Inter-Observer Variability of Radiologists and Gynecologists in Hysterosalpingogram Evaluation

Histerosalpingogram Değerlendirmede Radyolog ve Jinekologlar Arasındaki Gözlemciler Arası Değişkenlik

Ayşe GÜLER OKYAY,[a]
Gökalp ÖNER,[b]
Neşe ÇÖLÇİMEN,[c]
Arzu TURAN,[d]
Fatma BEYAZAL ÇELİKER[e]

[a]Department of Obstetrics and Gynecology, Mustafa Kemal University Tayfur Ata Sökmen Faculty of Medicine, Hatay
[b]Department of Obstetrics and Gynecology, Muğla University Faculty of Medicine, Muğla
[c]Department of Embryology, Yüzüncü Yıl University Faculty of Medicine,
[d]Clinic of Radiology, Özel Lokman Hekim Hospital, Van
[e]Department of Radiology, Recep Tayyip Erdoğan University Faculty of Medicine, Rize

Geliş Tarihi/*Received:* 02.07.2015
Kabul Tarihi/*Accepted:* 04.12.2015

Yazışma Adresi/*Correspondence:*
Ayşe GÜLER OKYAY
Mustafa Kemal University Tayfur Ata Sökmen Faculty of Medicine, Department of Obstetrics and Gynecology, Hatay, TÜRKİYE/TURKEY
doctorayseguler@yahoo.com.tr

doi: 10.5336/gynobstet.2015-47072

ABSTRACT Objective: Reading of hysterosalpingography (HSG) films is important for management of patients. The aim of this study was to evaluate and compare HSG interpretation of radiologists and clinicians. **Material and Methods:** Two clinicians and 2 radiologists, who were 35-43 years old and were fulfilled 5 years in their speciality, evaluated 116 hysterosalpingograms. HSG pictures were viewed at computer monitor and the observers were asked to evaluate them within a standard framework consisted of several questions for diagnosis of uterine and tubal disease. The consistency of each individual reader, the reliability of detecting specific abnormalities, and the consistency of clinicians compared with radiologists was measured. **Results:** There were statistically significant differences in the consistency of interpretations for contour of uterine cavity, uterine deviation, and uterine filling defect ($p<0.05$). Evaluation of uterine anomalies with HSG were similar between clinicians and radiologists ($p>0.05$). Although ratios of hydrosalphinx were similar between clinicians and radiologists ($p>0.05$), there were differences for the evaluation of bilateral tubal contrast falling ($p<0.05$). **Conclusion:** Gynecologists have more inter-observer variability than radiologists in hysterosalpingography evaluation. However, both clinicians and radiologists were compatible within themselves. Compatibility of radiologists was higher than that of clinicians. Better designed studies are needed in order to confirm the variability of HSG reports and to answer the question of "who should read HSGs of infertile women?".

Key Words: Hysterosalpingography; infertility

ÖZET Amaç: Histerosalpingografi filmlerinin doğru okunması hastaların yönetimi için önemlidir. Bu çalışmanın amacı radyologlarla HSG klinisyenlerin HSG yorumlamalarını karşılaştırmaktır. **Gereç ve Yöntemler:** Yaşları 35-43 arasında, kendi uzmanlık dallarında 5 yılını doldurmuş 2 klinisyen ve 2 radyolog 116 adet HSG filmini değerlendirdi. HSG filmleri bilgisayar monitöründen görüntülenerek gözlemcilere çeşitli sorulardan oluşan standart bir soru formuna göre uterin ve tubal hastalıları filmleri değerlendirmeleri istendi. Filmleri okuyan her bir gözlemcinin tutarlılığı, spesifik anormallikleri belirleyebilmedeki güvenilirliği ve klinisyenlerle radyologların kendi aralarındaki tutarlılıkları değerlendirildi. **Bulgular:** Uterin kavite, uterin deviasyon ve dolma defekti konusunda gözlemcilerin tutarlılığı istatistiksel olarak anlamlı derecede birbirinden farklı bulundu ($p<0.05$). Uterin anomalilerin değerlendirmesinde radyologlar ve klinisyenler arasında istatistiksel olarak anlamlı fark saptanmadı ($p>0.05$). Grupların hidrosalpinks yorumları her ne kadar benzer bulunsa da, bilateral tubal kontrast madde geçişi konusunda radyologlar ve klinisyenler arasında arasında fark bulundu ($p<0.05$). **Sonuç:** Jinekologlar arasındaki gözlemciler arası değişkenlik radyologlar arasındakinden daha fazlaydı. Fakat hem klinisyenler hem de radyologlar kendi içlerinde birbirleriyle uyumluydular. Radyologlar arası uyum klinisyenler arasındakinden daha yüksekti. HSG raporlarındaki değişkenliği doğrulamak ve "infertil kadınların HSG filmlerini kim okumalı?" sorusuna cevap bulmak için daha iyi planlanmış çalışmalara ihtiyaç vardır.

Anahtar Kelimeler: Histerosalpingografi; infertilite

Turkiye Klinikleri J Gynecol Obst 2016;26(1):18-22

Hysterosalpingography (HSG) is still a commonly used investigation in the evaluation of the female genital tract and the main indication for HSG is infertility. HSG is still the gold standard to show tubal damage in infertile women, and can be helpful in evaluating uterine cavity abnormalities. Tubal damage, the most important cause of infertility, has extrinsic (pelvic surgery, endometriosis) and intrinsic (salpingitis, isthmic nodosa) components and pelvic inflammatory diseases can cause tubal damage. In a meta-analysis by Swart et al., the estimate of sensitivity and specificity of HSG in detecting tubal patency were 0.65 and 0.83, respectively. HSG was found to be unreliable in diagnosing peritubal adhesions, with sensitivity below 50% (range 13%-83%).[1]

The other cause of infertility is uterine cavity abnormalities. About 10% of subfertile women have uterine cavity abnormalities and uterine cavity findings have reported as many as in 50% of recurrent implantation failure.[2] In the differential diagnosis of intrauterine filling defects by HSG includes polyps, endometrial hyperplasia, sub-mucosal fibroids, intrauterine adhesions and septa. HSG is considered to have a high sensitivity (60-98%) but low specificity (15-80%) in detecting uterine cavity abnormalities.[3]

HSG is crucial during investigations of infertility of women. However what is more important than HSG itself is its' comment or reading of the films. After showing tubal damage or uterine cavity abnormality by HSG as the cause of infertility, HSG helps to decide operational techniques (laparoscopy or hysteroscopy) that patients will undergo. However, inter-observer and intra-observer variability in reading and diagnosis of reproductive tract disease may also affect the interpretation of HSG results.[4] The results of HSG are pivotal to evaluate the infertile women because additional surgical attempts may be needed according to these results. In literature, limited published study has been found to examine the interpretation disparities or similarities of the radiologists and clinicians about HSG. The aim of this study was to evaluate the interpretation of HSG by radiologists and clinicians comparing with their results.

MATERIAL AND METHODS

The study was conducted in a private hospital in Van in June 2012. Computer registration system was scanned and 116 hysterosalpingograms (HSG) of infertile women performed during the past 1 year were determined. HSG pictures were evaluated by 4 observers who were 35-43 years old and were fulfilled 5 years in their speciality. Two of the observers were clinicians in the department of obstetrics and gynecology from other hospitals and the other 2 were radiologists. Clinicians were fulfilled 5 and 7 years and radiologists were 7 and 8 years in their speciality. HSG pictures were viewed at computer monitor and the observers were asked to evaluate them within a standard framework consisted of several questions for diagnosis of uterine and tubal disease (Table 1). Each observer evaluated the HSG films at different times and answered the questions. The protocol was approved by the Ethics Committee of the Faculty of Medicine at Yuzuncu Yil University.

After the completion of HSG readings, forms were collected from each observer and data was transported to the MedCalc 12.0 computer programme and SPSS for Windows 15.0 (Statistical Package for Social Sciences) for statistical analysis. The evaluations of the observers were coded as 0 (not present or normal) and 1 (present or abnormal). Then, the numbers were added up and the results gave us agreements as 0 (not present or normal and agreement of clinicians or radiologists), 1 (disagreement of clinicians or radiologists between themselves), 2 (present or abnormal and agreement of clinicians or radiologists). Thus, it might give us to observe and consider the differences both between radiologists and clinicians. To assess the relation between the answers McNemar test was used. To assess interobserver agreement for categorical variables, the kappa (κ) statistic was used. Finally, comparison of proportions were analyzed with MedCalc 12.0 and differences were shown. For statistical significance, p value was considered as ≤ 0.05.

RESULTS

Interpretations of 116 HSGs were evaluated and differences of clinicians and radiologists were

TABLE 1: Questions replied by the observers during evaluation of hysterosalpingography films.

Questions	0	1
1. Contour of uterine cavity?	regular	irregular
2. T-shaped uterus?	not present	present
3. Arcuate uterus?	not present	present
4. Uterus didelphis?	not present	present
5. Contrast material filling of the right fallopian tube?	not present	present
6. Passage of the contrast material to the peritone on the right side?	not present	present
7. Hydrosalphinx on the right?	not present	present
8. Contrast material filling of the left fallopian tube?	not present	present
9. Passage of the contrast material to the peritone on the left side?	not present	present
10. Hydrosalphinx on the left?	not present	present
11. Uterine deviation?	not present	present
12. Filling defect with contrast material in the uterine cavity?	not present	present

shown in Table 2. While there were statistically significant difference between the clinicians in interpreting uterine contour, tubal filling on the right, peritoneal passage on both sides, uterine deviation and uterine filling defect (the questions numbered 1, 5, 6, 9, 11, and 12) (McNemar $p<0.05$), there was not significant difference about the questions numbered 2, 3, 4, 7, 8 and 10 (McNemar $p>0.05$).

There was not significant difference between the interpretations of radiologists except the presence of arcuate uterus (question 3) (McNemar $p>0.05$).

Clinicians were significantly compatible with each other in answering all questions except the question 1 and 7 (Kappa $p<0.05$). Radiologists were significantly compatible with each other in all questions (Kappa $p<0.05$).

The highest number of comment differences and disagreements between clinicians and radiologists were revealed in the contour of uterine cavity and uterine deviation. Clinicians annotated that 16 of 116 (14%) patients had normal contour of uterine cavity; however radiologists annotated that 54 of 116 (47%) women had normal contour of uterine cavity and there was a statistically significant difference between the interpretations ($p<0.0001$). There was significantly higher discordance between the clinicians (42%) than that between the radiologists (24%) for the comment about contour of uterine cavity. The difference between the disagreement rates of clinicians and radiologists was also statistically significant ($p=0.0055$). Similarly, the clinicians claimed that there was statistically significantly higher number of uterine deviations present (52% vs. 21%, $p<0.0001$), also discordance was significantly higher between the clinicians than that between the radiologists (26% vs. 9%, $p=0.0012$). Comments about uterine anomalies such as T-shaped, arcuate uterus, and didelphys were similar between clinicians and radiologists ($p>0.05$). Although the ratios of hydrosalphinx were similar between the groups, there were significant differences between comments for bilateral tubal contrast passage to the periton. Concerning uterine filling defect, while discordance was higher within the clinicians (28% vs. 15%, $p=0.0244$), concordance on the presence of uterine filling defect was higher between the radiologists than that between the clinicians (22% vs. 10%, $p= 0.0207$).

DISCUSSION

HSG has been an important and first line diagnostic tool for evaluation of the uterine cavity, tubal patency and tubal disease in female fertility investigation. Results of HSG also greatly influences subsequent management. It is important to note that the performance and analysis of HSG is not restricted to fertility specialists. Often the test is per-

TABLE 2: Evaluations and diagnosis of clinicians and radiologists for hysterosalpingography.

			n (%)			
Reader	**0 (not present or normal)**	**P value**	**1 (disagreement)**	**P value**	**2 (present or abnormal)**	**P value**
1. Contour of uterine cavity						
Clinicians	51 (44%)	0.00253	49 (42%)	0.0055	16 (14%)	<0.0001
Radiologists	34 (29%)		28 (24%)		54 (47%)	
2. T-shaped ?						
Clinicians	101 (87%)	0.069	12 (10%)	0.063	3 (3%)	0.814
Radiologists	109(94%)		3 (3%)		4 (4%)	
3. Arcuate Uterus ?						
Clinicians	94 (81%)	0.057	14 (12%)	0.958	8 (7%)	0.056
Radiologists	80 (69%)		13 (11%)		23 (20%)	
4. Uterus didelphis ?						
Clinicians	112 (96%)	0.061	3 (3%)	0.064	1 (1%)	0.971
Radiologists	115 (99%)		0		1 (1%)	
5. Tubal filling on the right						
Clinicians	5 (4%)	0.803	9 (8%)	0.059	102 (88%)	0.061
Radiologists	3 (3%)		6 (5%)		107 (92%)	
6. Passage to the periton on the right ?						
Clinicians	7 (6%)	0.078	15 (13%)	0.0034	94 (81%)	0.0449
Radiologists	9 (7%)		2 (2%)		105 (91%)	
7. Hydrosalpinx on the right ?						
Clinicians	111 (96%)	0.064	5 (4%)	0.075	0	0.058
Radiologists	104 (90%)		10 (8%)		2 (2%)	
8. Tubal filling on the left ?						
Clinicians	5 (4%)	0.872	6 (5%)	0.792	105 (91%)	0.064
Radiologists	4 (3%)		8 (7%)		102 (90%)	
9. Passage to the periton on the left ?						
Clinicians	7 (6%)	0.756	24 (21%)	0.0039	85 (73%)	0.022
Radiologists	9 (7%)		9 (7%)		98 (86%)	
10. Hydrosalpinx on the left ?						
Clinicians	108 (93%)	0.931	6 (5%)	0.987	2 (2%)	0.849
Radiologists	106 (92%)		6 (5%)		4 (3%)	
11. Uterine deviation ?						
Clinicians	25 (22%)	< 0.0001	30 (26%)	0.0012	61 (52%)	< 0.0001
Radiologists	81 (70%)		11 (9%)		24 (21%)	
12. Uterine filling defect ?						
Clinicians	72 (62%)	0.786	32 (28%)	0.0244	12 (10%)	0.0207
Radiologists	73 (63%)		18 (15%)		25 (22%)	

formed and interpreted by radiologists. It is unknown whether there is greater or less variability in the interpretation of this test among radiologists compared with clinicians. Only one study in the literature was designed to assess interpretation of clinicians and radiologists for detecting abnormalities on HSG films.[5]

The purposes of this study were firstly to determine inter-observer variability of clinicians and radiologists separately and secondly comparison of clinicians as group with radiologists group for HSG interpretation. For inter-observer variability, difference between the answers of two observers in each group was determined. Also, compatibility of

the observers or the level of agreement in reading the films was evaluated. Afterwards, two groups were further compared for difference and concordance of their answers.

Renbaum et al. found that inter-reader reliability was high in the detection of normal uterine contour, normal tubal patency, and uterine filling defect and lower for the detection of a hydrosalpinx.[5] They found that inter-reader reliability was high in the detection of normal uterine contour, normal tubal patency, and uterine filling defect and lower for the detection of a hydrosalpinx. Similarly, in our study, this inter-observer reliability was high for determining uterine anomaly, contrast passage to the peritoneal cavity, uterine deviation, and uterine filling defect within clinicians and low for uterine contour and detection of hydrosalpinx. However, inter-observer reliability was high in general between radiologists and they were more consistent than clinicians.

The strongest agreements were those for readings of a normal uterus, uterine anomaly, and normal tubes as reported in literature.[1]

The third goal of this study was to compare the readings of clinicians with those of radiologists. Results of our study showed that radiologists were more compatible with each other in HSG interpretation than clinicians. According to the reports of clinicians, lower number of patients had normal contour of uterine cavity comparing with radiologists' reports and there was a statistically significant difference between the interpretations (14% vs. 47%, $p<0.0001$). Additionally, uterine deviation rates were significantly higher in radiologists comparing with clinicians (70% vs. 22%, $p<0.0001$). Clinicians reported that patients of uterine filling defect was lower in number and there was a statistically significant difference comparing with radiologists (10% vs. 22%, $p=0.0207$). Evaluations of tubes and uterine anomalies were usually similar. It might be due to that diagnosis of hydrosalphinx and uterine anomalies were usually clear and HSG has higher sensitivity and specificity to evaluate these anomalies. In the study of Renbaum et al., comments of clinicians and radiologists were generally similar.[5] This might be due to the low number of patients studied. However our study revealed that there might be some variability in the interpretations of clinicians and radiologists. Our findings might be explained by higher number of patients.

In conclusion; gynecologists have more inter-observer variability than radiologists in hysterosalpingography evaluation. However, both clinicians and radiologists were compatible within themselves. Compatibility of radiologists was higher than that of clinicians. The current study has been conducted in a prospective way, and each of the four observers was blinded to each reports as well as the identity and clinical history of the patient. Better designed studies are needed in order to confirm the variability of HSG reports and to answer the question of "who should read HSGs of infertile women?".

REFERENCES

1. Swart, P, Mol BW, van der Veen F, van Beurden M, Redekop WK, Bossuyt PM. The accuracy of hysterosalpingography in the diagnosis of tubal pathology: a meta-analysis. Fertil Steril 1995;64(3):486-91.
2. Brown SE, Coddington CC, Schnorr J, Toner JP, GibbonsW, Oehninger S. Evaluation of outpatient hysteroscopy, saline infusion hysterosonography, and hysterosalpingography in infertile women: a prospective, randomized study. Fertil Steril 2000;74(5):1029-34.
3. Schankath AC, Fasching N, Urech-Ruh C, Hohl MK, Kubik-Huch RA. Hysterosalpingography in the workup of female infertility: indications, technique and diagnostic findings. Insights Imaging 2012;3(5):475-83.
4. Glastein IZ, Sleeper LA, Lavy Y, Simon A, Adoni A, Palti Z, et al. Observer variability in the diagnosis and management of the hysterosalpingogram. Fertil Steril 1997;67(2):233-7.
5. Renbaum L, Ufberg D, Sammel M, Zhou L, Jabara S, Barnhart K. Reliability of clinicians versus radiologists for detecting abnormalities on hysterosalpingogram films. Fertil Steril 2002;78(3):614-8.