Tubal subfertility and ectopic pregnancy. Evaluating the effectiveness of diagnostic tests
Mol, B.W.J.

Citation for published version (APA):
2. Reproducibility of the interpretation of hysterosalpingography in the diagnosis of tubal pathology

Ben W.J. Mol, Patricia Swart, Patrick M.M. Bossuyt, Marc Van Beurden, and Fulco Van der Veen


Abstract

Objective: The aim of the study was to estimate the inter- and intra-observer reproducibility of the interpretation of hysterosalpingography (HSG) in the diagnosis of tubal pathology, and to assess the relation between this reproducibility and the diagnostic performance of HSG.

Methods: Four observers evaluated 143 HSGs twice, on proximal tubal occlusion, distal tubal occlusion, hydrosalpinx and peritubal adhesions. Reproducibility (inter- and intra-observer agreement) of the interpretation of HSG was expressed in terms of kappa-values. Diagnostic performance was expressed in terms of sensitivity, specificity and likelihood ratios (LR), using diagnostic laparoscopy with chromopertubation as the reference strategy.

Results: Kappa-values for reproducibility between observers were almost perfect for proximal occlusion, substantial for distal occlusion and hydrosalpinx, and moderate for adhesions. Kappa-values for reproducibility within observers were almost perfect for proximal occlusion and substantial for distal occlusion, hydrosalpinx and adhesions. Hysterosalpingography had a high specificity in the diagnosis of proximal occlusion, but a low sensitivity. The performance of HSG in the diagnosis of distal tubal occlusion, absence of hydrosalpinx and adhesions was poor. The LR for the presence of hydrosalpinx was high.

Conclusion: Proximal tubal occlusion detected on HSG changed the pre-test probability of proximal tubal occlusion from 16% to 50%, whereas proximal tubal patency detected on HSG changed the pre-test probability from 16% to 9%. It is unlikely that a lack of reproducibility of the interpretation of proximal tubal occlusion is responsible for the low sensitivity; alternative explanations are artifacts occurring while performing HSG or laparoscopy being an imperfect reference strategy. Hysterosalpingography has no value in the detection of peritubal adhesions.
Chapter 2

2.1 Introduction

The first reports on the test characteristics of hysterosalpingography (HSG) appeared early in the 1970s. Since then, many studies on the diagnostic performance of the HSG have been published. A meta-analysis of all studies published between 1968 and 1994 comparing HSG with diagnostic laparoscopy with respect to the detection of tubal pathology (presented in the next chapter of this thesis), showed that the sensitivity of HSG in the diagnosis of tubal occlusion was 65%, with a specificity of 83%. Hysterosalpingography appeared to be an inappropriate method for detecting peritubal adhesions.

A lack of reproducibility in test results is a potential explanation for a low diagnostic performance, i.e., sensitivity and specificity, of the corresponding diagnostic procedure. If observers disagree on the reading of a test result, the test is unlikely to have sufficient diagnostic performance. Therefore, a poor reproducibility limits the clinical value of a diagnostic test. If a low reproducibility can be held responsible for the low sensitivity of HSG in the diagnosis of tubal occlusion and the poor performance in the detection of peritubal adhesions, introducing guidelines for interpretation and training observers may enhance the diagnostic quality of HSG.

This study reports on the within and between observer reproducibility and diagnostic performance of the HSG, and the association between both, for the detection of tubal pathology.

2.2 Materials and methods

The study was performed in the department of reproductive medicine of the Academic Medical Center in Amsterdam. All patients in whom HSG was performed between May 1985 and November 1987 were included. Patients whose file or X-ray was not retraceable, and patients who did not undergo laparoscopy, were excluded.

Hysterosalpingographies were performed in patients with primary and secondary subfertility, as part of the routine subfertility work-up. Hysterosalpingographies were made in the follicular phase, with a water-soluble contrast medium (telebrix), by residents of the department. A spasmolytic was not used. Patients were lying on their backs during the whole procedure. Fluoroscopy was recorded on videotape and four X-rays were taken on each occasion.

In 1994, all X-rays were evaluated twice by four members of the center for reproductive medicine in two sessions with a time-interval of three months. Fluoroscopy was not available for the observers. The participants were a professor in gynecology, two gynecologists of the section reproductive medicine, and a resident in gynecology. None had previously been involved in performing the HSGs. The results of the laparoscopy were not available to the participants at the time of HSG evaluation. To avoid memory-bias, the X-rays were shown in a different order during the second session. The observers were not allowed to discuss HSGs during and between sessions. At the time this study was conducted, there were no written guidelines for interpretation of HSG in our department.
The observers evaluated the HSGs on four items, for each tube separately: (a) proximal tubal occlusion, i.e., absence of filling of the tube, (b) distal tubal occlusion, i.e., at least partly filling of the tube, without swelling of the tube and without contrast in the peritoneal cavity, (c) hydrosalpinx, i.e., at least partly filling of the tube, with swelling of the tube and without contrast in the peritoneal cavity, and (d) peritubal adhesions, i.e., contrast in the peritoneal cavity, but bowel contours not visible. Each item was diagnosed as ‘present’ or ‘absent’. Moreover, an item could be diagnosed as ‘inadequate for interpretation’ if no contrast could pass as a consequence of a proximal abnormality. For instance, if a tube was diagnosed as ‘proximally occluded’, the items ‘distal occlusion’, ‘hydrosalpinx’ and ‘peritubal adhesions’ were ‘inadequate for interpretation’.

Reproducibility for each of the four evaluated items was expressed using kappa statistics. A kappa-value of 0 indicated no agreement beyond chance, a kappa-value of 1 indicated perfect agreement between observers. The reproducibility in the case of kappa-values between 0 and 0.2 was regarded as ‘slight’, between 0.2 and 0.4 as ‘fair’, between 0.4 and 0.6 as ‘moderate’, between 0.6 and 0.8 as ‘substantial’, and between 0.8 and 1.0 as ‘almost perfect’.

Kappa-values and 95% confidence intervals (CI) for inter- and intra-observer reproducibility were calculated with four and two raters, respectively. For the calculation of the inter-observer reproducibility, only the results of the first session were used. In addition, if an item could be diagnosed as ‘present’, ‘absent’ or ‘inadequate for interpretation’, kappa-values were calculated for the inter- and intraobserver reproducibility of these three results separately.

Laparoscopy with chromopertubation was the reference strategy for the assessment of diagnostic performance. Laparoscopy was always performed after the HSG, with a double puncture technique, by a resident in gynecology, supervised by a staff member. Methylene Blue was injected at room temperature through a Foley catheter into the uterine cavity. The amount of Methylene Blue was variable, depending on the time necessary to assess tubal function. Abnormal findings were recorded on videotape. Laparoscopy was evaluated independently from the results of HSG. Results of diagnostic laparoscopy were reported systematically in the medical files.

Diagnostic performance (with 95% CI) was calculated as the mean of four observers, and expressed in terms of sensitivity, specificity and likelihood ratios (LR). Calculations of diagnostic performance were based solely on results of the first judgment session of the HSGs. Presence of disease on HSG (tubal occlusion/hydrosalpinx/peritubal adhesions) was regarded as a positive test result, whereas absence of disease was regarded as a negative test result. Therefore, sensitivity was defined as the percentage affected tubes correctly detected by HSG. Specificity was defined as the percentage of unaffected tubes properly classified by HSG.

A two-by-two table was constructed for proximal tubal occlusion, and sensitivity, specificity and LRs were calculated. For distal tubal occlusion, hydrosalpinx, and peritubal adhesions, a three-by-two table was constructed because some HSGs were diagnosed as ‘inadequate for interpretation’. Consequently, LRs of presence of abnormality, absence of
abnormality and 'inadequate for interpretation' were calculated. To assess the relationship between diagnostic performance and reproducibility, we stratified the results of the interpretation of HSG for the amount of agreement. The results of the two sessions were added, so there were eight observations on each item per tube.

2.3 Results

Between May 1985 and November 1987, 370 patients underwent HSG. In 201 patients laparoscopy with chromopertubation was also performed. The files and/or X-rays of 58 of them could not be retrieved. As a consequence HSGs of 143 patients were included in the study, so each observer evaluated 286 tubes. At laparoscopy, the prevalence of proximal tubal occlusion was 16%, the prevalence of distal tubal occlusion 29%, the prevalence of hydrosalpinx 13%, and the prevalence of peritubal adhesions 24%.

Reproducibility

The inter- and intra-observer reproducibility statistics for individual observers are presented in Table 1. The overall kappa-values (with 95 % CI) for inter-observer reproducibility were 0.85 (0.81 to 0.90) for proximal occlusion, 0.69 (0.66 to 0.73) for distal occlusion, 0.64 (0.60 to 0.68) for hydrosalpinx, and 0.55 (0.51 to 0.59) for adhesions. For distal occlusion, hydrosalpinx, and peritubal adhesions, the interpretation of an item could be divided in three different categories. In these cases it was possible to calculate the reproducibility for each interpretation separately. For distal occlusion, the kappa-values of the test results 'present', 'absent', and 'inadequate for interpretation' were 0.53, 0.71, and 0.85. For hydrosalpinx 'present', the kappa-value was fair, 0.24. For hydrosalpinx 'absent' and 'inadequate for interpretation' kappa's were 0.61 and 0.71, respectively. The test result 'adhesions present' was slightly reproducible (kappa-value 0.18). The results 'adhesions absent' and 'adhesions inadequate for interpretation' were better reproducible: 0.55 and 0.70, respectively.

The overall kappa-values (with 95% CI) for intra-observer reproducibility were 0.89 (0.84 to 0.95) for proximal occlusion, 0.72 (0.68 to 0.76) for distal occlusion, 0.68 (0.63 to 0.73) for hydrosalpinx, and 0.65 (0.60 to 0.69) for adhesions (Table 1). Intra-observer reproducibility of the separate test results of distal occlusion for the test result 'absent' was substantial.
Reproducibility of the interpretation of HSG

(kappa-value 0.72) and for the results ‘present’ and ‘inadequate for interpretation’ moderate, 0.57 and 0.58, respectively. For hydrosalpinx ‘present’, the kappa-value was fair, 0.37. For hydrosalpinx ‘absent’ and ‘inadequate for interpretation’ this was 0.69 and 0.72, respectively. For the result ‘adhesions present’ the kappa-value was only fair (kappa-value 0.38), and for the results ‘adhesions absent’ and ‘adhesions inadequate for interpretation’ substantial, 0.66 and 0.73 respectively.

Diagnostic performance

Data on the diagnostic performance of HSG are presented in Table 1. The LRs of proximal tubal occlusion were 6.0 (4.5 to 7.9) and 0.60 (0.53 to 0.69) for a positive and a negative test result, respectively, corresponding with a sensitivity of 44% (37% to 52%) and a specificity of 92% (91% to 94%). For distal occlusion the LR of a positive test result was 2.1 (1.8 to 2.3) and the LR of a negative test result was 0.43 (0.33 to 0.55). For hydrosalpinx the LR of a positive test result was 5.8 (4.4 to 7.6). However, the LR of a negative test result was 0.64 (0.54 to 0.76). For adhesions, the LR of a positive test result was 1.8 (1.3 to 2.6) and the LR of a negative test result 0.61 (0.53 to 0.71).

Reproducibility and diagnostic performance

Figure 1 shows a histogram for each item, in which the diagnosis on each tube as made at HSG is stratified by the amount of agreement at HSG. The white parts of the bars represent the presence of abnormality on laparoscopy, whereas the black parts represent a normal laparoscopy. Figure 1A shows the ideal situation: if, according to the reference standard, disease is present, all observers consider disease to be present, and vice versa.

For the item proximal occlusion the eight observers strongly agreed in their interpretation, resulting in only a few tubes in which no complete agreement was reached. Therefore, the lack of diagnostic performance for this item could not be explained by lack of reproducibility of the interpretation of proximal occlusion on HSG. Figure 1B shows that the lack of diagnostic performance was caused by false-positive and false-negative test results on the extremes (either 100% positive or 0% positive), probably caused by artifacts that occurred while performing HSG.

For distal tubal occlusion, hydrosalpinx and adhesions, there were many HSGs with <100% agreement (Figure 1C, 1D, and 1E). Moreover, the 100% positive and 100% negative diagnoses for distal occlusion contained a considerable amount of false-positive and false-negative test results. Lack of performance in the diagnosis of distal tubal occlusion can be explained by both lack of reproducibility and artifacts in the performance of HSG.

2.4 Discussion

This study showed that the within and between reproducibility of the interpretation of proximal tubal occlusion on HSG is almost perfect (kappa-value 0.85 and 0.89, respectively). For this item, HSG has a sensitivity of only 44% with a specificity of 92%, compared to laparoscopy. These data were in line with the results of the meta-analysis presented in the next chapter, in which summary point-estimates of 65% for sensitivity and 83% for specificity were calculated for tubal occlusion. In the present study, a finding of proximal
Figure 1: Relation between diagnostic accuracy and reproducibility. Each bar represents the number of observers that judged a tube as healthy or affected. The black bars represent healthy tubes, whereas the white bars represent affected tubes. (A. Test with perfect performance; B. Situation for proximal tubal occlusion; C. Situation for tubal occlusion; D. Situation for hydrosalpinx; E. Situation for peritubal adhesions)

A. Test with perfect performance

B. Proximal tubal occlusion

C. Distal tubal occlusion

D. Hydrosalpinx

E. Peritubal adhesions

tubal occlusion at HSG changed the probability of proximal occlusion from 16% to 50%. A finding of proximal tubal patency changed the probability of proximal tubal occlusion from 16% to 9%. The high reproducibility of the interpretation of proximal occlusion on HSG suggests that the low sensitivity of HSG cannot be increased by better interpretation of HSGs. For this reason, training of observers or the introduction of guidelines is unlikely to improve reproducibility and diagnostic performance. There remain three other possible explanations for the lack of sensitivity and specificity. First, laparoscopy might not be a perfect reference standard. Alternatively, the lack of diagnostic performance can be due the fact that the observers only were given access to X-rays, rather than fluoroscopy. Thirdly, artifacts might have occurred while performing HSG. Artifacts might include premature ending of the procedure, insufficient pressure because of vaginal reflux, or differences in
muscle tonus of the tubes. In that case, sensitivity and specificity can only be improved by optimizing the technique for performing HSG.

The obtained inter- and intra-observer reproducibility of the interpretation of distal occlusion was only substantial (kappa-value 0.69 and 0.72, respectively). As expected, assessment of overflow of contrast into the peritoneal cavity is more difficult than assessment of filling of the tube. The LRs were 2.0 for distal tubal occlusion and 0.43 for the absence of tubal occlusion, respectively. This combination limits the use of HSG to detect distal tubal occlusion. For evaluating distal tubal occlusion, training of observers or the introduction of guidelines might improve reproducibility and therefore performance of the diagnosis of distal tubal occlusion.

As for distal tubal occlusion, inter- and intra-observer reproducibility of the interpretation of hydrosalpinx on HSG were only substantial (kappa-value 0.64 and 0.68). The LRs of a positive and negative test result were 5.8 and 0.64, respectively. This means that HSG is a useful test to detect hydrosalpinx, but an imperfect test to rule out hydrosalpinx. In the meta-analysis presented in chapter 3 hydrosalpinx was not studied, due to a lack of data. Here also, training of observers or the introduction of guidelines could improve reproducibility and therefore performance of the diagnosis.

Likelihood ratios of 1.8 for a positive test result and of 0.61 for a negative test result were calculated for adhesions. HSG therefore is unreliable for the evaluation of peritubal adhesions. Overall, inter-observer reproducibility was moderate (kappa-value 0.55) and intra-observer reproducibility just substantial (kappa-value 0.65). However, inter-observer reproducibility of the test result 'adhesions present' was only slight, and its intra-observer reproducibility fair. This lack of reproducibility indicated that HSG is of no clinical significance for the diagnosis of peritubal adhesions.

In conclusion, we can say that HSG might have potential in the diagnosis of tubal occlusion. Lack of agreement with tubal occlusion at laparoscopy is not caused by a lack of reproducibility of its interpretation. Hysterosalpingography is a good test to detect hydrosalpinx, but not very useful in ruling out this condition. Improvement of reproducibility through the introduction of guidelines might improve diagnostic performance and therefore the practical value of the HSG with regard to distal occlusion and hydrosalpinx. Finally, HSG is of no use for the detection of adhesions.

Acknowledgements
The authors thank F.B. Lammes MD PhD and H.V. Hogerzeil MD PhD for the interpretation of the HSGs, and W.M. Ankum MD PhD and J.V.T.H. Hamerlynck MD PhD for critically reading the manuscript.

2.5 References


