Evaluation of cervical cytology

Authors' objectives
To evaluate the accuracy of conventional and new methods of Papanicolaou (Pap) testing when used to detect cervical cancer and its precursors.

Searching
MEDLINE (from 1966), EMBASE (from 1980), HealthSTAR (from 1975), Cancerlit (from 1983) and CINAHL (from 1983) were searched. EconLit was searched for economic data. Full details of the search strategy were given. The searches were limited to English language studies conducted in humans. Manual searches of newly published relevant journal issues, bibliographies of included studies, and recent systematic reviews were also conducted. Unpublished studies were sought by contacting relevant professional societies and the manufacturers of cytological devices. The World Wide Web was also searched.

Study selection
Study designs of evaluations included in the review
Study designs included in the review were diagnostic cohort test evaluations, meta-analyses and review articles. For studies comparing cytology with a reference standard, the tests had to be reasonably concurrent, up to 3 months apart, otherwise the studies were excluded.

Specific interventions included in the review
Studies that evaluated cervical cytology as a screening test were eligible for inclusion. Methods of Pap testing included conventional methods (with or without manual rescreening), primary computer screening (AutoPap or Papnet), computer rescreening (AutoPap or Papnet), and monolayer slide preparation (ThinPrep). Other recently developed methods, the AutoCyte PREP system and the AutoCyte SCREEN system (TriPath Imaging) were not evaluated.

Reference standard test against which the new test was compared
The reference standard to which the test results were compared had to be histology or colposcopy. Studies of new technologies that used cytology as a reference standard were also included.

Participants included in the review
No inclusion criteria relating to the study population were specified. Women undergoing Pap testing for primary screening and those undergoing evaluation for prior cytological abnormality were included. The candidates for Pap smear screening included women between the age of onset of sexual activity and age 85 years.

Outcomes assessed in the review
The included studies were required to report sufficient data for the construction of 2x2 tables. The main outcome measures were the sensitivity and specificity of the cytological test for detecting cytological abnormalities and 'cases'. Cytological abnormality was defined by one of three thresholds: atypical squamous cells of undetermined significance (ASCUS); low-grade squamous intra-epithelial lesions (LSIL); and high-grade squamous intra-epithelial cells (HSIL). 'Cases' were defined as histological diagnosis of cervical intra-epithelial neoplasia (CIN) grades I to III, or carcinoma.

How were decisions on the relevance of primary studies made?
Nine members of the study's working group developed a numeric quality score (maximum 11 points) by consensus; full details of the process and scoring according to the validity criteria were given. The authors did not state how many reviewers performed the validity assessment.

Assessment of study quality
Various methodological features were used as criteria for inclusion (see Study Designs of Evaluations Included in the
Validity was evaluated using the following criteria: type of reference standard; independence of assessment; independence of selection for verification; sample selection; spectrum of disease or non disease; publication type; and whether industry related. Nine members of the study's working group developed a numeric quality score (maximum 11 points) by consensus; full details of the process and scoring according to the validity criteria were given. The authors did not state how many reviewers performed the validity assessment.

**Data extraction**
Two reviewers independently completed 2x2 tables for each study and entered them into a computer database. Four different combinations of cytological and histological thresholds were used: ASUS/CIN-I; LSIL/CIN-I; LSIL/CIN-II-III; and HSIL/CIN-II-III. Cases that were indeterminate were documented but excluded from the calculation of sensitivity and specificity. Other data extracted from the included studies and reported in tables were: study design and characteristics, interventions, location and time period, patients and methods, outcomes measured, study results and limitation, number of subjects and prevalence of disease. Where more than one study population was included in a single report with data provided separately, they were treated as individual studies.

**Methods of synthesis**

*How were the studies combined?*
A standardised effectiveness score, which accounts for both sensitivity and specificity by fitting a receiver operating characteristic (ROC) curve through a logistic odds transformation of sensitivity and specificity, was calculated. In general, a score of 3 reflects a test with good discrimination, whereas a score of 1 reflects a test that does not discriminate between disease positives and disease negatives. Summary ROC curves and pooled sensitivity and specificity were estimated using a random-effects model. One-parameter (intercept only) logistic equations were used to generate summary estimates of sensitivity and specificity and to produce summary ROC curves.

Some studies that did not permit the direct calculation of sensitivity and specificity allowed the calculation of relative true and false positive rates according to the method of Chock et al. (see Other Publications of Related Interest no.1).

*How were differences between studies investigated?*
Heterogeneity was not formally tested, although the effects of possible sources of heterogeneity were investigated. The effect of disease prevalence on the summary estimates of sensitivity and specificity was estimated. Multiple logistic regression analysis was used to estimate the variation in overall effectiveness due to the variation in prevalence. The effect of study quality on the summary effectiveness scores was also examined, initially using the individual components of the score, then using the total score both as a continuous and a dichotomous (quality score <7, >7) variable.

**Results of the review**
Eighty-four studies of the conventional Pap test and 25 studies of monolayer cytology (6 of AutoPap, 11 of Papnet and 8 of ThinPrep) were included.

Thin-layer cytology relative to histological examination or conventional Pap smear.

AutoPap (n=6): for ASCUS or worse (n=3) the sensitivity ranged from 43 to 66%; for LSIL or worse (n=2) the sensitivity ranged from 66 to 77%.

Papnet (n=11): one study provided estimates of sensitivity and specificity; depending on the threshold, the sensitivity ranged from 38 to 41% and the specificity from 82 to 92%. Other studies looked only at sensitivity; very few details were provided.

ThinPrep (n=8): two studies permitted an estimation of the sensitivity and specificity; the sensitivity ranged from 85 to 94% and the specificity from 36 to 58%. A further study compared ThinPrep with conventional Pap smears and found that ThinPrep had higher sensitivity and a slightly lower specificity. Apart from the details in the data extraction tables, no information on the other studies was provided.
Summary ROC curves were presented in the report. The summary effectiveness scores for the different diagnostic thresholds were: ASCUS/CIN-1 (31 studies), 1.03 (95% CI: 0.78, 1.14); LSIL/CIN-1 (69 studies), 1.08 (95% CI: 0.94, 1.23); LSIL/CIN-2-3 (43 studies), 1.06 (95% CI: 0.87, 1.25); HSIL/CIN-2-3 (45 studies), 1.29 (95% CI: 1.08, 1.50).

Regression analysis: disease prevalence ranged from 0.02 to 0.98. Disease prevalence was significantly associated with effectiveness score for each different threshold. The effect of blinded evaluation of the screening test was investigated in a model that included prevalence for each of the four test and reference standard threshold combinations. A significant effect was observed only for the HSIL/CIN 2-3 threshold. The effect of quality score on the sensitivity, specificity and effectiveness was investigated; no association was found in any of the models investigated.

The subgroup analysis of studies (n=3) that identified patients undergoing initial Pap smear screening and verified all (or a random fraction of) test negative participants showed a pooled sensitivity of 0.51 (95% CI: 0.37, 0.66) and a pooled specificity of 0.98 (95% CI: 0.97, 0.99). The summary effectiveness score was 2.19 (95% CI: 1.66, 2.71). The prevalence of disease ranged from 10 to 19%.

Cost information
An extensive cost analysis was presented in the report. The cost-effectiveness ratios from published models comparing Pap smear screening with no screening fell into an acceptable range, but these models used parameters that overstated Pap test accuracy.

Authors' conclusions
Estimates of sensitivity of the conventional Pap test are biased in most studies; based on the least biased studies, sensitivity is near 50%, much lower than generally believed. Newer technologies improve sensitivity compared with conventional Pap screening; however, there are no precise estimates for their effect on specificity.

CRD commentary
This was a good review of the area. The aims and the inclusion criteria were clearly stated. A comprehensive literature search was conducted and attempts were made to locate unpublished studies. However, by restricting the search to English language articles some relevant studies might have been omitted. The methods used to select the primary studies and develop validity criteria were described. Appropriate study details were presented in tabular format. The statistical pooling methods used were appropriate for the type of data available. Possible sources of heterogeneity were investigated in a regression analysis. The evidence presented appears to support the authors' conclusions.

Implications of the review for practice and research
Practice: The authors stated that studies of Pap screening in low-prevalence samples found high specificity, but lower than generally believed sensitivity estimates.

Research: The authors stated that future decision models, cost-effectiveness studies, and health policy decisions should consider in their analyses the lower than generally believed sensitivity estimates found in this review.

Funding
Agency for Health Care Policy and Research, contract number 290-97-0014.

Bibliographic details
Other publications of related interest
1. Chock C, Irwig L, Berr G, Glasziou P. Comparing dichotomous screening tests when individuals negative on both
V, Hickey JD, et al. Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic
abnormalities: a systematic review. Ann Intern Med 2000;132:810-9. This additional published commentary may also
be of interest. Melnikow J, Nuovo J. Review: newer technologies improve sensitivity in detecting uterine cervical
lesions but at a substantially increased cost. Evidence-Based Medicine 2000;5:124.

Indexing Status
Subject indexing assigned by CRD

MeSH
Cervix Uteri /cytology /pathology; Diagnosis, Computer-Assisted; Female; Mass Screening /methods /standards;
Sensitivity and Specificity; Uterine Cervical Neoplasms /diagnosis; Vaginal Smears /methods /standards

AccessionNumber
12000008339

Date bibliographic record published
31/03/2005

Date abstract record published
31/03/2005

Record Status
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.