orange juice, the patients were placed in a supine position, and dynamic images were acquired anteriorly over the abdomen with a gamma camera (Siemens ECAM, USA) at 16 sec/frame, with a total of 167 frames (45 minutes) in the 64 × 64 matrix.

Scintigraphic images were interpreted with three different methods using an Xeleris workstation (GE Healthcare, USA): (1) visual interpretation (VI), (2) TAC, and (3) condensed image (CI). Three nuclear medicine specialists (P1, P2 and P3) and one resident (third year) (P4) performed the interpretations independently. The interpreters were not told at which clinic the patient examinations were performed or the decisions were made by other interpreters. The interpretations of the three different methods were performed separately to prevent inter-method influence. The discordant results between the interpreters were resolved by a consensus reading done together using all three methods. The consensus reading was accepted as the gold standard for individual comparisons.

The detected GERs were grouped according to the number of episodes (grade 0: no reflux, grade 1: 1-3 episodes, grade 2: ≥4 episodes), location (distal, middle or proximal esophagus) and intensity (low, moderate or high) (Fig. 1).

Visual interpretation was performed by evaluating the 167 frames (16 sec/frame) of each patient with a linear scale of different intensities. Any pathological activity that corresponded to the esophagus was reported as positive for GER.

TACs were generated by placing regions of interest (ROI) over the esophagus. The results of TAC method was interpreted as positive with high peaks twice above the baseline activity as previously defined.[10]

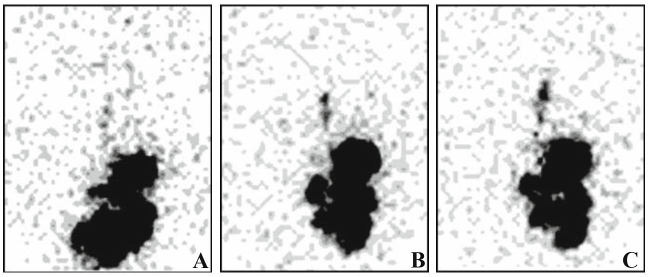

Fig. 1. The grades of gastroesophageal reflux intensity. **A:** low; **B:** moderate; **C:** high.

CIs were created by the Xeleris software; the program allows for the creation of CIs that summarize all of the acquired data into one image. Any significant activity above the baseline was interpreted as a reflux.

Statistical analysis

The correlations between the findings were calculated using the Spearman's correlation coefficient (SCC). The SCC values were evaluated as 0-0.25 for no or a very weak correlation, 0.25-0.5 for a weak to moderate correlation, 0.50-0.75 for a good correlation, and 0.75-1.0 for a very good correlation. The mean kappa (κ) values were calculated for intraobserver reproducibility. SPSS 15.0 software was used for statistical analysis.

Results

Comparison between interpreters and the consensus reading

There were 22 patients with negative (n) and 27 patients with positive (p) GERS as reported by the consensus evaluation. The results per interpreter were listed as n/p: physician 1 (P1) 21/28, physician 2 (P2) 20/29, physician 3 (P3) 26/23, and physician 4 (P4) 17/32.

The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy of each interpreter compared to the consensus reading are shown in Table 1. The PPV and specificity of the specialists were better than those of the resident (P4), but P3 had a lower sensitivity than the other interpreters.

The mean inter-observer reproducibility (κ values) was 0.910 for VI, 0.652 for TAC and 0.789 for CI when the results were interpreted as either positive or negative. When the GERS findings were scored according to the number of episodes, the κ values were 0.873 for VI, 0.623 for TAC, and 0.436 for CI.

Comparison between interpretation methods and the consensus reading

The correlation between the consensus reading and the interpretation methods is illustrated in Table 2. VI correlated better with the consensus reading than did the TAC and CI methods (Fig. 2). CI analysis was better correlated with P1, P2 and P3 than the TAC method. The correlation between the consensus reading and the CI method reached the level of VI for P3 (SCC: 0.767 and 0.750). There were 11 patients who had, at least by one interpreter, grade 0 reflux according to the VI method. Among these 11 patients, 10 were found to be positive by the TAC method and 4 were positive by the CI method. However, according to the consensus reading, only 2 of the 11 patients were said to have

GER. In these two patients, the TAC and CI methods found GER, but in the other 9 patients, these methods also lead to false-positive reflux interpretations.

When the true positive refluxes determined by the consensus reading were used, the cut-off value for peak/background was 1.4 for the TAC method.

Discordant GERS results among the readers

In 13 patients, at least one physician had discordant results with the others. As evaluated by the consensus reading, most of the discordant results included low grade (grade 1: 92%[12/13], grade 2: 8%[1/13]), location in the distal esophagus (distal: 54%[7/13], middle:15% [2/13]) and a low intensity (low: 77%[10/13], moderate: 15 [2/13], high: 8%[1/13]).

Table 1. The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of each interpreter (%)

Physician	Sensitivity	Specificity	PPV	NPV	Accuracy
1	96	90	93	95	94
2	96	86	90	95	92
3	81	95	96	81	88
4	96	73	81	94	86

PPV: positive predictive value; NPV: negative predictive value.

Table 2. The corelations between interpreters were evaluated according to the different interpretation methods

Physician	VI and consensus	TAC and consensus	CI and consensus
1	0.876*	0.511*	0.572*
2	0.837*	-0.010	0.213
3	0.767*	0.138	0.750*
4	0.721*	0.466*	0.466*

*: Correlation is significant when *P*<0.01 (2-tailed). VI: visual interpretation; TAC: time activity curve; CI: condensed image.

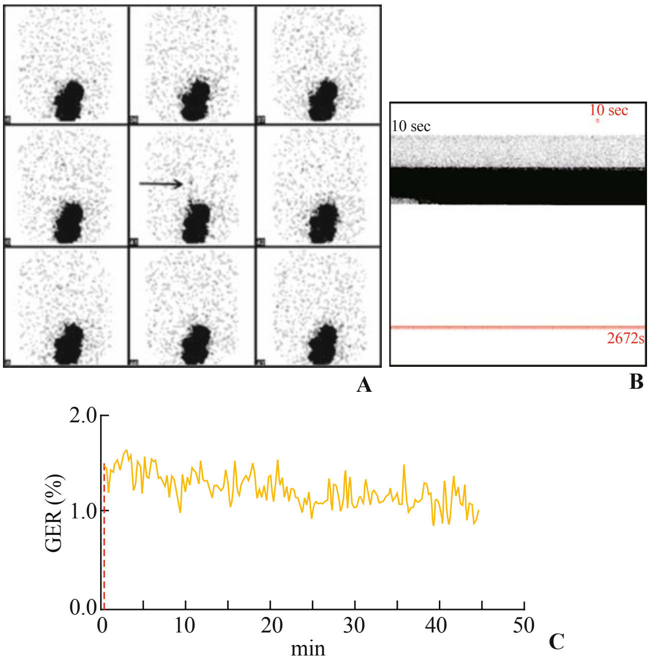

Fig. 2. An 8-year-old patient with gatroesophageal reflux detected by visual interpretation (black arrow) (**A**) but unrecognizable with condensed image (**B**) and time activity curve (**C**). GER: gastroesophageal reflux.

Discussion

The diagnostic modalities for the detection of GERD include upper gastrointestinal barium fluoroscopy, esophageal 24-hour pH probe monitoring, GERS, multichannel intraluminal impedance and ultrasonography.[2] Ambulatory 24-hour pH monitoring appears to be a gold diagnostic criterion in many studies.[12-14] This method can detect refluxes within 24 hours, and the nocturnal episodes of reflux provides a main advantage over scintigraphy. However, 24-hour pH monitoring is not ideal for routine clinical use because it is invasive, available only in tertiary care settings, and ill-equipped to detect alkaline reflux and acidic reflux.[14] Ultrasonography is a relatively new technique for detecting GER. It is non-invasive, relatively cheap and able to provide anatomical information related to GERD. However, this technique is operator dependent and has a lower sensitivity compared to 24-hour pH monitoring. The use of contrast with ultrasonography improves the sensitivity, but at an additional cost. On the other hand, GERS has been shown to be comparable with 24-hour pH-monitoring and ultrasonography, and GERS provides additional information about gastric emptying, aspiration, and abnormal esophageal contraction.[15] Unlike pH monitoring, GERS can also detect non-acid reflux, which is the predominant type in 16% of the children with GER.[16-19] The radiation dose from GERS is estimated to be in a range of 0.04-0.06 mSv, which is markedly lower than the annual environmental dose of 2.5-3 mSv.[20,21]

Although GERS is a rather straightforward technique to interpret, there is considerable variance in its sensitivity depending on the protocol used (e.g., the composition of the solution and acquisition parameters) and the interpreter (due to different experience levels).[5-8,19] The protocol introduced by Maurer et al[22] is preferred in many clinical studies including the present study, and the refluxes are graded according to the number of episodes, location and intensity. The interpreters in this study consisted of three specialists, all with more than 3 years of experience in nuclear medicine, and a resident in the last year of training. In this study, the sensitivity, specificity, PPV, NPV and accuracy of each interpreter were compared to those of the consensus reading, showing that the PPV and specificity improved as the interpreter's experience increased. In 26.5% of the patients, at least one physician had discordant results with the other interpreters, but those were the patients with a reflux localized in the distal esophagus with a low grade and intensity.

Using the optimal intensity and carefully evaluating each frame are crucial to avoiding false negative interpretations. In this study, most of the missed refluxes (77%) had a low intensity. The intensity of the reflux is primarily determined by the radioactivity that passes through the esophagus and the acquisition time per frame ratio. Seymour et al[7] reported that as the time/frame ratio decreases, the sensitivity increases. Increasing the time/frame ratio decreases the number of frames interpreted and eases the physicians' workload, but a longer time/frame may cause dilution of the reflux incidence, which may lead to a decrease in the sensitivity.[19] Although our acquisition rate was within the suggested limits,[22] a decrease in time/frame might further reduce the number of discordant cases.

The location of the reflux is also critical. Scatter artefacts due to high gastric uptake may appear to be a reflux to less experienced physicians (Fig. 3). However, an experienced physician may miss a possible reflux because it is considered to be a scatter artefact. Careful observation of the esophageal track activity and dynamic cine views may help solve this problem. Still, this problem causes discordant readings in distal refluxes, as shown in this study (54%). All of the false positive cases reported by P4 were distal, which explains the lower specificity of the resident and supports our observation.

VI method correlated better with the consensus than the TAC or CI methods. There were 11 patients who had, at least by one interpreter, grade 0 reflux according to the VI method. Among these 11 patients, 10 were found to be positive by TAC and 4 were positive by the CI method. Among these 11 patients, 2 were diagnosed with GER. Thus, in these patients the methods other than VI helped to avoid missing GER. However, in 9 of the 11 patients, the TAC and CI methods led to false-positive reflux interpretations. This discrepancy could be attributive to the physicians being more accustomed to VI because it is the routine interpretation method. The TAC and CI methods are relatively new techniques and lack standardization. For example, the cut-off value for a positive reflux in TAC was found to be 1.4 in our study when calculated from the retrospective reflux episodes, which is lower than 2.0 reported by Caglar et al.[8] TAC is prone to movement artefacts that may lead to a decrease or increase in curve peaks when the patient moves out of the ROI. CI provides a good overview of the whole study in one image, which presents a good outline of the study. Unfortunately, there is no standard in interpretation; the only positivity criterion is an uptake noticeably greater than the background, which is subjective.

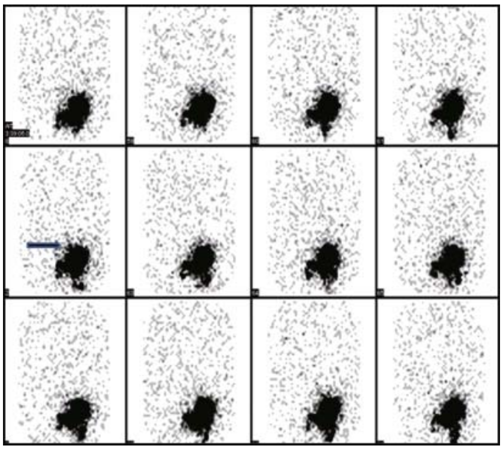

Fig. 3. A 9-year-old boy with intractable cough and weight loss underwent gatroesophageal reflux scintigraphy. The scintigraphic findings were interpreted as distal gatroesophageal reflux by P4 (black arrow) but consensus reading was negative. The false positive finding was considered to be due to a misinterpretation of scatter as a reflux.

Caglar et al[8] found good inter-observer reproducibility by the VI and TAC methods (κ=0.7065 for VI and 0.737 for TAC). In our study, the inter-observer reproducibility was κ=0.910 for VI and κ=0.652 for TAC. Additionally, CI had better reproducibility than TAC, with a κ=0.789. The primary difference between the two studies is the qualifications of interpreters (3 experienced and 1 last-year resident *vs.* 1 experienced physician with a resident in the second year of training).

The mean inter-observer reproducibility (κ) was 0.910 for VI, 0.652 for TAC, and 0.789 for CI when the results were interpreted as either positive or negative. VI showed a higher correlation with the consensus than TAC or CI, which are relatively new techniques without standardization regarding inter-observer reproducibility. When the GERS findings were scored according to the number of episodes, the κ values were 0.873 for VI, 0.623 for TAC and 0.436 for CI. This decrease in reproducibility also shows additional variance in the definition of episode grades.

This study has several limitations. Because of the retrospective nature of the study, it was impossible to include the complete clinical data of the patients. Although it was not an aim, the results of GERS were not compared with those of the gold standard methods in this field. Such comparison has been discussed in the literatures.[23]

In conclusion, GERS has been validated as a tool for the evaluation of patients with suspected GER. VI is better than the CI and TAC methods in terms of accuracy and inter-observer reproducibility. The intensity, location of reflux and number of episodes are the major factors affecting the final interpretation.

Funding: None.
Ethical approval: Not needed.
Competing interest: None declared.
Contributors: Tuncel M wrote the first draft of this paper. All authors contributed to the intellectual content and approved the final version. Kıratlı PÖ is the guarantor.

References

1 Berquist WE. Gastroesophageal reflux in children: a clinical review. Pediatr Ann 1982;11:135-142.
2 Hillemeier AC. Gastroesophageal reflux. Diagnostic and therapeutic approaches. Pediatr Clin North Am 1996;43:197-212.
3 Tsou VM, Bishop PR. Gastroesophageal reflux in children. Otolaryngol Clin North Am 1998;31:419-434.
4 Arasu TS, Wyllie R, Fitzgerald JF, Franken EA, Siddiqui AR, Lehman GA, et al. Gastroesophageal reflux in infants and children comparative accuracy of diagnostic methods. J Pediatr 1980;96:798-803.
5 Reyhan M, Yapar AF, Aydin M, Sukan A. Gastroesophageal scintigraphy in children: a comparison of posterior and anterior imaging. Ann Nucl Med 2005;19:17-21.
6 Braga FJ, De Miranda JR, Arbex MA, Haddad J, Zuolo Ferro S, de Oliveira RB, et al. A physiological manoeuvre to improve the positivity of the gastro-oesophageal reflux scintigraphic test. Nucl Med Commun 2001;22:521-524.
7 Seymour JC, West JH, Drane WE. Sequential ten-second acquisitions for detection of gastroesophageal reflux. J Nucl Med 1993;34:658-660.
8 Caglar M, Volkan B, Alpar R. Reliability of radionuclide gastroesophageal reflux studies using visual and time-activity curve analysis: inter-observer and intra-observer variation and description of minimum detectable reflux. Nucl Med Commun 2003;24:421-428.
9 Orenstein SR, Orenstein DM, Whitington PF. Gastroesophageal reflux causing stridor. Chest 1983;84:301-302.
10 Klotz SD, Moeller RK. Hiatal hernia and intractable bronchial asthma. Ann Allergy 1971;29:325-328.
11 Rode H, Millar AJ, Brown RA, Cywes S. Reflux strictures of the esophagus in children. J Pediatr Surg 1992;27:462-465.
12 Madan K, Ahuja V, Gupta SD, Bal C, Kapoor A, Sharma MP. Impact of 24-h esophageal pH monitoring on the diagnosis of gastroesophageal reflux disease: defining the gold standard. J Gastroenterol Hepatol 2005;20:30-37.
13 Argon M, Duygun U, Daglioz G, Omür O, Demir E, Aydogdu S. Relationship between gastric emptying and gastroesophageal reflux in infants and children. Clin Nucl Med 2006;31:262-265.
14 Malthaner RA, Newman KD, Parry R, Duffy LF, Randolph JG. Alkaline gastroesophageal reflux in infants and children. J Pediatr Surg 1991;26:986-991.
15 Piepsz A. Recent advances in pediatric nuclear medicine. Semin Nucl Med 1995;25:165-182.
16 Heyman S. Gastroesophageal reflux, oesophageal transit, gastric emptying and pulmonary aspiration. In: Treves ST, eds. Pediatric Nuclear Medicine, 2nd ed. New York, NY: Springer-Verlag, 1995: 430-452.
17 Heyman S. Pediatric gastrointestinal motility studies. Semin Nucl Med 1995;25:339-347.
18 Tolia V, Calhoun JA, Kuhns LR, Kauffman RE. Lack of correlation between extended pH monitoring and scintigraphy in the evaluation of infants with gastroesophageal reflux. J Lab Clin Med 1990;115:559-563.
19 Fisher RS, Malmud LS, Roberts GS, Lobis IF. Gastroesophageal (GE) scintiscanning to detect and quantitate GE reflux. Gastroenterology 1976;70:301-308.
20 Castronovo FP Jr. Gastroesophageal scintiscanning in a pediatric population: dosimetry. J Nucl Med 1986;27:1212-1214.
21 Hughes JS, Watson SJ, Jones AL, Oatway WB. Review of the radiation exposure of the UK population. J Radiol Prot 2005;25:493-496.
22 Maurer AH, Parkman HP. Update on gastrointestinal scintigraphy. Semin Nucl Med 2006;36:110-118.
23 Diaz DM, Winter HS, Colletti RB, Ferry GD, Rudolph CD, Czinn SJ, et al. Knowledge, attitudes and practice styles of North American pediatricians regarding gastroesophageal reflux disease. J Pediatr Gastroenterol Nutr 2007;45:56-64.

Received September 2, 2010
Accepted after revision March 17, 2011