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

Citation for published version (APA):
3. The diagnostic performance of hysterosalpingography in the diagnosis of tubal pathology, a meta-analysis

Patricia Swart, Ben W.J. Mol, Fulco van der Veen, Marc van Beurden, William K. Redekop, and Patrick M.M. Bossuyt
Fertility and Sterility 1995;64:486-91.

Abstract

Objective: The aim of this study was to assess the performance of hysterosalpingography (HSG) in diagnosing tubal occlusion and peritubal adhesions using laparoscopy with chromopertubation as the reference standard.

Methods: A meta-analysis was performed of 20 studies comparing HSG and laparoscopy for tubal occlusion and peritubal adhesions. In case homogeneity could not be rejected a point estimate of sensitivity and specificity was calculated. In case sensitivity and specificity showed a negative correlation a summary receiver operating characteristic (ROC) curve was estimated.

Results: For tubal occlusion the reported sensitivity and specificity differed between studies. In a subset of three studies that evaluated HSG and laparoscopy independently a point-estimate of 65% for sensitivity (95% CI 50% to 78%) and 83% for specificity (95% CI 77% to 88%) was calculated. For peritubal adhesions a summary ROC-curve could be estimated, that indicated a poor diagnostic performance.

Conclusions: The performance of HSG in the diagnosis of tubal occlusion is sufficient for its use as a triage before laparoscopy. After a normal HSG laparoscopy can be withheld for a considerable time. For the evaluation of peritubal adhesions HSG is not reliable.
3.1 Introduction

Hysterosalpingography (HSG) and laparoscopy with chromopertubation are two widely used methods to determine tubal function in subfertile women. The place of HSG in the diagnostic work-up, however, is still a matter for debate. Some authors consider HSG to be an indispensable test that should be performed before laparoscopy\textsuperscript{1-3}, whereas others favor omission of the HSG altogether.\textsuperscript{4,5} A frequently noted point in favor of the HSG is its presumed ability to improve conception chances.\textsuperscript{2,6-7}

To determine the performance of HSG in the diagnosis of tubal pathology, a systematic review of the literature was carried out. A meta-analysis was set up for studies evaluating the diagnostic performance of HSG, using laparoscopy as the reference standard.

3.2 Materials and methods

Search Strategy

A computerized MEDLINE search to identify all registered articles in the English, French, German and Dutch languages published between January 1968 and July 1994 was performed. Keywords used were ‘hysterosalpingography’ and ‘laparoscopy’. In addition, all volumes of *Fertility and Sterility, Human Reproduction, The British Journal of Obstetrics and Gynaecology, Obstetrics and Gynecology, The American Journal of Obstetrics and Gynecology, The New England Journal of Medicine, The Lancet, The Journal of the American Medical Association* and *The British Medical Journal* published between January 1974 and June 1994 were reviewed manually. Cross-references in all selected articles were checked.

Articles comparing HSG and laparoscopy for tubal pathology were included. Tubal pathology was defined in two ways, first, as the presence of tubal occlusion, i.e., the absence of filling or the absence of overflow, and, second, as the presence of peritubal adhesions. If studies reported on proximal and distal tubal occlusion, these data were analyzed separately.

Studies were excluded when data were not sufficient to construct a two-by-two table of the test (HSG) and the reference standard (laparoscopy) for tubal occlusion and/or peritubal adhesions. Studies that did not distinct between tubal occlusion and peritubal adhesions but only reported on ‘tubal pathology’ were also excluded. In the analysis of peritubal adhesions only those patients in whom patency of at least one tube was demonstrated were included in the two-by-two table.

Analysis

The meta-analysis was conducted according to the methodology described by Midgette \textit{et al.}\textsuperscript{8} For each study sensitivity, specificity and likelihood ratios (LR) were calculated from the published data. Thereafter, tests for homogeneity by means of the $\chi^2$-test statistic were performed for sensitivity and specificity separately.\textsuperscript{9} If homogeneity could not be rejected for both sensitivity and specificity, a more precise summary point estimate of sensitivity and specificity, with its 95\% confidence intervals (CI), was calculated. Sample sizes were used as the weight of each study.\textsuperscript{10}
In case of heterogeneity in the reported sensitivity and specificity, the differences could be caused by a shift in cutoff levels for test positivity. In that case, higher sensitivity should be accompanied by lower specificity values, and vice versa. This hypothesis was explored by calculating a Spearman correlation coefficient. If there appeared to be negative correlation (a correlation coefficient of -0.5 or lower was considered to be negative), the pairs of sensitivity and specificity could be thought of as originating from a single Receiver Operating Characteristic (ROC)-curve. A summary ROC-curve could then be estimated.

In absence of a negative correlation between sensitivity and specificity, alternative explanations for the heterogeneity were examined. Subgroups of studies were distinguished by means of pre-defined gynecological and methodological criteria. Subgroup analysis examines the possibility that the variations in study results found can be attributed to differences in study population, study methodology, and/or used techniques.

Subgroups were selected on the following criteria: 1. prevalence of disease, i.e., presence of tubal occlusion or presence of peritubal adhesions (prevalence <35% or prevalence ≥35%); 2. contrast medium used for HSG (oil- or water-soluble); 3. use of spasmolyticum while performing HSG; 4. setting (academic or non-academic); 5. judgement of the laparoscopy with or without knowledge of the HSG (dependent or independent); 6. use of criteria for judgement of HSG; 7. time between HSG and laparoscopy; 8. number of patients included in the study; 9. HSG judged by a gynecologist or by a radiologist, and 10. type of canula used for performing a HSG. Within each subgroup, the procedure reported earlier was repeated.

3.3 Results

Search Strategy

In total, 45 studies were identified, 38 by MEDLINE Search, six by cross-references and one by hand-search. Of the 35 studies that met the inclusion criteria, 15 studies were excluded according to the pre-defined criteria. Therefore 20 studies were analyzed. Study characteristics and results of these studies are listed in Tables 1 and 2.

Nineteen studies analyzed tubal occlusion. Twelve of these studies distinguished between patients with two open tubes and patients with one or two occluded tubes, whereas the other seven studies distinguished patients with two patent tubes, patients with one patent and one occluded tube, and patients with two occluded tubes. Defining disease as the presence of at least one occluded tube, sensitivity, specificity, and LRs were calculated. Thirteen studies reported on peritubal adhesions.

Analysis

The χ²-test statistic showed our data to be heterogeneous for tubal occlusion (P-value < 0.01, Table 3). For this reason the calculation of point estimates for sensitivity and specificity was not meaningful. Heterogeneity was confirmed by the plot of sensitivity against specificity, which showed that 95% CIs do not overlap (Figure 1). Only 2 of 20 studies distinguished distal and proximal tubal occlusion, so this distinction was not analyzed separately.
### Table 1: Hysterosalpingography versus laparoscopy for peritubal adhesions

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Number of Patients*</th>
<th>Prevt</th>
<th>Sens†</th>
<th>Spect</th>
<th>LR (pos)‡</th>
<th>LR (neg)‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swolin and Rosenkrantz</td>
<td>13</td>
<td>143</td>
<td>18</td>
<td>54</td>
<td>83</td>
<td>3.2</td>
<td>0.6</td>
</tr>
<tr>
<td>Maathuis et al.</td>
<td>14</td>
<td>207</td>
<td>38</td>
<td>72</td>
<td>83</td>
<td>4.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Duignan et al.</td>
<td>15</td>
<td>273</td>
<td>25</td>
<td>85</td>
<td>89</td>
<td>8.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Hutchins</td>
<td>1</td>
<td>409</td>
<td>20</td>
<td>76</td>
<td>90</td>
<td>7.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Taylor</td>
<td>4</td>
<td>34</td>
<td>12</td>
<td>100</td>
<td>70</td>
<td>3.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Philipsen and Hansen</td>
<td>17</td>
<td>152</td>
<td>36</td>
<td>84</td>
<td>86</td>
<td>5.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Donnez et al.</td>
<td>25</td>
<td>974</td>
<td>34</td>
<td>100</td>
<td>99</td>
<td>72</td>
<td>0.0</td>
</tr>
<tr>
<td>Duff et al.</td>
<td>7</td>
<td>52</td>
<td>40</td>
<td>67</td>
<td>84</td>
<td>4.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Snowden et al.</td>
<td>23</td>
<td>77</td>
<td>48</td>
<td>73</td>
<td>88</td>
<td>5.8</td>
<td>0.3</td>
</tr>
<tr>
<td>Rice et al.</td>
<td>12</td>
<td>143</td>
<td>49</td>
<td>83</td>
<td>85</td>
<td>5.5</td>
<td>0.2</td>
</tr>
<tr>
<td>La Sala et al.</td>
<td>3</td>
<td>100</td>
<td>31</td>
<td>65</td>
<td>62</td>
<td>1.7</td>
<td>0.6</td>
</tr>
<tr>
<td>Loy et al.</td>
<td>21</td>
<td>77</td>
<td>26</td>
<td>80</td>
<td>72</td>
<td>2.9</td>
<td>0.3</td>
</tr>
</tbody>
</table>

*Studies that compare two groups of hysterosalpingography results

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Number of Patients*</th>
<th>Prevt</th>
<th>Sens†</th>
<th>Spect</th>
<th>LR (pos)‡</th>
<th>LR (neg)‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keirse and Vandervellen</td>
<td>16</td>
<td>73</td>
<td>50</td>
<td>54</td>
<td>78</td>
<td>4.5</td>
<td>0.3</td>
</tr>
<tr>
<td>El Minawi et al.</td>
<td>6</td>
<td>78</td>
<td>352</td>
<td>26</td>
<td>89</td>
<td>8.9</td>
<td>0.1</td>
</tr>
<tr>
<td>Montanari et al.</td>
<td>14</td>
<td>82</td>
<td>85</td>
<td>35</td>
<td>83</td>
<td>7.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Nordenskjöld and Ahlgren</td>
<td>22</td>
<td>83</td>
<td>326</td>
<td>17</td>
<td>78</td>
<td>3.1</td>
<td>0.3</td>
</tr>
<tr>
<td>La Sala et al.</td>
<td>20</td>
<td>86</td>
<td>174</td>
<td>29</td>
<td>34</td>
<td>89</td>
<td>3.0</td>
</tr>
<tr>
<td>WTIO</td>
<td>24</td>
<td>86</td>
<td>121</td>
<td>60</td>
<td>82</td>
<td>55</td>
<td>1.8</td>
</tr>
<tr>
<td>Ismajovich et al.</td>
<td>19</td>
<td>86</td>
<td>215</td>
<td>78</td>
<td>74</td>
<td>68</td>
<td>2.3</td>
</tr>
</tbody>
</table>

*Donnez et al. and Duff et al. = number of tubes
†Prevt = prevalence; sens = sensitivity; spec = specificity.
‡LR (pos) = Likelihood ratio of positive test result; LR (neg) = Likelihood ratio of negative test result.

### Table 2: Hysterosalpingography versus laparoscopy for peritubal adhesions

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Excluded patients</th>
<th>Number of Patients*</th>
<th>Prevt</th>
<th>Sens†</th>
<th>Spect</th>
<th>LR (pos)‡</th>
<th>LR (neg)‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swolin and Rosencrantz</td>
<td>13</td>
<td>29</td>
<td>114</td>
<td>41</td>
<td>79</td>
<td>61</td>
<td>2.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Maathuis et al.</td>
<td>14</td>
<td>61</td>
<td>146</td>
<td>43</td>
<td>51</td>
<td>73</td>
<td>1.9</td>
<td>0.7</td>
</tr>
<tr>
<td>Duignan et al.</td>
<td>15</td>
<td>0</td>
<td>273</td>
<td>7</td>
<td>3</td>
<td>98</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Keirse and Vandervellen</td>
<td>16</td>
<td>25</td>
<td>25</td>
<td>43</td>
<td>83</td>
<td>85</td>
<td>5.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Hutchins</td>
<td>1</td>
<td>93</td>
<td>316</td>
<td>12</td>
<td>39</td>
<td>99</td>
<td>36.6</td>
<td>0.6</td>
</tr>
<tr>
<td>El Minawi et al.</td>
<td>6</td>
<td>108</td>
<td>244</td>
<td>46</td>
<td>54</td>
<td>89</td>
<td>4.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Philipsen and Hansen</td>
<td>17</td>
<td>60</td>
<td>92</td>
<td>30</td>
<td>46</td>
<td>61</td>
<td>1.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Montanari et al.</td>
<td>18</td>
<td>31</td>
<td>54</td>
<td>22</td>
<td>17</td>
<td>95</td>
<td>3.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Duff et al.</td>
<td>7</td>
<td>19</td>
<td>33</td>
<td>33</td>
<td>64</td>
<td>77</td>
<td>2.8</td>
<td>0.5</td>
</tr>
<tr>
<td>La Sala et al.</td>
<td>20</td>
<td>31</td>
<td>143</td>
<td>30</td>
<td>30</td>
<td>93</td>
<td>4.3</td>
<td>0.8</td>
</tr>
<tr>
<td>Loy et al.</td>
<td>21</td>
<td>77</td>
<td>77</td>
<td>53</td>
<td>29</td>
<td>92</td>
<td>3.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Ismajovich et al.</td>
<td>19</td>
<td>53</td>
<td>162</td>
<td>58</td>
<td>77</td>
<td>50</td>
<td>1.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Reshef et al.</td>
<td>26</td>
<td>0</td>
<td>215</td>
<td>35</td>
<td>13</td>
<td>87</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

*Donnez et al. and Duff et al. = number of tubes
†Prevt = prevalence; sens = sensitivity; spec = specificity.
‡LR (pos) = Likelihood ratio of positive test result; LR (neg) = Likelihood ratio of negative test result.
Table 3: Results of tests for homogeneity (P-values of the $\chi^2$-test statistic) and correlation (Spearman correlation coefficient).

<table>
<thead>
<tr>
<th>Subgroups of tubal occlusion</th>
<th>sensitivity</th>
<th>specificity</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>tubal occlusion</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.22</td>
</tr>
</tbody>
</table>

**Subgroups tubal occlusion**

- Prevalence <35%: $P = 0.05$, $r = 0.05$
- Prevalence $\geq 35%$: $P < 0.01$, $r = 0.30$
- Contrast medium oil-soluble: $P < 0.01$, $r = 0.45$
- Contrast medium water-soluble: $P < 0.01$, $r = 0.80$
- Spasmolyticum: $P < 0.01$, $r = 0.14$
- No spasmolyticum: $P < 0.01$, $r = 0.39$
- Independent judgement: $P = 0.35$
- Dependent judgement: $P < 0.01$, $r = 0.35$
- Criteria for judgment of HSG: $P = 0.53$, $r = 0.01$
- No criteria for judgment of HSG: $P < 0.01$, $r = 0.25$
- Time between HSG - laparoscopy $\leq 3$ months: $P < 0.01$, $r = 0.50$
- Time between HSG - laparoscopy $> 3$ months: $P < 0.01$, $r = 0.39$
- Number of patients $\leq 250$: $P < 0.01$, $r = 0.14$
- Number of patients $> 250$: $P < 0.01$, $r = 0.10$
- Judgement by gynecologist or radiologist: not possible
- Type of canula: not possible
- Peritubal adhesions: $P < 0.01$, $r = 0.69$

Visual inspection of the sensitivity and the specificity made it unlikely that the pairs of sensitivity and specificity originated from the same ROC-curve (Figure 1). This was confirmed by the absence of a negative correlation (Spearman correlation coefficient $-0.22$, Table 3).

At the next step, analysis of the subgroups, homogeneity could not be rejected for the three studies that judged HSG and laparoscopy independently ($P$-values 0.18 and 0.99 for sensitivity and specificity, respectively, Table 3). The calculated point estimates for these studies were 65% for sensitivity (95% CI 50% to 78%) and 83% for specificity (95% CI 77% to 88%) (Figure 1A). It was not possible to distinguish subgroups for HSG-judgement by a gynecologist or a radiologist and for the type of canula used for performing a HSG. We could identify only one study in which the judgement was performed by a radiologist. Ten of 20 studies did not mention which type of canula was used whereas the other 10 used 7 different canulas. All other subgroups were heterogeneous and correlation was insufficient to justify estimation of a summary ROC-curve.

The $\chi^2$-test statistic showed our data to be heterogeneous for peritubal adhesions, too. However, visual inspection showed that the studies could be thought of as originating from the same ROC-curve. This was confirmed by a negative correlation (Spearman correlation coefficient $-0.69$, Table 3). A summary ROC-curve could be estimated for the ranges of
reported sensitivity (0% to 83%) and specificity (50% to 99%). Figure 1B shows sensitivity, specificity and 95% CIs for peritubal adhesions with the estimated ROC-curve.

3.4 Discussion

In this meta-analysis we evaluated the performance of HSG in the diagnosis of tubal pathology. For tubal occlusion an overall point estimate of sensitivity and specificity could not be calculated and a summary ROC-curve could not be constructed. Subgroup analysis indicated that for studies in which HSG and laparoscopy were judged independently, a point-estimate of 65% for sensitivity (95% CI 50% to 78%) and of 83% for specificity (95% CI 77% to 88%) could be calculated for tubal occlusion. This point estimate is the best assessment of the truth, firstly, since it represents a pooled estimate of several studies and is therefore more accurate compared with the result of a single study, and, secondly, since independent judgement should have prevented observer-bias. For peritubal adhesions data allowed the construction of a summary ROC-curve (Figure 1B). This summary ROC-curve represents the best assessment of the truth for peritubal adhesions.

Two methodological issues are of importance in the interpretation of the data. The first issue to be considered is that all studies were performed in retrospective and included only patients who underwent both HSG and laparoscopy. As a consequence women who conceived after HSG and women who refused further investigations following HSG were not included.

In the work-up of subfertility it is widely accepted to perform laparoscopy shortly after HSG when findings are abnormal and to postpone laparoscopy with 6 months when HSG findings are normal, and only perform laparoscopy in case pregnancy does not occur. Therefore, abnormal HSGs will be over-represented in the populations studied. As a consequence, there is verification or work-up bias. Verification bias occurs when only a subset of all tested subjects are selected for further work-up and verification of the test result. Several methods for correction of verification bias have been suggested. These correction methods all assume that the decision to verify the diagnosis depends only on the result of the test under study and clinical information, and not on the clinical context or random variation. Since verification of the diagnosis with laparoscopy does not only depend on the HSG findings, but also on the fact that pregnancy did not occur between HSG and laparoscopy, the proposed correction methods can not be applied in this situation. Additionally, a practical limitation that hampers correction for verification bias is that exact information on the number of patients with a normal HSG, in whom laparoscopy had not been performed was never retraceable from the data supplied. However, since a normal HSG will not be verified after conception, the presence of tubal pathology in such patients conceiving after a normal HSG will never be known. The presence of tubal pathology at laparoscopy in these patients would rather be unlikely, and since these patients would be classified as 'true-positive', the verification bias is likely to underestimate the specificity, although we can not quantify this estimation.

One could overcome verification bias by performing HSG and laparoscopy in one session. Only in three studies laparoscopy was performed shortly after HSG. However,
Figure 1AB: Sensitivity, specificity and 95% CI in a Receiver Operating Characteristic (ROC) sheet for the performance of hysterosalpingography in the diagnosis of tubal occlusion (1A) and in the diagnosis of peritubal adhesions (1B). For the diagnosis of tubal occlusion a point estimate could be calculated from three studies that judged hysterosalpingography and laparoscopy independently. For the diagnosis of peritubal adhesions, a ROC-curve could be estimated.

in one of these studies a substantial fraction of the patients had no laparoscopy, whereas there was no independent judgment in another study. The third study had a sensitivity of 54% and a specificity of 83%, in accordance with that of the other two studies that judged HSG and laparoscopy independently.

The second issue is the fact that laparoscopy is not a perfect reference test. Sometimes tubal occlusion at laparoscopy is due to artifacts, because of technical failure and differences in resistance between the two tubes. The sensitivity of 65% that we found could partly be due to artifacts of laparoscopy. However, laparoscopy is the best standard available.

Apart from the diagnostic importance of HSG, it is also thought to have a therapeutic value. A meta-analysis of four studies assessing the therapeutic role of oil- versus water soluble contrast media are compared, showed that oil-soluble contrast media have a pregnancy-inducing effect compared with water soluble contrast media.

The clinical significance of HSG cannot be concluded from the results of this meta-analysis. If HSG is used as a triage before laparoscopy, it should be compared with Chlamydia Antibody Testing (CAT). A meta-analysis on the diagnostic performance of CAT in the diagnosis of tubal pathology is presented in the next chapter. Apart from its use in the triage before laparoscopy, HSG can also be used in the assessment of the fertility prognosis. This aspect is addressed in chapter 5 and 6.

Acknowledgements
The authors thank W.M. Ankum MD PhD, J.V.T.H. Hamerlynck MD PhD and F.B. Lammes MD PhD from the department of Obstetrics and Gynecology, Academic Medical Center, University of Amsterdam, for critically reading the manuscript.
3.5 References


