The accuracy of endometrial sampling in the diagnosis of patients with endometrial carcinoma and hyperplasia: a meta-analysis

Dijkhuizen F P, Mol B W, Brolmann H A, Heintz A P

Authors' objectives
To review the literature, and to evaluate the accuracy of endometrial sampling in the diagnosis of patients with endometrial carcinoma or atypical hyperplasia.

Searching
MEDLINE was searched from January 1966 to May 1999 using the terms 'endometri', 'biopsy' and/or 'Pipelle'. The references in selected articles were checked.

Study selection
Study designs of evaluations included in the review
Diagnostic accuracy studies were eligible for inclusion.

Specific interventions included in the review
Studies of endometrial sampling by any method were included. The most common methods found were the Pipelle device, Z-sampler, the Vabra or Novak device, and a technique based on lavage. More than one sampling method was sometimes used.

Reference standard test against which the new test was compared
The reference standard was at least one of the following: dilation and curettage (DC), hysteroscopy, and hysterectomy. Hysterectomy was the preferred 'gold' standard.

Participants included in the review
Women with possible endometrial carcinoma or atypical hyperplasia were included. The participants in the studies included premenopausal and postmenopausal women, and women both with and without symptoms.

Outcomes assessed in the review
True-positive and false-negative results were recorded for patients with each disease. Where data were reported for women without the disease, 2x2 tables were constructed. The number of women for whom endometrial sampling failed was also recorded.

How were decisions on the relevance of primary studies made?
The authors do not state how the papers were selected for the review, or how many of the reviewers performed the selection.

Assessment of study quality
No formal method for assessing validity was specified but the following were discussed in the text: whether the study was prospective; whether recruitment was consecutive; and whether there were independent assessments of the test and the reference standard.

Data extraction
The authors do not state how the data were extracted for the review, or how many of the reviewers performed the data extraction.

The data extracted included: patient characteristics (menopausal status and symptomatic status); type of sampling (direct biopsy or indirect); reference test; study characteristics (whether prospective, whether recruitment was
consecutive, whether assessment was independent); the number of patients where the test failed because an adequate sample could not be obtained; the number of true-positive and false-negative results, and also the false-positive and true-negative results where available.

The percentage of failed tests, and the sensitivity of endometrial sampling in the detection of endometrial carcinoma, were calculated. Specificity was calculated where possible. These calculations were repeated for the detection of either carcinoma or atypical hyperplasia where possible.

**Methods of synthesis**

How were the studies combined?
The results were combined by calculating the weighted sensitivities, specificities and failure rates, with weighting by sample size. The sensitivities and specificities were pooled by sampling method, by type of reference test, and by participant characteristics (menopausal status and symptomatic status). Correlations between the sensitivities and specificities were calculated in order to determine whether a summary receiver operating curve could be plotted.

How were differences between studies investigated?
Using a chi-squared test, the authors assessed whether the sensitivity was dependent on the method of sampling used or on the reference strategy used, in relation to the diagnosis of (1) carcinoma alone, and (2) carcinoma or atypical hyperplasia. Differences according to study validity measures were also investigated. No formal tests for heterogeneity were reported.

**Results of the review**

Thirty-nine studies (N=7,814) were included, of which one was a diagnostic case-control and the remainder were diagnostic cohort studies. The number of patients in the studies ranged from 25 to 902.

Five studies were prospective and seven reported consecutive recruitment. Methodological weaknesses were noted in all studies, and only one study reported an independent assessment. Only 27% of studies used hysterectomy, the 'gold' standard test, as the reference standard. DC was most commonly used, while hysteroscopy was the least common and was not used exclusively in any study.

Biopsy failure rates (32 studies). The biopsy failure rate ranged from 0 to 54%. The weighted failure rates were 10.4% where the Pipelle was used (15 studies) and 9.5% for other methods (17 studies) (p=0.12).

Diagnosis of endometrial carcinoma.

The sensitivities ranged from 25% (with specificity 99%) to 100% (specificity range: 90 to 100%) (39 studies). The specificities ranged from 93% (with sensitivity 100%) to 100% (sensitivity range: 50 to 100%) (31 studies).

The weighted sensitivities by reference test were 68% for hysterectomy, 78% for DC, and 81% for both (p=0.61). The corresponding weighted specificities were 99.7% for hysterectomy, 99.6% for DC, and 99.9% for both (p=0.01). The weighted sensitivities by menopausal status were 95% for postmenopausal only, and 75% for mixed or unknown (p=0.05). The corresponding weighted specificities were 99.5% for postmenopausal only (9 studies), and 99.7% for mixed or unknown (30 studies) (p=0.01).

There were no statistically-significant differences according to the presence or absence of symptoms (sensitivity p=0.31, specificity p=0.12) or according to study validity measures (no p-values reported).

For studies of postmenopausal women only, the weighted sensitivity was 99.6% when using the Pipelle (4 studies) and 97.1% when using the Vabra device (3 studies) (p=0.045). For studies of mixed menopausal status, the weighted sensitivity was 91% using the Pipelle, 80% using the Vabra device, and 57% using lavage (6 studies) (p=0.16).

The correlation between sensitivity and specificity was low (Spearman correlation coefficient -0.15).

Diagnosis of atypical hyperplasia or endometrial carcinoma.
The sensitivities ranged from 39% (with specificity 99%) to 100% (specificity range: 93 to 100%) (19 studies). The specificities ranged from 93% (with sensitivity 100%) to 100% (specificity range 67 to 100%) (17 studies).

The weighted sensitivities by reference test were 74% for hysterectomy, 75% for DC, and 45% for both (p=0.01). The corresponding weighted specificities were 100% for hysterectomy, 99.1% for DC, and 100% for both (p=0.01).

There were no statistically-significant differences according to study validity measures (no p-values reported).

There was one study of postmenopausal women only, (using the Pipelle); this reported a sensitivity of 88% and a specificity of 98%. There was one study of asymptomatic women only; this reported a sensitivity of 91% and a specificity of 98%. For studies of mixed menopausal status, the weighted sensitivity was 82.3% using the Pipelle (9 studies), 66.7% using the Vabra device (3 studies), and 53.4% using lavage (4 studies)(p=0.01). The corresponding weighted specificity was 100% using the Pipelle 100%, 99.8% using the Vabra device, and 98.9% using lavage (p=0.01).

The correlation between sensitivity and specificity was low (Spearman correlation coefficient -0.09).

**Authors’ conclusions**

Endometrial biopsy with the Pipelle was superior to other endometrial techniques in the detection of endometrial carcinoma and atypical hyperplasia. The accuracy of the Pipelle was higher in postmenopausal women than in premenopausal women.

**CRD commentary**

The review question was clearly stated, although the inclusion of several reference standards made the interpretation of the results more difficult. Only one database (MEDLINE) was searched, so it is likely that some studies were missed. The possibility that publication bias might have led to overstating the value of the test was noted. The method of conducting the review was not described (e.g. whether two reviewers carried out the screening and data extraction), so it is not possible to judge whether its conduct was satisfactory. No formal validity assessment was specified, but the effect of adherence to certain methodological standards was examined. Relevant details of the studies were reported.

Pooling of sensitivities and specificities weighted by sample size is not the recommended method of meta-analysis for diagnostic tests where there is significant heterogeneity. Heterogeneity was not assessed before pooling, although pooling was stratified according to potential mediating factors and the effect of each factor was tested. However, within-group heterogeneity was not investigated. The subgroup results quoted in the review's conclusions were based on a small number of studies and should be treated with some caution, as they were not comparisons specified in the review objectives and may have occurred by chance.

The authors do not draw definite conclusions for practice based on these results, and their conclusion that the results warrant further research is appropriate.

**Implications of the review for practice and research**

Practice: The authors did not state any implications for practice.

Research: The authors state that the sensitivity and specificity of the Pipelle device are high enough to justify further assessment of its clinical value, including its possible use with premenopausal women. They also suggest that the use of endometrial biopsy and/or transvaginal sonography for women with postmenopausal bleeding should be the subject of further research.

**Bibliographic details**

This is a critical abstract of a systematic review that meets the criteria for inclusion on DARE. Each critical abstract contains a brief summary of the review methods, results and conclusions followed by a detailed critical assessment on the reliability of the review and the conclusions drawn.