

NEW ZEALAND HEALTH TECHNOLOGY ASSESSMENT (NZHTA)
THE CLEARING HOUSE FOR HEALTH OUTCOMES AND
HEALTH TECHNOLOGY ASSESSMENT

Department of Public Health and General Practice
Christchurch School of Medicine, Christchurch, New Zealand

Effectiveness and cost effectiveness of automated and semi-automated cervical screening devices

A systematic review of the literature

Marita Broadstock

This report should be referenced as follows:

Broadstock M. Effectiveness and cost effectiveness of automated and semi-automated cervical screening devices: A systematic review. *NZHTA Report* 2000; **3**(1).

2000 New Zealand Health Technology Assessment Clearing House (NZHTA)
ISBN 1-877235-13-X
ISSN 1174-5142

CONTRIBUTORS

NZHTA staff

This report was developed by the staff of NZHTA. The review was undertaken by Ms Marita Broadstock (Research Fellow) who conducted the critical appraisals, wrote the report and coordinated the project. Mrs Susan Bidwell (Information Specialist) developed and undertook the search strategy, coordinated retrieval of documents and managed the Endnote bibliographic library. Dr Ray Kirk (Director) provided critical input, participated in project meetings with consultants and provided peer review of report drafts. Dr Barbara Nicholas (formerly Research Fellow at NZHTA) contributed to protocol development, and Dr Phil Hider (Senior Research Fellow, NZHTA) provided helpful comments on methodological issues.

Ms Cecilia Tolan (NZHTA Administrator) provided administrative support and document formatting. Clerical assistance was provided by other NZHTA staff including Miss Becky Mogridge, Ms Tracy Smitheram and Mrs Joan Downey. Sub-editing was performed by Mr Ainslie Talbot (Medical Journalist).

Project consultants

The following experts acted as consultants¹ throughout the project, providing critical input and guidance on technical and methodological issues, participating in project meetings, and providing peer review of various drafts of the report:

- Dr Peter Fitzgerald, pathologist, Southern Community Laboratories, Dunedin. Representative on the Cytology Disciplinary Advisory Committee of the Royal College of Pathologists of Australasia (RCPA).
- Dr Terri Green, health economist, Senior Research Fellow, Department of Public Health and General Practice, Christchurch School of Medicine, and also Senior Lecturer, Department of Management, University of Canterbury. Dr Green is also a member of the Health Funding Authority's Advisory Group for Population Based Screening Programmes.
- Dr Ann Richardson, epidemiologist, Senior Lecturer, Department of Public Health and General Practice, Christchurch School of Medicine. Member of the Independent Monitoring Group for the National Breast Screening Programme.

Dr Clint Teague, pathologist, Wellington Medical Laboratory, acted as a consultant during the development of the proposal for this review, as well as providing expert advice during deliberations concerning appropriate thresholds for reporting outcomes.

Editorial review

We are grateful for editorial review provided by the following external consultants:

- Professor Les Irwig (Professor in Epidemiology, Department of Public Health and Community Medicine, University of Sydney). Professor Irwig also provided invaluable advice relating to modifications to the hierarchy of evidence used for studies evaluating liquid-based slide preparation devices.
- Dr Gabriele Medley (pathologist, Victorian Cytology Service, Melbourne).
- Mr Des O'Dea (health economist, Wellington School of Medicine, Wellington).

¹ All consultants warranted that at the time of their involvement in the project, they held no commercial interest in the technologies connected with this review.

ACKNOWLEDGEMENTS

This review was commissioned by Dr Julia Peters, Prevention Programmes Manager, on behalf of the New Zealand Health Funding Authority.

The following people are thanked for their assistance:

- Ms Ruth Herbert (New Zealand Health Funding Authority) and Dr Julia Peters (Prevention Programmes Manager, HFA) provided background material for the review, including government reports;
- Mr Christopher N. Bowden, Section Head, Cytology Laboratory, Canterbury Health Laboratories, Christchurch), provided a tour of the Laboratory as well as access to Acta Cytologica;
- Dr Bill Mackey (Medical Director, Biotek, NZ distributors for ThinPrep), and Fiona Diversi and Ann-louise Weaver (Dade Behring, Australasian distributors for AutoCyte Prep and AutoPap) provided information upon request about the availability of devices in New Zealand.

Many researchers responded to queries (usually by e-mail) relating to their research; many also engaged in additional debate concerning methodological issues. We are grateful to the following individuals:

- Dr Ulrik Baandrup, Consultant, University Institute of Pathology, Aarhus University Hospital, Denmark;
- Mr Adalsteinn D. Brown, Assistant Professor, Department of Health Administration, University of Toronto, Toronto, Canada;
- Dr Tony Bierre, pathologist, Diagnostic Laboratory, Auckland, New Zealand.
- Dr Damian Coburn, Director, Health Technology Assessment, Medical Services Advisory Committee (MSAC), Canberra, Australia;
- Professor Jack Cuzick, Imperial Cancer Research Fund, London, UK;
- Mr Gary W. Gill, Senior Science Advisor, Diagnostic Cytology Laboratories, Inc, Indianapolis, USA;
- Dr Tony Hanselaar, Associate Professor, Department Pathology, University Medical Centre Nijmegen, The Netherlands;
- Dr Chris Hyde, Director of Aggressive Research Intelligence Facility (ARIF), and Senior Lecturer in Public Health, Dept of Public Health & Epidemiology, University of Birmingham, UK;
- Dr Douglas C. McCrory, Assistant Professor of Medicine, Duke University Medical Center, Durham, USA;
- Dr Chris Meijer, Chairman and Director, Dept of Pathology, Vrije Universiteit, Amsterdam, The Netherlands;
- Dr Evan Myers, Assistant Professor, Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, USA;
- Dr Kavita Nanda, Associate Medical Director, Clinical Research Department, Family Health International, Durham, USA;
- Dr George F. Sawaya, Assistant Professor, Departments of Obstetrics, Gynecology and Reproductive Sciences, Epidemiology and Biostatistics, University of California, San Francisco, USA;
- Ms Karyn Tappe, Senior Research Analyst, Health Technology Assessment Information Service, ECRI, USA.

DISCLAIMER

NZHTA takes great care to ensure the information supplied within the project timeframe is accurate, but neither NZHTA, the University of Otago, or the contributors involved can accept responsibility for any errors or omissions. The reader should always consult the original database from which each abstract is derived along with the original articles before making decisions based on a document or abstract. All responsibility for action based on any information in this report rests with the reader. NZHTA and the University of Otago accept no liability for any loss of whatever kind, or damage, arising from reliance in whole or part, by any person, corporate or natural, on the contents of this report. This document is not intended as personal health advice. People seeking individual medical advice are referred to their physician. The views expressed in this report are those of NZHTA and do not necessarily represent those of the University of Otago, New Zealand Ministry of Health or the New Zealand Health Funding Authority.

NZHTA is a Research Unit of the University of Otago and funded under contract by the New Zealand Health Funding Authority and the New Zealand Ministry of Health.

CONTACT DETAILS

New Zealand Health Technology Assessment (NZHTA)
The Clearing House for Health Outcomes and Health Technology Assessment
Department of Public Health & General Practice
Christchurch School of Medicine
PO Box 4345
Christchurch
New Zealand
Tel: +64 3 364 1152 Fax: +64 (3) 364 1152

Email: nzhta@chmeds.ac.nz

Web Site: <http://nzhta.chmeds.ac.nz/>

EXECUTIVE SUMMARY

Objectives

1. To systematically review the international evidence for *clinical effectiveness* (primarily sensitivity and specificity) and *cost effectiveness* of introducing automated and semi-automated devices available for cervical screening in New Zealand in place of conventional testing.
2. To consider the above evidence in terms of its applicability to New Zealand's population-based cervical screening programme.

The review aimed to update the report produced by the Australian Health Technology Advisory Committee (1998).

Data sources

The literature was searched using the following databases: Medline, Embase, Healthstar, Current Contents, Science Citation Index, Cancerlit and Econlit. Other electronic and bibliographic sources searched included: Cochrane Library, Database of Abstracts of Reviews of Effectiveness, NHS Economic Evaluation database, Health Technology Assessment database, US National Library of Medicine, North Thames Regional Library (UK), World Health Organisation. In New Zealand, databases were accessed from the National Bibliographic Database, Ministry of Health website and library, university and medical library catalogues and the NZHTA in-house collection. Several Internet websites were also searched. "Grey" (unpublished) literature not accessed from the above sources was sought through personal contact with staff in the Health Funding Authority. Material referenced in publications obtained in the course of research on the topic was identified.

Searches were limited to English language material from January 1 1997 to May 31 2000, to update the search dates of the AHTAC review, which was completed in July 1997.

Study selection

Studies were included if they compared the clinical effectiveness, or cost effectiveness, of new devices with conventional Pap testing methods. Technologies considered were automated or semi-automated devices suitable for ready introduction into the New Zealand cervical screening programme. Specifically, these devices included two liquid-based slide preparation methods, ThinPrep and AutoCyte Prep, and a semi-automated imaging device for primary screening and rescreening, AutoPap.

Included studies required a reference standard for verification of cytological diagnoses of either histology (usually by biopsy), or cytology by adjudicated panel review by cytology professionals. Economic studies were included which investigated the effect of screening by the new device on life expectancy or quality, the number of cases of cervical cancer avoided, or total health care costs, compared to conventional Pap testing in a three year screening cycle. Systematic reviews or meta-analyses were also appraised, principally as background information.

Excluded studies included abstracts, single case studies, studies with poor description of methods or results.

Of over 700 articles identified by the search strategy, 58 articles were retrieved as full text from which a final group of 26 papers were identified as eligible for inclusion in the review. Of these, 20 reported primary research (15 on *clinical effectiveness* and five on *cost effectiveness* of new devices), and six papers reported secondary research (in addition to economic primary research studies already identified that also had systematic review components).

Data extraction and synthesis

A systematic method of literature searching, appraising and grading was employed in the preparation of this report. Systematic reviews and meta-analyses were described and critiqued in terms of their search strategy, as well as their criteria for inclusion/exclusion, data synthesis and interpretation of papers considered. Papers of clinical effectiveness were coded on aspects of study quality including recruitment of study sample, blind verification, reference standard employed, extent of verification, and whether industry provided financial support. These papers were ranked in the evidence tables according to a “hierarchy of evidence” indicating design quality. Economic studies of cost effectiveness were described and appraised in terms of their design, outcomes, data sources, assumptions, limitations, key results and sensitivity of the model to value changes in variables.

Key outcomes for reports of clinical effectiveness included, where possible, test sensitivity (Se), test specificity (Sp), relative true positive rate (TPR), relative false positive rate (FPR), and positive predictive value (PPV) at a threshold of HSIL+ for both cytology and reference standard. This threshold was chosen as of clinical importance in a population based screening programme given that they have the highest likelihood of progression to invasive tumour if undetected. Key outcomes for economic studies included cost-effectiveness ratios.

Key results

Clinical effectiveness

Only 15 eligible papers were identified reporting on the clinical effectiveness of new devices compared with the conventional screening: ThinPrep=10, AutoCyte Prep=3, AutoPap=2. Many studies were excluded because there was no or an inadequate reference standard (e.g. single pathologist review), they lacked a comparison group of manual screening, or they were not used as the device was intended (e.g. AutoPap used for high-risk slides). The majority (9/15) of studies appraised were at least partially funded by the industry producing the devices considered.

There were no randomised controlled trials using an outcome of invasive cancer incidence or mortality. Whilst there was some evidence that new devices may marginally increase detection of low grade abnormalities, estimates of test sensitivity and specificity could not be reliably determined from the current evidence base. Studies were severely limited by design, inadequate reference standards, and incomplete verification of cytological diagnoses. Moreover, there was no reliable evidence for improved detection of high-grade abnormalities by semi-automated and automated devices for cervical screening.

Liquid-based slide preparation (ThinPrep and AutoCyte Prep)

Six systematic reviews and/or meta-analyses were identified and 15 primary research studies appraised relevant to liquid-based slide preparation. All used histology by biopsy as their reference standard to verify positive diagnoses, though no study verified negatives at a threshold of HSIL. Verification of positives was commonly limited to test results which were discordant (i.e. one screening test gave a positive result and the other gave a negative result). This leads to a lack of evidence on specificity and inflated estimates of sensitivity.

Given these and other limitations in study quality, the clinical effectiveness of ThinPrep and AutoCyte Prep for detection of high-grade abnormalities cannot be reliably determined from the current evidence base. Moreover, it is not possible to say whether one device has advantages over another in terms of considered outcomes. Valid estimates of test sensitivity and specificity of these devices await further research employing better designs.

Semi-automated devices for primary screening and re-screening

Six systematic reviews and meta-analyses were identified and only one primary research study (reported in two papers) was appraised relevant to AutoPap. This prospective trial evaluated the AutoPap System as a combined primary screener and rescreener. There was only limited verification of cytological diagnoses and therefore the study did not permit direct estimates of test sensitivity and specificity. Instead, indirect estimates verified by panel cytology review of discordant test results (leading to an overestimation of diagnostic performance) were reported. Results suggest that whilst there may potentially be increases in detection of low grade abnormalities for AutoPap compared with

conventional screening followed by 10% random rescreening, there is no evidence to suggest an increase in detection of high grade abnormalities. There was inadequate evidence concerning the specificity of AutoPap. To be relevant to New Zealand's rescreening practices (which usually include targeted full rescreening of high risk slides, as well as rapid rescreening of all negative smears) comparisons with alternative rescreening strategies to random 10% review are recommended.

Cost effectiveness

Six economic studies including the AHTAC review were appraised, all of which were *disease state transition models* which track movement of a hypothetical cohort of women between health states (representing progression of disease) over repeated screening intervals until death (the AHTAC review only considered a single screening phase). All cost effectiveness models were severely limited by the uncertainty surrounding estimates for improved sensitivity and the lack of information on changes to specificity that may occur with the introduction of new devices into screening programmes. When improved detection at all grades of abnormality was assumed, the impact of new devices on days-of-life saved was extremely small for women screened at three yearly intervals. As most (assumed) additional abnormalities found are low-grade, these are likely to regress, or if they persist are very likely to be detected at the next regular screen. Given the very slow growth of cervical cancer from pre-cancerous abnormalities which progress, women screened regularly at laboratories meeting minimal quality standards will, in the vast majority of circumstances, have any abnormalities missed at one screen detected at a subsequent screen and potentially treated before cancer develops.

The possibility that new devices may decrease specificity, and therefore increase false positive diagnoses, has not been comprehensively evaluated in cost effectiveness models. False positives are likely to have a significant impact on quality of life in terms of a screened woman's inconvenience, discomfort and distress arising from unnecessary investigations. This would drastically reduce the cost effectiveness of screening devices that reduce specificity.

Conclusions

The following conclusions are based on the current evidence available from this report's systematic reviews of literature published on the clinical and cost-effectiveness of new devices for population cervical screening (i.e. semi-automated and automated devices: ThinPrep, AutoCyte Prep, and AutoPap). Note that the impact of Human Papilloma Virus (HPV) DNA testing (to distinguish a low-risk group who require diminished surveillance) has not been considered in this review.

1. Estimates of test sensitivity and test specificity for the new devices could not be reliably determined. The research reviewed here provides no evidence for improved detection of high-grade abnormalities by new devices for cervical screening. New devices should demonstrate clinical effectiveness gains in detecting higher grade abnormalities. High-grade squamous intraepithelial lesions have a much higher probability of progressing to cancer than low grade abnormalities that are likely to regress.
2. Estimates of test sensitivity and specificity were the main source of uncertainty in the economic models investigating the cost effectiveness of new devices. In economic models where improved detection from the introduction of new devices was assumed, the impact of new devices on days-of-life saved was extremely small for women screened at three yearly intervals. Cost effectiveness may be even poorer in New Zealand where more effective rescreening practices are employed for conventional screening.
3. Any increases in sensitivity resulting from the introduction of new devices may come at the cost of decreased specificity. This would lead to increases in false positive results (where slides are read as abnormal when the woman does not have a cervical abnormality). False positives lead to health sector costs of unnecessary diagnosis, treatment and follow-up that may lead to pressure on health services to the detriment of women with true abnormalities. Investigations of false positives also are associated with social and psychological costs for women including inconvenience, discomfort and distress. The potential for increases in false positives would, if realised, have a profound impact on quality of life and the related cost effectiveness of the devices.
4. Higher quality research is required to generate valid estimates of test sensitivity and specificity. Methodological limitations to address include the application of appropriate reference standards for verification of cytological diagnoses, including test negatives. Economic modelling studies will be

more meaningful with more valid estimates of test characteristics, and a comprehensive measurement of costs of screening from a societal perspective, including careful investigation of the impact of screening and clinical management on quality of life.

5. It is important that promotional information for new devices is balanced by material for health professionals and for women based on key findings of independent evidence such as found in this report. Additionally, the New Zealand Health Funding Authority/Ministry of Health should investigate legal avenues to restrict advertising making unsubstantiated claims for new devices.
6. The vast majority of missed abnormalities will be detected at subsequent screens for women who are routinely screened appropriately, assuming acceptable levels of smear taking and laboratory performance. Three yearly cervical screening using the conventional Pap smear can be highly effective, preventing 93% of cervical cancer, assuming all eligible women are screened. Therefore, the Pap test should remain the standard of care in population cervical screening.
7. The introduction of new devices for cervical screening cannot be recommended for the New Zealand national cervical screening programme at this time.
8. Rather than committing resources to the introduction of new devices into the national screening programme, better outcomes may be achieved for women screened if resources are directed to other ways of improving the programme. These strategies could include the following:
 - *increasing up-take of routine screening by eligible women,*
 - *ensuring that women are screened at appropriate intervals,*
 - *implementing standards for smear taking, and ensuring the use of the most effective smear taking instruments,*
 - *implementing strict laboratory standards and quality assurance,*
 - *and ensuring adequate follow-up and treatment where required.*
9. Resources should be directed to the appropriate monitoring of the national cervical screening programme.
10. This review is based on research published to end of May 2000. It is recommended the conclusions of this report be revisited in 12 months (October 2001).

MeSH Headings

cervix neoplasms, vaginal smears, mass screening, cytological techniques, image processing-computer assisted, cervix neoplasms-diagnosis,

Additional key words

thinprep, papanicolaou, autopap, papnet, autocyte prep, autocyte screen, cytorich, rescreen*, ((vagina* or cervi*) near (screen* or smear*)), cervi* near cytology, pap* near (screen* or smear*)

TABLE OF CONTENTS

CONTRIBUTORS	i
<i>NZHTA staff</i>	i
<i>Project consultants</i>	i
<i>Editorial review</i>	i
ACKNOWLEDGEMENTS	ii
DISCLAIMER	iii
CONTACT DETAILS	iii
EXECUTIVE SUMMARY	iv
<i>Objectives</i>	iv
<i>Data sources</i>	iv
<i>Study selection</i>	iv
<i>Data extraction and synthesis</i>	v
<i>Key results</i>	v
<i>Conclusions</i>	vi
<i>MeSH Headings</i>	vii
<i>Additional key words</i>	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	xi
LIST OF FIGURES	xi
CHAPTER 1: INTRODUCTION	1
1.1 BACKGROUND	1
<i>The National Cervical Screening Programme</i>	1
<i>Conventional practice for cervical screening in New Zealand</i>	1
<i>Limitations of cervical screening</i>	4
<i>Cost effectiveness</i>	5
<i>Liquid-based slide preparation devices</i>	5
<i>Semi-automated slide analysis devices</i>	6
<i>Diffusion of devices in New Zealand</i>	6
<i>Need for proposed Systematic Review</i>	8
1.2 REVIEW SCOPE	8
1.3 OBJECTIVES	8
1.4 STRUCTURE OF REPORT	9
CHAPTER 2: METHODOLOGICAL ISSUES	11
2.1 STUDY DESIGNS	11
<i>Both tests applied to the same woman</i>	11
<i>Tests randomly allocated to different women</i>	12
<i>Tests non-randomly allocated to different women</i>	12
2.2 SPECTRUM OF DISEASE IN STUDY POPULATION	13
2.3 SAMPLING DEVICE	14
2.4 CYTOLOGIST ASSESSMENT OF SLIDES	14
2.5 APPROPRIATE COMPARISON TESTS FOR RESCREENING	15
2.6 REFERENCE STANDARD	15
<i>Need for a reference standard</i>	15
<i>Histology</i>	16
<i>Panel review by expert cytologists</i>	16
<i>Acceptable reference standards for this review</i>	17
2.7 VERIFICATION BIAS	17
2.8 REPORTING OF OUTCOMES	18
<i>Sensitivity and specificity</i>	18
<i>Thresholds for reporting positive diagnoses</i>	18
<i>Outcomes reported in this review</i>	19
2.9 COMPARISONS BETWEEN STUDIES	20
<i>Interdependence of primary screening and rescreening</i>	20
<i>Industry funding</i>	21
CHAPTER 3: METHODOLOGY	23
3.1 STUDY SELECTION	23

	<i>Study Inclusion criteria</i>	23
	<i>Study Exclusion Criteria</i>	24
	<i>Excluded emerging technologies</i>	24
3.2	SEARCH STRATEGY	25
	<i>Principal sources of information</i>	25
	<i>Search terms used</i>	26
3.3	SELECTION AND APPRAISAL.....	27
3.4	APPRAISAL OF STUDIES	27
3.5	KEY OUTCOME MEASURES FOR PRIMARY STUDIES	28
3.6	METHODOLOGICAL QUALITY OF PRIMARY STUDIES	29
3.7	LIMITATIONS OF THE REVIEW	30
CHAPTER 4: EFFECTIVENESS OF LIQUID-BASED SLIDE PREPARATION DEVICES		31
4.1	THINPREP.....	31
	<i>Secondary research</i>	31
	<i>Primary research: Study designs and quality assessments</i>	39
	<i>Primary research: Study results</i>	39
	<i>Conclusions</i>	40
4.2	AUTOCYTE PREP.....	51
	<i>Secondary research</i>	51
	<i>Primary research: Study designs and quality assessments</i>	51
	<i>Primary research: Study results</i>	51
	<i>Conclusions</i>	51
CHAPTER 5: EFFECTIVENESS OF AUTOMATED DEVICES FOR PRIMARY SCREENING AND RE-SCREENING		57
5.1	AUTOPAP	57
	<i>Secondary research</i>	57
	<i>Primary research: Study results</i>	64
	<i>Conclusions</i>	65
CHAPTER 6: ECONOMIC EVALUATIONS OF CERVICAL SCREENING DEVICES		69
6.1	INTRODUCTION	69
	<i>Study designs</i>	69
	<i>Markov models</i>	69
	<i>Assumptions</i>	70
	<i>Sensitivity analyses</i>	70
	<i>Outcomes</i>	70
	<i>Health system perspective</i>	71
6.2	STUDY RESULTS.....	71
	<i>AHTAC (1998)</i>	71
	<i>Brown and Garber (1998, 1999)</i>	72
	<i>McCrorry et al. (1999)</i>	74
	<i>Smith, Lee, Leader and Wertlake (1999)</i>	75
	<i>Payne, Chilcott & McGoogan (2000)</i>	77
6.3	DISCUSSION.....	79
	<i>Costs of new devices uncertain</i>	79
	<i>Uncertainty in sensitivity estimates</i>	79
	<i>Uncertainty in specificity estimates</i>	80
	<i>Impact of screening on quality of life</i>	80
	<i>Conclusions</i>	81
CHAPTER 7: DISCUSSION		89
7.1	SUMMARY OF EVIDENCE	89
	<i>Clinical effectiveness of new devices</i>	89
	<i>Cost effectiveness of new devices</i>	90
7.2	FUTURE RESEARCH	91
	<i>Clinical effectiveness research</i>	91
	<i>Cost effectiveness research</i>	92
	<i>Future developments</i>	92
7.3	IMPLICATIONS OF RESULTS FOR THE NATIONAL CERVICAL SCREENING PROGRAMME	93

<i>Conventional screening can be highly effective</i>	93
<i>Impact of introducing new devices into a population based cervical screening programme.....</i>	94
<i>Alternative ways of improving effectiveness of cervical screening</i>	95
CONCLUSIONS	99
REFERENCES	101
LIST OF ABBREVIATIONS AND ACRONYMS	109
GLOSSARY	111
APPENDIX 1	115
CALCULATION OF TEST CHARACTERISTICS	115
APPENDIX 2	117
CALCULATION OF RELATIVE TRUE POSITIVE RATES AND RELATIVE FALSE POSITIVE RATES	117
APPENDIX 3	119
SEARCH STRATEGIES	119
<i>Medline</i>	119
<i>Healthstar</i>	119
<i>Embase</i>	120
<i>Current Contents</i>	121
APPENDIX 4	123
RETRIEVED STUDIES EXCLUDED FROM REVIEW.....	123
APPENDIX 5	125
ADDITIONAL BACKGROUND PAPERS	125

LIST OF TABLES

Table 1.	Secondary research appraised relevant to liquid-based slide preparation devices.....	33
Table 2.	Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep	42
Table 3.	Evidence table of primary research studies investigating liquid-based slide preparation devices – AutoCyte Prep	53
Table 4.	Secondary research appraised relevant to AutoPap.....	60
Table 5.	Evidence table of primary research investigating AutoPap	67
Table 6.	Economic evaluations appraised.....	69
Table 7.	Evidence table for economic evaluations	82
Table 8.	Possible impact of the introduction of new devices into a population based cervical screening programme.....	98

LIST OF FIGURES

Figure 1.	Various classification schemes for cervical cytology.....	3
Figure 2.	Alternate pathways for cervical screening at the laboratory: conventional testing and new devices.....	7

Chapter 1: Introduction

1.1 BACKGROUND

The National Cervical Screening Programme

In 1985 a working group recommended routine cervical screening for New Zealand to replace the opportunistic screening then available (Skegg et al., 1985). In 1988, the Cartwright Inquiry investigated allegations concerning treatment of cervical cancer at the National Women's Hospital. The resultant report recommended the urgent establishment of a population-based cervical screening programme (New Zealand Committee of Inquiry into Allegations Concerning the Treatment of Cervical Cancer at National Women's Hospital and into Other Related Matters, 1988). Subsequently, the New Zealand National Cervical Screening Programme was launched in 1990. Over the next decade, various working parties, expert groups and ministerial reviews contributed to the development of Government policy on cervical screening (Adams, 1991; Members of the Working Party on Cervical Screening, New Zealand, 1998; New Zealand Ministry of Health, 1993; New Zealand Department of Health, 1991; New Zealand Ministry of Health, 1996; Paul et al., 1991). Guidelines for the management of women with abnormal cervical smears were released (New Zealand Health Funding Authority, 1999). This was followed by a draft document on policy and quality standards (New Zealand Health Funding Authority, 1999c), an evaluation and monitoring plan (New Zealand Health Funding Authority, 1999a) and proposed national indicators for the National Cervical Screening Programme New Zealand (New Zealand Health Funding Authority, 1999b).

Current recommendations for cervical screening include three yearly screening of women aged between 20 and 69 years (Members of the Working Party on Cervical Screening, New Zealand, 1998). If it is a woman's first smear or there is a gap of five years or more since her last one, the second smear should take place a year later. Women who have a hysterectomy for a benign condition with complete removal of histologically normal cervical epithelium and a clear smear history do not require further screening.

Local coordination and delivery of the National Cervical Screening Programme are provided from 14 sites around New Zealand. The Health Amendment Act (1993) requires that the results of all cervical cytology (and histology) tests be forwarded to the National Cervical Screening Register unless a woman "opts off". The Register aims to identify groups who are not being reached, track individuals who move to another provider, recall women who are overdue for a cervical smear, and has the potential to provide data for monitoring, quality assurance and evaluation. Following concern amongst Māori women about the control of that register and access to the data, a Kaitiaki group was established to act as guardian to data on Māori women (New Zealand Ministry of Health, 1997).

Conventional practice for cervical screening in New Zealand

Aim of the Pap test

The conventional screening test used in the National Cervical Screening Programme is the Papanicolaou smear ("the Pap test"). The Pap test is a cytological screening test that aims to detect precancerous changes in the cervical epithelium (cervical intraepithelial neoplasia or CIN) which can be detected and treated before invasive squamous cell carcinoma develops. That is, its main objective is to prevent cervical cancer. Squamous cell carcinoma comprises 85% of all cervical cancers and is more preventable by the Pap test than adenocarcinoma and mixed adeno-squamous carcinoma (Australian Health Technology Advisory Committee, 1998). It is important to understand that not all detected abnormalities will develop into cancer. Lower grade abnormalities (CIN I) regress in about 60% of cases, persist in 30% and progress to major precancerous changes in about 10% of cases. Only 1% progress to true invasive cancer. This progression tends to take place over a long time, averaging 12 years from progression from minor to major changes, and five years from a major precancerous change to invasive carcinoma (Östör, 1993). By comparison, 40% and 33% of higher grade

abnormalities (CIN II and III, respectively) regress and 5% and 12%, respectively, progress to invasion. The classification of grades of abnormality will be explained shortly.

Taking the smear

The test involves taking a Pap smear; that is, a sample of cells from the cervix collected at the “transition zone” of the ectocervix and endocervix, providing cells from both areas. A general practitioner, nurse or trained lay smear taker takes the cervical smear using a broom-type device, or combination of spatula and cyto-brush. Smear taking is accompanied by a visual inspection of the lower genital tract. Cells are transferred onto a slide then sprayed with, or immersed in, a fixative to preserve the cells and then sent to a pathology laboratory accompanied by a request form that details identification of the woman, the smearer and includes relevant past history and clinical information.

Reading the slides at the laboratory

At the laboratory, the slide is stained using the Papanicolaou method and screened by a cyto-technologist using a microscope. This phase is often termed primary screening. All slides read as abnormal are reviewed by a cyto-pathologist². Slides that are read as negative; that is, clear of any apparent abnormality, are often reported as being “within normal limits” (WNL). Negative slides may be screened a second time according to laboratory quality control (QC) protocols, termed re-screening. Manual methods of rescreening internationally include one or more of the following:

- *full* rescreening,
- *partial* (e.g. 10%) *random* rescreening,
- *targeted* or directed rescreening,
- and *rapid* (the cytologist taking 1 –3 minutes per slide) rescreening.

Generally targeted rescreening of slides occurs where negative slides are fully rescreened when there is significantly abnormal clinical or cytological history (e.g. previous abnormal smears, symptoms such as abnormal bleeding, or a note that the cervix appears abnormal to the smear taker whilst taking the smear). Although strategies for QC rescreening are not consistent across laboratories in New Zealand, they usually consist of targeted full rescreening of high risk slides, as well as rapid rescreening of all negative smears (Dr Peter Fitzgerald, pathologist, March 2000, *personal communication*). Partial random rescreening (e.g. of 10% of negative smears) is not used in New Zealand, but is the QC strategy federally mandated in the USA by Clinical Laboratory Improvement Amendments (CLIA).

The laboratory produces a report for each smear which:

- (a) comments on specimen adequacy (whether the smear is satisfactory, limited, or unsatisfactory);
- (b) may give a general categorisation of whether the smear is within normal limits, LSIL, HSIL, etc;
- (c) provides a descriptive diagnosis where relevant;
- (d) and gives an appropriate recommendation indicating when the next smear should be taken, or if further investigation is indicated.

Various classification schemes may be used for reporting of cytological abnormalities. **Figure 1** maps the various schemes together so that alternative classifications to those described above can be interpreted (note that this a simplification of actual systems).

² For convenience, in this review the term cytologist is generally used to refer to either a cyto-technologist or cyto-pathologist who reads a slide and makes a diagnosis of any abnormality.

Figure 1. Various classification schemes for cervical cytology (from Nanda et al., 2000)

Classification System	Cytology Classification							
The Bethesda System (TBS)	Normal	Infection	ASCUS	Squamous Intraepithelial Lesion (SIL)			Invasive Carcinoma	
Richart		Reactive Repair		Low Grade (LSIL) (including HPV)		High grade (HSIL)		
Reagen (World Health Organisation)		Atypia		Condyloma	Cervical Intraepithelial Neoplasia (CIN)			
					CIN I	CIN II		CIN III
Papanicolaou	I	II		III			IV	V

NOTES: ASCUS: atypical squamous cells of undetermined significance. In the UK, ASCUS/AGUS cells are described as “borderline changes”, HPV is borderline rather than ASCUS, and the term dyskaryosis is used instead of dysplasia.

New Zealand has adopted the Bethesda system (International Academy of Cytology, 1992) for reporting of cytological abnormalities with modifications as approved by the National Cervical Screening Programme. The descriptive categorisation includes the following terms:

- benign cellular changes (including infection, inflammation, reactive repair),
- atypical squamous cells of undetermined significance (ASCUS),
- squamous cell abnormalities (SIL) which are divided into two sub-categories:
 - (i) low grade squamous intraepithelial lesions (LSIL) which includes CIN I (HPV)
 - (ii) high grade squamous intraepithelial lesions (HSIL) which includes CIN II and CIN III
- invasive squamous cell carcinoma (SCC); and
- glandular cell abnormalities.

Clinical diagnosis and management

The report of screening results is provided to the referring practitioner and appropriate management/monitoring is scheduled where necessary. Definitive diagnosis of a lesion is made by histological examination of cervical tissue obtained by biopsy following colposcopy. Depending on the degree of abnormality, women can be managed with more frequent screening or treated. Recent guidelines recommend that LSIL detected by screening be followed by a repeat smear in six months (Jones et al., 2000). If the abnormality persists or progresses at the repeat smear, a colposcopy is advised to visualise the cervix with any visualised lesions biopsied. When a patient receives a diagnosis of HSIL or higher on a Pap smear, she undergoes colposcopy and possible biopsy. Treatment options after a positive biopsy may include cervical excision, ablation, freezing of the transformation zone, and for invasive cancer, hysterectomy, radiotherapy and/or chemotherapy.

Limitations of cervical screening

The cervical screening programme has been effective in reducing mortality and incidence of invasive cervical cancer in New Zealand. However, like any cancer screening programme, some women screened will still develop invasive cancer. In countries with organised screening programmes, the largest category of women to be diagnosed with invasive cervical cancer, or to die from invasive cervical cancer is women who have never been screened. The next largest category is women whose abnormal smears have not been adequately followed up, then women with a long interval between smears, and finally women with false negative smears (Chamberlain, 1986).

False negative smears can arise when a woman is screened as being without cervical abnormality (“negative” or “clear”) when she actually has abnormal cells in the cervix. Screening errors can occur at three stages for a woman once she decides to have a routine smear: *smear taking* affecting smear sample quality, *preparation* of the slide from the smear, and *reading* of the slide at the laboratory to detect abnormalities (during primary screening and rescreening). About two-thirds of false negatives are a result of errors in sampling (taking the sample, and preparing the slide) and the remaining one-third relate to detection error (McCrorry et al., 1999)³

Concerns about the Pap test have tended to focus on false negatives, perhaps in part because of media coverage of prominent cases of alleged misreading of abnormal slides, sometimes with devastating results for the women involved. Most recently in New Zealand, a Ministerial Inquiry has investigated a Gisborne pathologist’s screening; rescreening of 23,000 slides suggested that there had been a substantial under-reporting of high grade abnormalities (<http://www.csi.org.nz>). In the UK, the High Court upheld a ruling that three women who developed cancer despite negative smear tests were victims of medical negligence (Dyer, 1999). Increased litigation, particularly in the United States, relating to false negatives as well as competitive laboratories seeking ways of increasing market share, have also contributed to a call for ways of improving test sensitivity.

³ There are also “false negatives” where the smear is normal but the woman develops abnormal cells soon afterwards. Strictly speaking the *smear* was read correctly, but the screening *programme* failed to detect the abnormality, because it occurred in the interval between screens.

False negatives are minimised when a screening test's "sensitivity" is maximised; that is, the probability of receiving a positive test result in the presence of true abnormality. Once a smear has been taken, low screening test sensitivity leads to a high *false negative rate* (FNR) where slides that do include abnormal cells have been diagnosed as negative or within normal limits (WNL). Typical FNR is estimated at being around 10%, and 3-5% is generally accepted as irreducible (Eddy, 1990), though a standard and acceptable FNR has yet to be defined by the medical profession for a routine Pap test (Davey, 1997).

Whilst there have been concerns about the Pap test's sensitivity, the other aspect of test accuracy, "specificity" appears to be very high. Specificity refers to the probability of a negative test result in the presence of no abnormality. Low specificity leads to a high *false positive rate* (FPR) where slides including *no* abnormal cells have been diagnosed as abnormal or positive. A recent systematic review of Pap test accuracy reported sensitivity at a threshold of LSIL+ ranging from 30% – 87% (mean=47%) and specificity ranging between 86% – 100% (mean=95%) (Nanda et al., 2000)⁴.

Automated and semi-automated devices (also referred to in this report as "new devices") have been developed to address screening errors relating to the Pap smear's slide preparation and/or slide reading. Some of these devices are already being actively promoted in New Zealand, and they are the subject of this review. Thinprep™ (Cytoc Corporation) and AutoCyte Prep™ (TriPath Imaging) are both automated liquid-based slide preparation systems designed to provide more representative cell samples of evenly dispersed cells. The AutoPap Primary Screening System™ (TriPath Imaging) is an automated cervical screening device for primary screening and quality control rescreening of slides previously screened as negative or WNL. These new devices will be described more fully shortly.

Cost effectiveness

The new devices described above aim to improve the detection of abnormalities. However, their introduction into a population-based screening programme would need to be considered firstly, with respect to whether the devices do improve sensitivity and specificity, and secondly, in the context of the additional benefit to women for the additional cost involved. It is important to recognise that the impact of a false negative on mortality is reduced by regular screening as any abnormality missed at one screen may be detected at a subsequent screen. This is particularly the case for lower grade abnormalities that, as discussed earlier, rarely progress to cancer and where they do, do so over a period of several years. There may be downstream implications of detecting more abnormalities, such as changes to the number of investigations undertaken.

Costs of screening with the new devices are also likely to be greater than for the Pap smear, when considering initial start-up and consumable costs as well as training and maintenance costs for the equipment. False positives result in costs of unnecessary repeat smears and possible investigations, as well as a woman's inconvenience, and potential discomfort and distress from these interventions. There has been concern in some quarters that screening uptake may be reduced, especially for poorer women, if there is a cost to the consumer of new screening devices which is too high (Sawaya and Grimes, 1999) and generally if consumer confidence in the conventional Pap test is undermined. These issues are relevant to studies of the cost effectiveness of new devices.

Liquid-based slide preparation devices

Liquid-based screening (LBS) involves the use of automated liquid-based sample collection and slide preparation. LBS devices include ThinPrep (Cytoc Corporation) and AutoCyte Prep (previously known as CytoRich) (TriPath Imaging). These devices have been designed to provide more representative cell samples of evenly dispersed cells. This involves sample cells, collected in the standard way, being suspended in a fixative solution rather than smeared on a slide. In the laboratory, the solution is dispersed, collected selectively on a filter, and then transferred to a microscope slide for interpretation. This process produces cells in a thin layer (sometimes described as monolayer) which

⁴ These estimates come from nine studies of high standard involving low-risk patients undergoing screening. Cytology diagnoses (including all or a random fraction of negative slides) were verified using a reference standard of histology or negative colposcopy. These aspects of study quality are discussed at length in Chapter 2. Unfortunately, results were not reported at a threshold of HSIL.

aims to reduce material such as blood, pus and mucus from obscuring cells during cytological examination.

ThinPrep Beta and ThinPrep 2000 Processor (a later version) have been the subject of most research studies completed to date. ThinPrep 2000 has a capacity of 50,000 samples per year and was cleared by FDA for marketing as a replacement for conventional Pap smear slide preparation in 1996. In June 2000, ThinPrep 3000 Processor was given FDA premarket approval. ThinPrep 3000 differs from ThinPrep2000 in being able to process samples as unattended batches, and being slightly faster, with a capacity of 140,000 smears per year.

AutoCyte Prep was given FDA premarket approval in June 1999.

Semi-automated slide analysis devices

AutoPap (TriPath Imaging) is a computerised image-processing device. It uses algorithms to select a proportion of slides for manual screening by a cytologist at the microscope. High speed video microscopy, image interpretation software and field-of-view computers are used to analyse and classify the images of a conventional Pap smear (Halford, 1998). In September 1995, AutoPap 300 QC received FDA premarket approval (PMA) as a rescreener of at least 10% of Pap slides previously read as negative or within normal limits (WNL). This device is the subject of most studies published to date evaluating AutoPap. However, AutoPap 300 QC was not rendered Y2K compliant and is no longer commercially available internationally.

The AutoPap 300 QC has been replaced by the AutoPap (Primary Screening) System (referred to here as the AutoPap System) which adds a new algorithm to the AutoPap 300 QC. The AutoPap System received FDA PMA in January 1998 as a primary screener *and* QC rescreener. When applied in both modes, the AutoPap System designates slides in two categories. *No Further Review* slides are those classified as “within normal limits” requiring no manual review (up to 25% of all slides processed). *Review* slides are ranked into quintiles according to their likelihood of being abnormal and then are screened manually by cytologists aware of these relative score rankings (at least 75% of slides processed). Of the *Review* slides subsequently read as within normal limits, at least 15% ranked by AutoPap as most likely to be abnormal are selected for manual QC rescreening or *QC Review*. Some slides fail processing by AutoPap due to physical characteristics or insufficient cellularity. These are called *Process Review* and *ReRun* slides and if they cannot be repaired and reprocessed, need to be read manually.

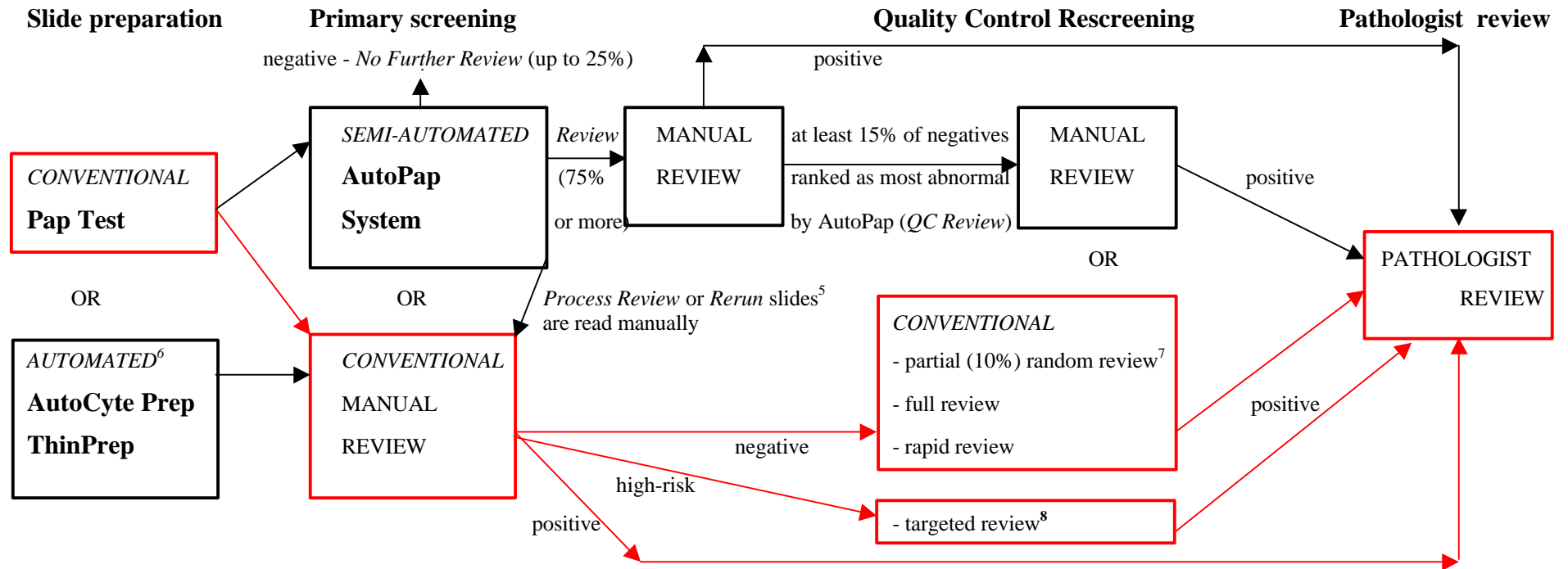
This selected rescreening of WNL slides is an alternative to other rescreening methods (full manual rescreening, partial random rescreening, targeted rescreening, and rapid rescreening). As mentioned earlier, random rescreening of 10% of WNL slides is mandated practice in the USA. This has led to most research evaluating AutoPap devices to compare their performance with 10% random rescreening as the “conventional practice”. However, in New Zealand other rescreening strategies such as rapid rescreening and targeted “high risk” review are conventional practice.

Figure 2 presents how automated (ThinPrep and AutoCyte Prep) and semi-automated (AutoPap System) devices may be used as alternatives to conventional cervical screening at the laboratory.

Diffusion of devices in New Zealand

In New Zealand, approximately 450,000 – 500,000 cervical smears are processed annually. According to New Zealand distributors of ThinPrep, about 17% of these are currently processed using ThinPrep (with materials supplied for about 7000 tests per month). There are seven ThinPrep 2000 Processors in New Zealand, five in the North Island and two in the South Island. ThinPrep smears are available to the woman screened for an additional fee which ranges from \$10 to \$25, and averages about \$15 to \$20 according to industry sources (Dr Bill Mackey, Medical Director of Biotek, New Zealand distributors of ThinPrep devices, May 2000, *personal communication*).

Figure 2. Alternate pathways for cervical screening at the laboratory: conventional testing and new devices



⁵ Process Review or Rerun slides are those that fail processing due to physical characteristics or insufficient cellularity. These are excluded from the total in describing proportions as *No Further Review* or *Review*.

⁶ Supplements for FDA PMA are being sought for use of AutoCyte Prep with the AutoPap System. Cytoc are also developing an image processing system for primary screening of ThinPrep slides.

⁷ This method is not generally used in New Zealand but is federally mandated in the US.

⁸ Slides from women at high-risk for abnormalities must be read manually as AutoPap System is not FDA approved to process these slides in either primary or rescreening modes.

Dade Behring, distributors of TriPath Imaging's products in Australasia, have not actively marketed AutoCyte Prep or AutoPap in New Zealand, to date, but both are commercially available here (Ann-Louise Weaver, Dade Behring (NZ), May 2000, *personal communication*).

Need for proposed Systematic Review

In an effort to consider the effectiveness of new screening devices being available, the Australian Health Technology Advisory Committee (AHTAC) published a comprehensive review entitled "Review of automated and semi-automated cervical screening devices" in April 1998 (Australian Health Technology Advisory Committee, 1998). This reviewed material published between 1990 and July 1997. It concluded that available studies indicated that the new devices decreased the number of significant lesions missed. Moreover the new devices increased the rate of detection of low grade changes, and slide preparation techniques reduced the proportion of slides that could not be interpreted (i.e. "unsatisfactory" or "inadequate" slides) as well as facilitating the reading of slides. However, scientific evidence available was limited at that time and there were also significant cost implications for the introduction of the screening devices into routine practice. The Australian review did not recommend increased uptake of semi-automated or automated devices in a national screening programme, from a public health perspective, at the time of the review.

The development and evaluation of new devices for cervical screening is a rapidly evolving area. There was recognition in the AHTAC report that as devices develop and become more widely used, cost structures would change and cost effectiveness might improve. This review aims to update the AHTAC report in reviewing the international evidence for effectiveness and cost effectiveness of currently available, feasible, automated cervical screening devices. Since publication of the AHTAC report, there have been a number of significant international studies and reviews relating to cervical screening devices.

1.2 REVIEW SCOPE

The present review aims to update the systematic review produced by AHTAC (Australian Health Technology Advisory Committee, 1998) of the clinical effectiveness of currently available semi-automated and automated devices for cervical screening.

Studies were included for review if they report experimental studies of liquid-based slides preparation techniques and devices suitable for primary screening and rescreening suitable for ready introduction into the New Zealand national cervical screening programme, namely, ThinPrep, AutoCyte Prep, and the AutoPap System. Economic evaluations of these new devices were also systematically reviewed. As this review updates the AHTAC review the search was limited to full reports published between January 1997 to May 2000 inclusive, and available in English. Selection criteria relating to aspects of study design (including use of an adequate reference standard) also were employed for studies of clinical effectiveness. Full details of inclusion and exclusion criteria are provided in Chapter 3.

Studies relating to the effectiveness of DNA testing for HPV were excluded from the review.

1.3 OBJECTIVES

1. To systematically review the international evidence for *clinical effectiveness* (primarily sensitivity and specificity) and *cost effectiveness* of introducing automated and semi-automated devices available for cervical screening in New Zealand in place of conventional testing.
2. To consider the above evidence in terms of its applicability to New Zealand's population-based cervical screening programme.

1.4 STRUCTURE OF REPORT

This report includes seven chapters, divided into Sections. Chapter 2 presents the methodological issues relating to research appraised in the review of clinical effectiveness, and informed decisions made concerning the appraisal and reporting of studies. In Chapter 3 the methods and limitations of the review are detailed. The effectiveness of liquid-based slide preparation devices, and primary and re-screening devices, are investigated in Chapters 4 and 5 respectively. Each chapter includes a summary of relevant findings of systematic reviews and meta-analyses, where available. This is followed by a description of the quality of studies reviewed, a summary of results, and discussion of findings. Detailed “evidence tables” are provided which describe appraised papers. Chapter 6 presents the systematic review of analyses of cost effectiveness appraised. Chapter 7 provides an overall summary, a discussion of implications of the review’s findings for New Zealand’s National Cervical Screening Programme, and review conclusions. A glossary and list of abbreviations used in the report are provided prior to the Appendices.

Chapter 2: Methodological issues

In this chapter, methodological issues are discussed relevant to studies of the clinical effectiveness of cervical screening devices. Issues have been identified from review papers, editorials, task force deliberations, working group guidelines, correspondence, etc. Key issues include requirements for new diagnostic or screening tests described by Sackett et al. (1997). These are that tests must:

- (i) have a reference standard for the verification of diagnoses as accurate,
- (ii) evaluate the test in an appropriate spectrum of subjects, and
- (iii) must apply a reference standard regardless of test results.

Issues presented here informed the selection criteria used for papers in the review, as well as codes of study quality, reporting of outcomes, and the “hierarchy of evidence” used to rank studies according to design quality. This chapter also provides a background to how the papers were appraised.

2.1 STUDY DESIGNS

In appraising papers it is useful to rank them in terms of quality according to a pre-determined “evidence hierarchy”, as described in the next chapter. Levels of evidence in this hierarchy relate to key study design elements discussed below.

Both tests applied to the same woman

The most valid design generally considered for diagnostic and screening tests is where both tests are applied to the same woman and compared in a **within-subjects** analysis (Irwig and Glasziou, 1996). The main advantage of within-subjects comparisons is that many characteristics of test conditions are equivalent for both tests (including patient attributes, timing of smear in the menstrual cycle, clinician factors, etc.). As many of these potential confounders are made constant, one can be more confident that any difference in test performance is related to the tests alone.

Within-subjects studies are common in automated primary and rescreening studies as slides from the same woman can easily be screened again by a different test. However, there have been concerns about the validity of this design for liquid-based slide preparation studies where the split-sample technique is employed in sample delivery.

Split-sample studies involve the Pap smear prepared in the conventional way onto a slide followed by the transfer of remaining cells on the sampling instrument to the vial of liquid used to prepare the liquid-based slide preparation slide. The split sample technique therefore represents a within-subjects, “matched pair” design.

It has been argued that split-sample studies were used initially due to ethical constraints as ThinPrep 2000 had not received FDA approval for clinical use (Rosenthal, 1998). However, this approach, according to the manufacturers, does not replicate intended use of the device⁹. The split-sample technique has also been reported as disadvantaging the liquid-based smear and underestimating its sensitivity (Brown and Garber, 1998). Some evidence supporting this view comes from a study by Corkill et al. (1997) which demonstrated that when a sample was rinsed directly into the vial, the percentage of slides that contained endocervical component was the same as for the conventional smear. This contrasts with earlier split-sample studies which have found ThinPrep smears lacking in

⁹ The Centre for Devices and Radiological Health (CDRH) of the FDA have written a warning letter (http://www.fda.gov/foi/warning_letters/m3659n.pdf) to TriPath Imaging which suggests that promotional material from TriPath has been false and misleading in representing a direct-to-vial study (Vassilakos et al., 1999) as representing intended use of the AutoCyte Prep System. Moreover, the FDA objected to the implication that use of the split-sample protocol was not adequate to establish the safety and effectiveness of the device.

endocervical cells compared to the Pap test (Australian Health Technology Advisory Committee, 1998).

An alternative sample delivery method to the split sample technique is the direct-to-vial (DTV) approach used for between-subjects comparisons (described in the following two sections). This involves the sample being rinsed directly into the vial for the liquid-based test. This means that only one test (Pap test or LBS) is performed for each woman. Within-subjects designs (as used in split-sample studies) are usually more valid than between-subjects designs (as used in direct-to-vial studies), as is the case for primary and rescreening devices. However, problems of sampling described above lead us to suggest that the most valid design for studies of liquid-based screening devices is a randomised controlled trial (RCT), described below.

Tests randomly allocated to different women

For evaluations of screening and diagnostic tests, a study design of less validity than within-subjects designs involves tests being randomly applied in different women in **between-subjects** comparisons. (Note however, that for liquid-based slide preparation studies, the reverse is true, for reasons described above). However, such studies are only valid if the two study arms compared are sampled from populations of similar risk, which may be achieved through random allocation of sufficiently large samples of women to test conditions. Without randomisation, one cannot be confident that the spectrum of disease is equivalent between women receiving each test, as may be the case for example, if the ages of the women differ significantly between the two arms of the study (Sawaya and Grimes, 1999).

Ideally women should be randomised on an *individual* basis; that is, every woman should have the same “50-50” chance of receiving the conventional test or the device-assisted test. However, this is difficult to apply for liquid-based slide preparation devices when patient and physician choice commonly play a role, especially when an additional cost is involved in applying the liquid-based test. Allocation of *groups* of women (e.g. randomly allocating tests to practices or health providers) is a less valid alternative.

Tests non-randomly allocated to different women

In the absence of randomisation, the conventional test (“the control”) is performed for one group of women who are tested either: (i) over the same time period as those receiving the device-assisted test (i.e. *contemporaneously*), or (ii) at a period before that in which the device-assisted test is used (i.e. *historically*).

Both methods are regarded as less valid than the other designs described above regardless of whether liquid-based slide preparation devices or primary or rescreening automated devices are evaluated. This is because achieving equal prevalence of abnormalities in both groups (i.e. of women receiving the conventional Pap test or the device-assisted test) is difficult to control. Differences in patient age, screening history, and socioeconomic status between groups affect women’s risk for cervical abnormalities and threaten the assumption of equivalent spectrum of disease (Brown and Garber, 1998).

These designs are common in evaluations of liquid-based slide preparation devices and open to systematic biases. The provider may offer the liquid-based device to women who are perceived as being at greater risk for an abnormal smear (e.g. for repeat or surveillance smears), or to women who are willing to pay, or can be reimbursed, for the smear. Alternatively, the choice of test offered could be made at the provider or practice level where a provider has decided to “convert” to liquid-based smears and may offer all women this technique. The type of health provider who decides to convert to a new device may be consulted by women with a different spectrum of disease to those consulting a provider who has not changed their practice. One study (Weintraub and Morabia, 2000) revealed that ThinPrep smears were offered by physicians who were more likely to be female, and older, than those using conventionally prepared smears. Ideally, random recruitment of women to screening test can address this bias, either through consecutive enrolment or another system. This feature is reported in studies reviewed as a study quality code, although it is recognised that such biases are difficult to control. Studies rarely discuss whether women recruited into a study were asked for informed consent.

Studies using an historical control group (for the conventional smear) represent the least valid design in the evidence hierarchy of considered studies. This approach is open to the same recruitment biases described above for concurrent screening in that, for the liquid-based slide preparation series, the provider may select women to be offered the liquid-based smear from a larger group of women who request smears. However, the difference in the time period introduces further potential biases in the population being screened or the way they are screened, such as the sampling instruments used (see Section 2.3 below).

Effectiveness of a new device may also be affected by the original performance of laboratories. The performance of some laboratories for conventional Pap tests prior to conversion to liquid-based smears is sometimes noticeably poor, which may be reflected in the extent of increased detection reported for liquid-based slide preparation devices. High level performing laboratories may not show the same improvements as poorly performing laboratories for new devices (Bartels and Vooijs, 1999; Brown and Garber, 1998).

Without adequate randomisation, detailed descriptions of population characteristics are essential. This can assist in determining whether confounders, or other differences in the groups apart from that they have received different screening tests, are responsible for any differences in detection found. However, this is only possible for known, and measured confounders; only randomisation can control for both known and unknown confounders. In sum, studies must be carefully controlled and require very large and very similar populations for a valid comparison (IAC Task Force No 3 on “Sampling, sampling errors and specimen adequacy”, draft May 2000, Dr Ulrik Baandrup, *personal communication*).

2.2 SPECTRUM OF DISEASE IN STUDY POPULATION

Regardless of whether the study is a within-subjects or between-subjects design, comparability between studies or between centres in a multi-centre trial also requires detailed description of the population from which the sample is drawn. To enhance external validity and generalisability to New Zealand’s national cervical screening programme, the sample should ideally be from screening populations rather than referral and/or teaching hospital (e.g. University-based) settings. In the latter where there is higher prevalence of disease, test sensitivity is likely to be higher and specificity lower (McCroory et al., 1999). In a study involving women recruited from screening centres (with low risk populations) and hospital settings (with high risk populations), there was a higher yield of abnormalities detected by ThinPrep in the screening centres, but not in the hospital settings (Lee et al., 1997). Authors suggest that ThinPrep may have an advantage over conventional smears in detecting lower grade lesions more prevalent in the screening centres. Whilst this study was excluded for appraisal as no valid reference standard was used, it does suggest that when interpreting results account must be taken of the population screened and its likely spectrum of disease. Studies involving screening populations in developed countries will be more relevant to the New Zealand screening programme, although should be interpreted with care as the screening interval and uptake may differ (as discussed in Chapter 7).

Depending on the prevalence of disease in a population, there may be few cases of higher abnormalities such as HSIL and cancer, which are the most important outcomes for a test to detect (see Section 2.8 below). In smaller studies it may be necessary to “enrich” the sample by including women screened at high-risk sites, or “seed” the sample with known archived cases. However, such strategies do not represent usual working conditions.

Another approach involves tracing the screening history of women who have later been diagnosed with cancer (i.e. “cases”) in a “case control study”, commonly identified within a cohort or case series (see glossary under “nested case control study”). This is problematic, as a new abnormality may have developed in the interval between the screening test and cancer developing¹⁰. There is also the possibility of rapid onset cancers that progress after a cervical screen, though their existence is controversial and if they exist their occurrence is likely to be rare (Halford, 1998). Such a method is

¹⁰ Including a recent screening test result prior to diagnosis is also contentious, as it is known that the false negative rate for cytology is higher in the presence of cervical cancer and smears taken when a woman has symptoms may be “diagnostic” rather than “screening” smears. For this reason many studies of this kind exclude results from smears made within a few months of diagnosis of cancer (Dr Ann Richardson, July 2000, *personal communication*).

therefore a better means of auditing the effectiveness of a screening programme, rather than measuring the clinical effectiveness of the test *per se*.

2.3 SAMPLING DEVICE

A review of split-sample liquid-based slide preparation studies by Austin and Ramzy (1998) concluded that results appear to have been influenced by sample collection instruments. The traditional wooden Ayre spatula was associated with *decreased* detection for ThinPrep compared with Papette, Cervex Brush, and “cytobrush-plastic spatula” combination. Plastic collection instruments are advised for use with liquid-based slide preparation devices. For AutoCyte Prep studies, use of a cotton-tipped swab, which is now virtually obsolete (IAC Task Force No 3 on “Sampling, sampling errors and specimen adequacy”, draft May 2000, Dr Ulrik Baandrup, *personal communication*) was *less* favourable than the Cervex brush. These results are consistent with a recent systematic review of instruments used to take smears which found that more absorbent, older methods lead to poorer detection, with the wooden Ayre spatula least effective (Martin-Hirsch et al., 1999). As the type of collection device used frequently varies based on provider choice, and is not always recorded by the smear-taker or researcher, this introduces a bias that is difficult to appraise (Rosenthal, 1998). In split-sample studies where the instrument is consistent, older studies may underestimate detection rates of smears generally. More problematic are direct-to-vial studies that use an historical control. This design may lead to older, less effective instruments being used for conventional smears than for the more recent liquid-based screening group (e.g. Vassilakos et al., 2000). Direct-to-vial studies where sampling instruments vary or are unreported are also problematic. The historical control in several studies (Papillo et al., 1998; Weintraub and Morabia, 2000) used a variety of unspecified sampling instruments whereas the more recently applied ThinPrep smears were collected by newer sampling instruments such as plastic spatulas.

Another potential confounder is whether one or two Pap smears are taken. Austin et al. (1998) suggest that the two slide smear is a recognised enhanced sampling method and should be considered as a comparison for liquid-based smears, especially with regard to HSIL detection.

2.4 CYTOLOGIST ASSESSMENT OF SLIDES

Another area prone to bias relates to the assessment of slides by different cytologists. It has been suggested that cytologists should have comparable experience and abilities as determined by skill proficiency in both arms of a trial comparing screening methods (Coleman, 1998). Familiarity with a different set of morphological parameters is required for reading liquid-based slides from that used in conventional cytology. For example, reading of ThinPrep slides requires three days training because of differences in how to detect abnormalities (Rosenthal, 1998). The learning curve may take many months and will vary between individuals, laboratories and populations being served (McGoogan et al., 1998). For these reasons, a cytologist may be selected from a laboratory to receive training in reading liquid-based smears and may be solely responsible for reading such smears in an investigation (e.g. Bishop et al., 1998). This practice introduces several possible biases into a study. The training itself is likely to encourage revision and improvement of slide reading skills generally regardless of any unique advantages of the device. Those cytologists chosen for training may also be selected because they are more experienced or skilled in the first place.

A common “observer effect” in research occurs where people act differently when they know their behaviour is being measured. When cytologists are aware of testing conditions, research suggests that their behaviour is altered (Spitzer, 1998). In prospective trials, differences in alertness should have the same effects for both study arms (Linder and Zahniser, 1998). However, this is not the case in studies where historical controls are compared with a laboratory that has later converted to a new device. Cytologists returning from training in this situation may be more enthusiastic about the new approach that could lead to their screening with greater vigilance. Indeed, an “invigorating spirit of change and innovation” may apply to a whole laboratory if liquid-based devices have been taken on board completely (Austin, 1998) which may disadvantage comparisons with smears read conventionally before conversion to the new device.

With respect to studies investigating rescreening devices (e.g. AutoPap), some studies have compared the performance of the automated device on previously screened smears. To offset the effect of altered vigilance for the device, a second arm in the study may include slides randomly selected for review by cytologists at the same percentage that the rescreening instrument selects slides (Bedrossian et al., 1998) ; but see Section 2.5 for debate on this point. In this situation cytologists' vigilance will be heightened for both arms.

Blind interpretation of smears is another means of reducing bias by cytologists aware of the study hypotheses. However, in studies evaluating liquid-based slide preparation, it is not possible to completely remove this bias due to the different appearance of smears (Diaz-Rosario and Kabawat, 1999; Sherman et al., 1998). Blinding to the comparison test's result can also be complicated in diagnostic discrepancy studies where discrepant smears are reviewed by a pathologist for final diagnosis. This means that the pathologist will know that a smear being looked at is positive from at least one of the tests. For these reasons, evidence tables for studies evaluating liquid-based screening devices only report "blind verification" as a *study quality code*, and do not report "blind interpretation". Blind verification of smears by expert panel is not possible for liquid-based slide preparation devices for the reasons described above.

2.5 APPROPRIATE COMPARISON TESTS FOR RESCREENING

Performance of automated rescreening techniques should be assessed against other QC methods such as random rescreening, rapid rescreening, or directed rescreening of high risk groups (Wain, 1997). Some researchers suggest using a traditional rescreening method, such as 10% random rescreening as the comparison manual screening test (the comparison conventional testing group is also called the "control") as it was mandated federally in the USA by amendments to the Clinical Laboratory Improvement Act (CLIA) (Sawaya and Grimes, 1999). However, a study of this approach found that it yielded almost no gain in life expectancy (Raab, 1998). This method is not widely used in New Zealand where rapid review is standard practice (Dr Peter Fitzgerald, March 2000, *personal communication*). It has been argued that a more balanced comparison would require the same percentage of slides randomly selected for rescreening as selected by the rescreening instrument for review (e.g. 20% of slides) (Bedrossian et al., 1998). Another view is that the control group for this comparison should be "targeted rescreening" where negative slides at higher risk are targeted for review (e.g. from women who have not been regularly screened, or have a history of unsatisfactory or abnormal smears) (Melamed et al., 1998). According to Rosenthal (1998) to make a nearly equivalent comparison, full manual review of all the negative slides should be performed. This has been suggested as the gold standard for rescreening (Hutchinson, 1996; Spitzer, 1998).

2.6 REFERENCE STANDARD

Need for a reference standard

A "reference standard" is regarded as providing an accurate or "truth" diagnosis. It is an independently applied test that is compared to the screening test being evaluated in order to verify test accuracy. Many (especially earlier) research studies evaluating cervical screening devices have used no reference standard at all. Without verification of positive and negative diagnoses, one cannot be truly sure that all readings represent accurate diagnoses; that is positives are "true positives" and negatives are "true negatives". Instead some of these diagnoses may represent "false positives" (slides read as positive in the absence of cervical abnormality) and "false negatives" (slides read as negative in the presence of cervical abnormality).

Studies exhibiting this problem commonly compare the "yield" of abnormalities between the conventional test and the device-assisted test. Results such as "110% more HSIL detected by test" assume that these additional abnormalities are true positives and that the test is therefore more sensitive. In studies comparing tests applied to *different* women, differences in yields may reflect a different prevalence of disease in the two populations. For example, a study excluded from this review (Dupree et al., 1998) reported a higher detection of abnormalities by ThinPrep when compared to the conventional smears performed over the previous year. As ThinPrep was only applied for women who

could pay or be reimbursed for the additional costs, such women may reflect a different level of risk than those not given a choice of whether to receive ThinPrep testing in the previous year. Comparison of abnormality yields without verification in such studies is not very informative. Problems also arise in studies comparing tests applied to the *same* woman, where an abnormality detected by one test is assumed to be true but may actually be false (i.e. a false positive). This design also commonly assumes that where both tests agree or are concordant, the diagnosis is accurate. However, concordant assessments may simply reflect a false diagnosis by both tests of falsely positive or falsely negative¹¹.

Histology

Whilst the need for a reference standard in evaluating screening tests is widely accepted, there has been great debate about what a valid standard is. Established clinical outcome, or the eventual development of cancer, is considered the “gold standard”. However, this would require a longitudinal study involving many women tested repeatedly over many years. Due to the time, expense and difficulty of this approach, an acceptable surrogate suggested is histology. A key reason for this is that whereas the Pap smear is a screening test, in clinical practice diagnosis is based on biopsy confirmation (Austin, 1998). This forms a basis for most clinical management decisions (McCrorry et al., 1999).

There are limitations of histological reference standards. Cytological diagnoses, especially of low-grade abnormalities, are often not confirmed histologically because of inconsistencies in histopathological interpretation, the location of the biopsy, and the possible regression of lesions (Sherman et al., 1998). Whilst these are valid concerns, the effects of these factors in studies comparing the Pap test with an alternative screening device would be expected to affect both tests similarly, as long as conditions for histology diagnosis were the same for both tests (Carpenter and Davey, 1999). Strategies may also be implemented to limit the influence of these errors on the accuracy of diagnosis. For example, variances in histological interpretation can be minimised by using consensus diagnosis of an expert panel (Intersociety Working Group for Cytology Technologies, 1997). The error resulting from regression of lesions can also be minimised by employing concurrent biopsy, such as ensuring the biopsy is taken within three months of the smear of interest (McCrorry et al., 1999). Studies commonly do not report the interval between cytology and biopsy and when “available” follow-up data is retrieved from records this delay can be significant. This could disadvantage conventional Pap tests when applied as historical controls before the use of a device-assisted test. An example of this bias is a study by Papillo and colleagues (1998) where biopsy follow-up was obtained for ThinPrep smears between one and seven months after the smear reading, whereas follow-up for the historically applied Pap test occurred between nine and 21 months later. The authors argue that this design would have disadvantaged ThinPrep as positives in the longer follow-up period would be more likely to progress to higher grade lesions and be easier to detect. However, at the threshold used of LSIL, only 11% of cases at that more common grade would be expected to show disease progression and 60% would regress, and progression from low to high grade abnormality would take 12 years (Östör, 1993). Therefore a bias toward lower correlation between cytology and histology would be expected for the conventional smears.

Panel review by expert cytologists

Whilst it has been suggested that an expert cytology review by a single pathologist provides a valid alternative to confirmation by biopsy (Linder, 1998), this form of verification is relatively unreliable and is not considered an acceptable reference standard in this review. A methodologically acceptable reference standard (though arguably less valid than histological verification) is adjudicated panel cytology review (McCrorry et al., 1999). Ideally this involves consensus review by two or more expert cytologists after discussion at a multi-headed microscope. The Intersociety Working Group for Cytology Technologies (ISWG) have proposed guidelines for identifying false negatives, which can be problematic for biopsy follow-up approaches (Bedrossian et al., 1998). They suggest that all putatively normal slides read by cytologists should be read by at least two experienced cytology professionals. Any positives should then be given consensus review at the microscope with at least one additional expert. The ISWG acknowledge that consensus diagnosis of biopsy can be an appropriate reference

¹¹ Whilst follow up (by repeat smear or histology) may indicate that the tests were both false, in this design follow-up results are not retrieved for the purposes of verification and are therefore not presented in the study’s results.

standard in follow-up of a significant subset of patients with a positive cytological diagnosis, although they prefer that “cytology devices should be compared to a cytological gold standard”.

Acceptable reference standards for this review

As will be discussed in the next chapter, an adequate reference standard was an important criterion for inclusion of studies for appraisal in this review. “Reference standard” is also reported as a study quality code in evidence tables for appraised papers. Given the above information, histological diagnosis is accepted in this review as a valid reference standard. In the absence of histological verification, especially for verification of negatives where colposcopy and biopsy may not be acceptable ethically (see Section 2.7 below), consensus cytology review by an independent panel of at least three cytologists is considered to be an acceptable standard in this review. Note that “independent” here means different cytologists on the panel from those who read the smears to produce the initial test diagnoses. Regardless of which reference standard is used, it is important to compare studies with similar reference standards for validating positive diagnoses due to their varied impact on sensitivity and specificity estimates (Brown and Garber, 1998).

2.7 VERIFICATION BIAS

Investigation of asymptomatic people with negative test results is “invasive, impractical, time consuming, costly and probably unethical and unacceptable” (Frommer et al., 1988). This means that it is not generally acceptable for women whose smears are read as negative to receive colposcopy or histological follow-up in research studies. A consequence of this is that histological reference standards are usually only applied to smears read as positive. That is, verification is limited to women referred for evaluation of a cytological abnormality. Verification of only cytological positives is susceptible to *verification* or *work-up* bias where a high frequency of histological abnormalities is included in the sample verified. This bias can lead to elevated estimates of sensitivity and lowered estimates of specificity (Nanda et al., 2000). Whether “verification” of positives and/or negative diagnoses is performed is reported in evidence tables as a study quality code.

The lack of histological verification of test negatives is a major drawback of research conducted in this area. In this situation, sensitivity and specificity cannot be directly determined (though relative true positive rates and relative false positive rates can, as described in Section 2.8 below). Positive predictive value (PPV) may be reported where the same threshold for positives is used on cytology as histology. However, there are important limitations to this outcome discussed in the next section (see Appendix 1 for calculation).

Screening tests are subject to a trade-off between sensitivity and specificity (Fahey et al., 1995), as new devices become more sensitive, they may become less specific (Brown and Garber, 1998). Determining specificity is problematic in cervical screening because few studies are able to provide meaningful estimates due to their study design limitations. Brown and Garber (1998) argue that when colposcopy-directed biopsy is used as a reference standard, validation of negative diagnoses will be based on a non-random choice of cases which will bias downwards estimates of specificity. And so, negatives according to a threshold of HSIL+ may be based on a subset of slides that were positive for at least one screening test at a lower cytological threshold (e.g. ASCUS). As these abnormalities may be correlated with the presence of histologically confirmed abnormalities, one would expect lower specificity than if one also verified all cytological negatives. It is difficult to determine accurately the true proportion of women without abnormality. In view of these difficulties, Brown and Garber (1998) assumed in their evaluation of liquid-based slide preparation devices and automated screening devices that choice of device does not affect the specificity of screening. They argued that manual rescreening is likely to be the most important determinant of specificity as this is constant for liquid-based slide preparation devices and primary screening methods.

In studies where no negative test results are verified with the reference standard, exact sensitivity and specificity cannot be determined, as one hasn't verified all the cases in the sample. However, one can determine the incremental characteristics of a test by directly comparing independently applied conventional and new tests (Chock et al., 1997). In such studies the reference standard is only applied when either or both tests on the same woman reveal a positive result. Verification of all positive results

on either test (i.e. concordant and discrepant positives) can allow the calculation of the **relative true positive rate** (relative TPR) of one test over another test, and the **relative false positive rate** (relative FPR), using the formulae given in Appendix 2. These outcomes are reported in our review where appropriate.

There are a few points of caution to note about interpreting these outcomes (Chock et al., 1997). Where one test has higher relative TPR and lower relative FPR than another test, the former test will be clearly more accurate. There is often a trade-off such that one test has a higher relative TPR at the expense of a higher relative FPR. In such cases, relative TPR and relative FPR are insufficient to determine the practical impact of this trade-off¹².

Finally, discrepancy studies involve verification of discrepant cases only when comparing diagnoses of two tests. This means that concordant negatives *and* concordant positives are not verified and are assumed to be true. It is likely that tests may have similar problems (such as sampling methods, interpretation of borderline or mild abnormalities) and therefore a significant proportion of concordant assessments may represent false results for both tests. This study design consistently under-estimates both sensitivity and specificity of the tests (Miller, 1998).

2.8 REPORTING OF OUTCOMES

Sensitivity and specificity

In reviewing the clinical effectiveness of devices for cervical screening, the key outcomes were the devices' test characteristics. The primary outcomes required to determine a screening test's accuracy are its sensitivity (Se) and specificity (Sp) (see Appendix 1 for calculation). A diagnostic test should minimise the false negative rate (where abnormalities are missed by screening) by having higher *sensitivity* and lower the false positive rate by increasing the *specificity*. However, these characteristics are related such that if sensitivity is increased, there *may* be a decrease in specificity. To use an extreme example, if all cervical smears were read as positive, one would demonstrate 100% sensitivity as all true positives would have been diagnosed. However, one would also have numerous false positives leading to women receiving investigations and follow-up unnecessarily. Achieving an increase in true positives and increasing false positives is achieved by simply lowering the threshold for a positive diagnosis.

As mentioned in Chapter 1, specificity is of concern because false-positive tests lead to additional investigations including repeated smears or colposcopy-directed biopsy which bring with them costs in terms of patient anxiety and inconvenience as well as resource costs to the health system generally. Therefore, the assessment of specificity has been viewed as being as important as sensitivity in any cost-benefit analysis (McGoogan et al., 1998). The impact of changes in specificity are particularly profound because the vast majority of smears read are negative. And so, a 1% drop in specificity represents a far greater number of false positive tests than a 1% drop in sensitivity increases the number of false negative tests.

Thresholds for reporting positive diagnoses

Estimates of sensitivity and specificity are based on whether a screening or diagnostic test is positive or negative. However, cervical screening tests detect various degrees of abnormalities. The choice of *threshold* at which smears (at this level of abnormality or higher) are reported as positive (which also determines the levels below which smears are negative) is therefore of crucial importance. Choice of threshold is dependent on the purpose of the outcomes and therefore the event of interest.

The event of interest for cervical screening programmes is ultimately to reduce the incidence and mortality of invasive cervical carcinoma. However, evaluations of cervical screening have used cytologically detected pre-cancerous abnormalities as outcomes. Verifying the relationship between cervical dysplasias and carcinoma would require prospective studies (preferably randomised controlled

¹² In this situation, Chock et al suggest calculating the ratio of extra false positive to extra true positives (FP:TP ratio) but the comparison may be difficult if the target and sample populations have different prevalences of abnormalities.

trials) designating invasive cancer or cancer death as outcome of interest, though this is impractical for such rare outcomes (Melamed et al., 1998; Sawaya and Grimes, 1999). Such a study was not undertaken prior to cervical screening being introduced and as results from observational studies have verified the success of the Pap test it has been regarded as unethical to conduct an RCT of cervical screening (Koss et al., 1963). Evaluations of cervical screening have used cytologically detected pre-cancerous abnormalities as outcomes.

International working groups of cytologists have argued that new methodologies must demonstrate sensitivity and specificity across a wide variety of diagnostic categories, including HSIL and cancer (McGoogan et al., 1998). Others have advised that separate analyses should be performed for sensitivity, specificity and positive predictive value for ASCUS/AGUS, LSIL, and HSIL plus cancer (Intersociety Working Group for Cytology Technologies, 1997). However, HSIL plus cancer remains “the most important to consider in terms of decreasing morbidity/mortality from cervical cancer,” (p. 1313; Bedrossian et al., 1998). A narrower definition of positive is also advantageous as it includes those patients most highly likely to harbour a significant lesion (Krieger et al., 1998).

The progression of lower grade abnormalities to cancer is significantly reduced below the threshold of HSIL (Östör, 1993). This is particularly the case for ASCUS/AGUS, which is notorious for its lack of consistent criteria for diagnosis and is particularly open to subjective biases in interpretation (Krieger et al., 1998). It has been argued that this threshold for positives should not be used because of the unreliability of defining these lower grade abnormalities and the lack of inter-observer reproducibility of this category (Bartels et al., 1998; Brown and Garber, 1998; Davey, 1997).¹³ For these reasons, “missed cases” at this threshold should not be considered errors for inter-laboratory comparison or liability purposes, although they are useful for intra-laboratory comparisons for quality assurance and educational efforts (Krieger et al., 1998).

Most debate surrounds the usefulness of LSIL (or simply SIL) as a threshold. The International Academy of Cytology (IAC) task force on automation (Bartels et al., 1998) argued that it is not clear whether a less progressed lesion always poses a lesser risk to a patient than a more-progressed lesion. However, knowing precisely how to interpret the relevance of detection of lower grade lesions is likely to require larger trials with longer surveillance and histology confirmation (Leiman, 1999). In some countries such as the United States, a diagnosis of LSIL is frequently an indication for colposcopy. However, this is not the case in New Zealand where recent guidelines recommend that LSIL detected by screening be followed by a repeat smear in six months (Jones et al., 2000). Only about 11% of LSILs are likely to show disease progression, with 60% showing spontaneous regression (Östör, 1993). Therefore, if LSIL has not regressed in this six-month period and is detected again, colposcopy and another smear is performed because of the risk of HSIL. After follow-up of three clear smears (at six months and then annually), the women returns to three yearly smears.

Outcomes reported in this review

In the review of clinical effectiveness of devices (Chapters 4 and 5), detection of HSIL+ is regarded as the primary objective of the national cervical screening programme. These lesions are far more clinically significant than lower level abnormalities and if undetected have the highest likelihood of progression to invasive tumour (Syrjanen et al., 1992). Therefore outcomes (sensitivity, specificity, positive predictive value, relative true positive rate, and relative false positive rate) are reported at this threshold in evidence tables and are discussed in appraising studies investigating the clinical effectiveness of cervical screening devices. Efforts to increase sensitivity are likely to decrease specificity, leading to an increase in smears reporting low-grade change which require further investigation for relatively non-threatening lesions (Australian Health Technology Advisory Committee, 1998). Therefore, the diagnosis of LSIL does have cost implications for the NCSP as they are followed by repeat smears and in some cases, colposcopies. As well as leading to greater health sector costs, such investigations are likely to raise women’s anxiety and cause them inconvenience. Such impacts on cost are especially important given the relatively high number of LSILs compared to higher-grade lesions. Therefore studies of cost effectiveness which are based on estimates of

¹³ The issue is also complicated by variations of The Bethesda system (TBS). Unlike in the United States, Australia (since 1996) and NZ (since 1998) allow a code of “ASCUS: possible HSIL” which is not part of TBS (Dr Clinton Teague, May 2000, *personal communication*). These diagnoses lead to colposcopy and follow-up identical to that for HSIL and higher grade lesions in New Zealand (Jones et al., 2000).

sensitivity and specificity at the threshold of LSIL+ are included in this review (see Chapter 6). This review also reports outcomes at the LSIL+ threshold in evidence tables of appraised clinical effectiveness papers for those who are interested in this threshold (including readers from other countries and those considering indicators for laboratory performance quality).

Positive predictive value (PPV) is only useful in comparisons between tests applied to populations with the same prevalence of disease. In high-risk samples, PPV is inflated. This means that PPV is generally only useful in comparing tests' performance within a study, and is only valid in between-subjects studies where populations carry the same spectrum of disease (which is more likely in a RCT, as discussed in Section 2.1 above). Comparing PPV rates *per se* across studies is therefore not useful, although it is useful to look at whether a significant difference exists between PPV's for different tests within a study when the populations are comparable.

Studies have tended to focus on detection of pre-invasive squamous lesions, which have been the primary objective of cervical screening. Whilst cervical cytology is effective at reducing the incidence and mortality of invasive squamous cell lesions, cervical screening may also have an important role to play in preventing glandular carcinomas of the cervix (Roberts et al., 1999). Therefore, studies reporting detection of these lesions are also included in this review.

Finally, smear adequacy rates have been reported as outcomes in some evaluations of new devices for cervical screening. As the clinical significance of the unsatisfactory smear is widely variable, this outcome was not reported in this review as its affect (on the need for repeat smears) is regarded primarily as a cost issue (McCroly et al., 1999). However, smear adequacy rates are included in some models reviewed in the cost effectiveness chapter.

2.9 COMPARISONS BETWEEN STUDIES

Comparisons between studies have used varying experimental designs and thresholds for positive diagnoses making them hard to compare directly (Australian Health Technology Advisory Committee, 1998). There is also a lack of studies comparing the devices with each other, which will continue to be the case due to competitive commercial pressures (Halford, 1998).

Interdependence of primary screening and rescreening

When evaluating the new devices in isolation one must remember that they are used in one aspect of screening (primary or rescreening) and therefore have distinct roles to play. Primary screening devices including liquid-based slide preparation devices relate to enhancing the initial manual screening by cytologists. Computerized automated screening devices such as AutoPap can be used for primary screening, as well as rescreening of smears read initially as negative (to reduce false negatives). These screening modes are not independent. When evaluating liquid-based slide preparation devices, rescreening practices will also have been applied. However, what these methods are is rarely stated and several are possible (as discussed in Chapter 1). The test characteristics will therefore not be generalisable to laboratories employing different quality control (QC) procedures, such as rapid review, directed rescreening or full rescreening. Moreover, in studies employing historical controls for conventional screening, QC procedures or laboratory performance may have changed over time and may differ between the tests being compared.

The non-independence of primary and rescreening tests is particularly important in evaluating automated devices such as AutoPap that can be used in both modes. In a recent meta-analysis of AutoPap, the authors observed that two types of abnormal slides may be analyzed in the determination of sensitivity for this device (Abulafia and Sherer, 1999). In primary screening mode, one must consider sensitivity to all abnormal slides, whereas in rescreening mode, one must consider sensitivity to false negative slides read in the initial screen. One would expect false negative slides to be harder to diagnose as they were undetected in the initial screen, and therefore sensitivity would be lower. Studies employing an automated device in both modes in the same design are problematic because the tests are not being applied independently (Nanda et al., 2000).

Industry funding

A final note on comparing studies relates to the researchers involved and potential biases affecting their work. The financial stakes are high in this area. Many researchers were involved in developing the devices in the first place, and continue to have roles in these companies. It has been argued by Dorothy Rosenthal, who was someone involved in early development of automated screening devices, that early research in the academic context became “controlled by the profit margin” once venture capitalists entered the equation (Rosenthal, 1998). This has led to misleading advertising, presentation of incomplete data, and the exaggeration of the effectiveness of devices by the manufacturers. Given these pressures, one must be particularly wary of research that is conducted on behalf of manufacturers or with their financial support. Such industry funding is commonplace in the literature and pathologists with financial connections with the devices are commonly, due in part to their expertise and interest, over-represented in editorial committees, working parties and task forces, as well as research teams. In a meta-analysis of AutoPap, it was found that in 13 of 14 studies reviewed, the same basic group of researchers were involved with different subsets authoring papers. It was concluded that these studies may not be truly independent and that systematic interests/biases may exist. Given these concerns, a study quality code employed in this review, as used in the Duke University review (McCrary et al., 1999) for AHRQ describes level of industry funding for the research.

Chapter 3: Methodology

3.1 STUDY SELECTION

Study Inclusion criteria

Publication date

Papers were included for review if they were published from January 1997 onwards (excluding papers from early 1997 that were appraised in the AHTAC review). Earlier papers were accessed where required to provide background material for the review.

Publication type

Publications included primary research (published as full original reports) and secondary research (systematic reviews and meta-analyses).

Context

Studies were included for review if they report automated or semi-automated devices suitable for participants in a population-based cervical screening programme.

Study design

For primary research studies relevant to the clinical effectiveness of devices, the following criteria were used:

- A reference standard was used of either
 - i. histology (or negative colposcopy¹⁴), or
 - ii. cytology by adjudicated panel cytology review where discrepancies are resolved by a consensus diagnosis made by an independent panel of cytology professionals (ie, three or more cytotechnicians/cytopathologists not involved with the original reading of the specimen)

For primary research, economic studies relevant to cost effectiveness, the following criteria were used:

- The study assessed the effect of screening by the new device on life expectancy or quality, the number of cases of cervical cancer avoided, or total health care costs, compared to conventional Pap testing in a three year screening cycle.

Secondary research studies reporting systematic reviews or meta-analyses must include a methods section that describes how the relevant studies were identified, including the search terms used, databases searched, and the dates searching was conducted. However, it is important to note that no other selection criteria were used. Therefore, these papers may not employ the same inclusion and exclusion criteria as has been used here and the results must be interpreted with caution.

Devices

Automated and semi-automated devices for slide preparation, primary screening and rescreening feasible for ready introduction into the New Zealand NCSP were included. This criterion was operationalised as devices which (i) had received USA FDA (Food Drug Administration) Pre Market Approval (PMA), (ii) were used in a way that corresponded to the device's FDA approved use, and (iii) were commercially available as of May 2000. Studies are included which investigate FDA approved devices which may be used in ways that are not currently FDA approved (e.g. direct-to-vial delivery of

¹⁴ In the verification of negative smears or lower grade abnormalities, a negative colposcopy may not be followed up by a biopsy

liquid-based screening, use of AutoCyte Prep prior to AutoPap), which may be the subject of applications for Supplements to existing FDA PMA's.

Devices included relate to the following three modalities of cervical screening:

(i) Preparation techniques

Liquid-based slide preparation devices which aim to produce samples which are both more representative cell samples and also easier to interpret. Devices included AutoCyte Prep (previously known as CytoRich) and ThinPrep.

(ii) Primary screening devices

Semi-automated devices to perform primary screening and identify slides that require further evaluation by a cytologist (AutoPap).

(iii) Re screening devices

Semi-automated devices to perform re-screening of negative or within normal limits (WNL) slides after primary screening (AutoPap).

Study Exclusion Criteria

Studies were excluded if:

- they were not published in English
- they were “correspondence”
- they were reported only as abstracts
- they reported single case studies
- their methods and results were not clearly described, or had significant discrepancies
- they evaluated slides taken *solely* from the following samples: women who have had previous abnormal smears, women who have never been screened, pregnant women, women younger than 18 years old
- they were cited in the AHTAC review (Australian Health Technology Advisory Committee, 1998).

In addition, the review excluded studies evaluating the following devices, techniques or strategies:

- HPV testing devices
- strategies for recruiting women to a cervical screening program
- strategies for improving the uptake of screening
- variations to the screening interval
- sample collection instruments used to take the smear or view the cervical area
- location guidance systems (LGS) designed solely to enhance manual rescreening.

Excluded emerging technologies

There are many developments in cervical screening technologies beyond the devices considered in this review. Two devices initially considered in the search strategy but subsequently excluded from the review were Papnet and AutoCyte SCREEN. Neither device was FDA approved for primary screening, and they are both no longer commercially available from TriPath Imaging¹⁵. However, features of both are being incorporated into a new device, including features of AutoPap, called “AutoCyte SCREEN 2” (Fiona Diversi, Cytology Specialist for Dade Behring, Australasian distributors for TriPath Imaging, August 2000, *personal communication*).

¹⁵ TriPath was established as an umbrella company from the merger of AutoCyte (which had purchased the intellectual property rights of Neomedical Systems (NSI), producers of Papnet) and Neopath (producers of AutoPap 300QC).

Telepathology has the potential to provide remote access pathology that may have useful applications in rural and underserved areas. Telecommunication systems such as TriPath's LINK allow high resolution images to be captured and sent to a remote site for teaching, research, QC diagnostic and consulting purposes. It enables users at both sites the ability to interactively review images concurrently on their computers (Fiona Diversi, Cytology Specialist for Dade Behring, Australasian distributors for TriPath Imaging, August 2000, *personal communication*).

Other excluded technologies identified from our search strategy are listed below that may emerge as important alternatives to be considered in the future.

Laboratory based tests

- infrared spectroscopy, a method using the infra-red spectra of exfoliated cervical cells
- devices to assist manual screening including AcCell Series 2000 and AutoCyte's Slide Wizard (similar to NeoPath's "Pathfinder" which is no longer supported by TriPath Imaging, the umbrella company)

In vivo tests

- naked-eye visualisation of the cervix following the application of acetic acid (speculoscopy)
- cervicography: a method of visual inspection of the cervix using photography
- polarprobe: electro-optical sensor technique. Draft Submission Guidance for an IDE/PMA was released for public comment by the FDA on 25 August 1999:
<http://www.fda.gov/cdrh/ode/cervcan.pdf>

3.2 SEARCH STRATEGY

A systematic method of literature searching, grading and appraising was employed in the preparation of this report.

Searches were limited to English language material from 1997 onwards, to supplement the searching for the AHTAC review (Australian Health Technology Advisory Committee, 1998) which included articles published until July 1997. The original searches were performed in October 1999. The *Current Contents* and *Science Citation Index* searches were repeated in February, April, and May 2000 to locate articles which had been published subsequently.

Principal sources of information

The following databases were searched using the search strategy outlined in **Appendix 3**:

Bibliographic databases

- Medline
- Embase
- Current Contents
- Healthstar
- Science Citation Index
- Health Technology Assessment database
- Cancerlit
- Econlit

Review databases

- Cochrane Library

- Database of Abstracts of Reviews of Effectiveness
- NHS Economic Evaluation database

Library Catalogues

- US National Library of Medicine
- New Zealand National Bibliographic Database
- New Zealand Ministry of Health
- North Thames Regional Library (UK)
- World Health Organisation

Websites

- Health Canada
- UK Department of Health publications
- Australian Department of Health and Aged Care
- US Centers for Disease Control
- Minnesota Health Technology Advisory Committee
- American Society of Cytopathology
- International Academy of Cytology
- US Agency for Healthcare Research and Quality or AHRQ (formerly Agency for Health Care Policy and Research or AHCPR)
- Canadian Coordinating Office for Health Technology Assessment

Other

- Hand search of NZHTA print collection
- General Internet searching
- Citation searching: reference sections of retrieved papers, including background papers, were scanned for relevant publications

Note that hand searching of Journals, contacting of manufacturers, or contacting of authors for unpublished research were not undertaken in this review. However, regional distributors for reviewed devices including ThinPrep (Biotek) and AutoPap and AutoCyte Prep (Dade Behring) were contacted by NZHTA to obtain additional information on current availability of these devices in New Zealand as well as product lines being developed or awaiting FDA pre-market approval.

Search terms used

- Index terms from Medline/Healthstar (MeSH terms): Cervix neoplasms, vaginal smears, mass screening, cytological techniques, image processing, computer assisted
- Index terms from Embase: Uterine cervix cytology, mass screening, vagina smear, Papanicolaou test, cancer screening
- Additional keywords (not standard index terms) were used on Medline and Embase in addition to the index terms given above and for the remaining sources which do not have controlled vocabulary:
 - (vagina* or cervi*) and (screen* or smear*); (pap or papanicolaou or papnet¹⁶) and (screen* or smear*);
 - (automat* or rapid*) and (screen* or smear*); thinprep; cytorich; autocyte; (automat* or method*) and

¹⁶ Note that Papnet and AutoCyte Screen were initially included in search strategies although these devices were excluded from the review at the study selection stage once it was determined that they were no longer commercially available.

(screen* or smear*) and (cervi* or vagina*); rescreening.

Filters for study design (e.g. randomised controlled trials) were not included in the strategies as the number of references was small enough to be screened manually for relevant study designs. Full search strategies are provided in **Appendix 3**.

3.3 SELECTION AND APPRAISAL

The search strategy and supplementary search updates (to end of May 2000) identified over 700 articles. From reading abstracts/titles, the reviewer (MB) identified 58 publications as potentially eligible for inclusion, which were retrieved as full text. The inclusion and exclusion criteria were applied to these to select the final group of 26 papers, including systematic reviews, meta-analyses, and original reports of clinical and/or cost effectiveness, for critical appraisal and inclusion in the evidence tables. Included studies and other cited publications are presented in the References (including reviews, methodological discussions, guidelines, commentaries, editorials, correspondence, screening programme reports, and other background papers). Excluded retrieved studies of eligible devices are presented in Appendix 4 (note: this list excludes papers relating to Papnet and AutoCyte Screen devices which were subsequently excluded from the review when their commercial unavailability was ascertained). Additional background papers retrieved are presented in a supplementary bibliography in Appendix 5 (these provide a resource list of relevant literature).

Critical appraisal was informed by the Recommended Methods of the Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests (Irwig and Glasziou, 1996).

3.4 APPRAISAL OF STUDIES

Details are provided for critically appraised studies in tabular form in evidence tables (Tables 2, 3, 5 and 7).

Evidence tables for primary research studies of clinical effectiveness employed column headings described below.

- **Source of the study** including authors, and year published
- **Study design** including a grading of the study's level of evidence according to a hierarchy (see Section 3.6 in this Chapter, below), and a brief description of the study design and sample delivery (e.g. split-sample, direct-to-vial)
- **Comparison interventions** including description of the conventional screening strategy employed such as the Pap test and the semi-automated or automated device it is compared with, using the manufacturer's name
- **Location and dates of testing** including country and city if applicable
- **Sample** including number of smears in the comparison groups, source of sample, and the spectrum of disease reflecting whether the woman sampled come from a population at high or low risk for cervical abnormalities. Whether or not detailed information of the sample characteristics are provided in the study is also mentioned
- **Outcomes and verification** including the threshold for positive test results and positive verification results, what reference standard was used and how it was applied in verifying diagnoses, delay between histological verification and cytology where applicable
- **Results** including where relevant and appropriate the following: differences in detection rates, sensitivity, specificity, positive predictive value, false negative rate, health care costs, and p values to indicate the level at which significance was found. Non-significant results are also reported. Note that data on concurrence of diagnoses between conventional and tested devices, and data on slide adequacy were not reported. Where not provided, some results were calculated from data provided in studies.

- **Quality** assessments were made and coded as follows (with coding options)¹⁷:

Recruitment: How was the study sample collected? (random, consecutive, not random)

Blind verification: Was the test and reference standard measured independently (blind to each other)? (yes, no/not reported)

Reference Standard: Was the test compared with a valid reference standard? {histology (colposcopy/biopsy); panel review (i.e. cytological review by independent panel of at least two cytologists)}

Verification: Was the decision to perform the reference standard independent of the test results? (positives and negatives; positives and random fraction of negatives; positives and selected sample of negatives; positives only; none)

Industry: What was the industry's relationship to the study? (no support (not done or funded by industry), partial support (some funding from industry to authors individually or to the project), total support (done on behalf of industry)).

Primary research, economic studies of cost effectiveness were described and appraised in terms of their design, outcomes, data sources, data assumptions, limitations, key results and sensitivity of the model to value changes in variables. Greater detail concerning these categories is provided in Section 6.1.

Systematic reviews and meta-analyses were described and critiqued in terms of their search strategy, as well as their criteria for inclusion/exclusion, data synthesis and interpretation of papers considered. Note that these papers were appraised principally as background information to the report's systematic reviews of clinical and cost effectiveness.

3.5 KEY OUTCOME MEASURES FOR PRIMARY STUDIES

Where studies of clinical effectiveness report yields of abnormalities detected without adequate verification, it is not possible to establish whether increased detection demonstrates true positives or false positives. For this reason, concurrence of diagnoses between newer devices and conventional screening is not reported in this review. Also, whilst significant differences between yield rates of tests are reported, the actual percentage increases or decreases in yield rates are not reported in the evidence tables. In studies which involve tracing back to the cytology records of women with histologically confirmed abnormalities (e.g. invasive carcinoma) one is examining case prediction by screening tests. Results are presented as the proportion of histological positives detected by each screening test. This is a problem because abnormalities that are correctly identified and treated should not get to the stage of invasive carcinoma. Thus, only a sub-set of women with abnormal smears will be investigated by this approach.

As discussed in Section 2.8, it is only possible to report Se and Sp when one has reference standard verification of cytology positives and negatives. In within-groups comparisons where only concordant positives or discrepant results from the two tests are compared with a reference standard, one cannot directly determine sensitivity and specificity, as one hasn't verified all cases in the sample. However, the *relative* true positive rate (relative TPR) and *relative* false positive rate (relative FPR) can be determined based on comparison of all positive diagnoses on either test with a reference standard (Chock et al., 1997).

In studies where histological follow-up results are only available for cytology positives, only PPV can be determined although comparisons are only valid when populations are comparable in terms of prevalence. Therefore differences in PPV between tests are only reported in within-subjects studies or randomised between subjects studies.

Key findings for studies of clinical effectiveness (Chapters 4 and 5) are reported in terms of estimates of sensitivity (Se), specificity (Sp), relative true positive rate (TPR), relative false positive rate (FPR), and positive predictive value (PPV) at a threshold of HSIL+ for both cytology and reference standard. That is, in this review, diagnostic positives for cytology are abnormalities graded as HSIL or higher, and diagnostic positives for histology are abnormalities graded as HSIL or higher (CIN II/III or

¹⁷ Further discussion of issues which relate to study validity are provided in Chapter 2.

higher). Results at the threshold of LSIL+ are included in the evidence tables as these are relevant to cost effectiveness and laboratory quality estimates (see 2.8 of Chapter 2 for further justification).

Outcomes for economic models (Chapter 6) include cost effectiveness (CE) ratios (see Glossary), such as the cost per health outcome gained; for example, cost per year of life saved, or cost per cancer prevented. Outcomes may be reported as total costs or as additional costs that are relative to the cost effectiveness ratio of the conventional Pap test (also termed *marginal* or *incremental* cost-effectiveness ratios).

3.6 METHODOLOGICAL QUALITY OF PRIMARY STUDIES

Studies of *primary and rescreening* were rated according to a “hierarchy of evidence” described below.

- (1) All tests done on each person (within-subjects design; i.e. each slide read by both tests)
- (2a) Different tests done on randomly allocated individuals (between-subjects design, randomised controlled trial)
- (2b) Different tests done on randomly allocated groups (between-subjects design)
- (3a) Different tests done on different individuals, not randomly allocated, and recruited concurrently (between-subjects design)
- (3b) Different tests done on different individuals, not randomly allocated, with historical cohort (between-subjects design)

This hierarchy has been adapted from the Recommended Methods of the Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests (Irwig and Glasziou, 1996).

The hierarchy was altered for liquid-based device evaluations in consultation with Professor Les Irwig, co-convenor of the Cochrane Working Group. Professor Irwig suggested that randomised controlled trials (between-subjects comparisons) should provide more valid evidence than within-subjects comparisons when there is an issue of how samples are taken (Prof. Les Irwig, March 2000, *personal communication*). That is, for liquid-based screening methods, split-sample techniques used for within-subjects comparisons are thought to disadvantage the liquid-based smear taken after the conventional smear. Therefore between-subjects comparisons as demonstrated in direct-to-vial preparation of the liquid-based slide are ideally better, but only when one is confident that both study arms represent the same spectrum of disease. This would be achieved by well-randomised allocation of woman to test condition (see Section 2.1 for further discussion of these issues).

For *liquid-based screening studies*, a revised hierarchy was employed as follows:

- (1a) Different tests done on randomly allocated individuals (between-subjects design, randomised controlled trial)
- (1b) Different tests done on randomly allocated groups (between-subjects design)
- (2) All tests done on each person (within-subjects design)
- (3a) Different tests done on different individuals, not randomly allocated, and recruited concurrently (between-subjects design)
- (3b) Different tests done on different individuals, not randomly allocated, with historical cohort (between-subjects design)

For both hierarchies, it is important to note that within these levels, studies may be carried out with greater or lesser care as indicated by quality codes reported in the evidence tables. Also note that

studies representing some of these levels are omitted based on inclusion/exclusion criteria, including studies employing less valid reference standards than histology or panel cytology review.

3.7 LIMITATIONS OF THE REVIEW

This study has used a structured approach to review the literature. However, there are some inherent limitations with this approach.

Publications included in this review were limited to January 1997 – May 2000, inclusive.

This review has been limited by the need to restrict the analysis to English language studies and references presented in the database, cited by these papers, or suggested by individuals consulted during preparation of this review.

The studies included in this review were all conducted outside New Zealand. The transferability of effectiveness and cost effectiveness evidence to the New Zealand NCSP is discussed in Chapter 7.

The impact of HPV testing in predicting progression of disease has not been considered in this review.

For a detailed description of interventions and evaluation methods and results used in the studies appraised, the reader is referred to the original papers cited.

Chapter 4: Effectiveness of liquid-based slide preparation devices

4.1 THINPREP

The search strategy identified seven systematic reviews in addition to the AHTAC review (Australian Health Technology Advisory Committee, 1998) which considered ThinPrep's effectiveness. The methods and conclusions are described in Table 1 (p. 33). As discussed in the Methods chapter above, these papers may not employ the same inclusion and exclusion criteria as applied in this review and the results must be interpreted with care.

Secondary research

The current review is broadly an update of that conducted by the Australian Health Technology Advisory Committee (AHTAC) (1998). Based on research published up until July 1997, the AHTAC review found that there was little information on sensitivity and specificity due to lack of biopsy verification. Their review found some evidence that ThinPrep had slightly higher sensitivity than conventional Pap test slide preparation methods, and sampling devices used may impact on ThinPrep's performance. The review concludes that interpretation of evidence of ThinPrep's accuracy is difficult given limitations of study designs and implementation. AHTAC suggested further local population-based research was required and did not recommend increased uptake of ThinPrep at that time.

Since the AHTAC review, there have been a number of other systematic reviews. A study of data from split-sample studies published no later than 1997 concluded that there was an increased detection of abnormalities by ThinPrep compared to the Pap test (Austin and Ramzy, 1998). However, this review provided little information about its search strategy and did not consider aspects of study quality, including use of a reference standard in its consideration of studies.

A study of cost effectiveness conducted by Brown and Garber (1998) for the Blue Cross & Blue Shield Association, adapted for a later publication in JAMA (Brown and Garber, 1999a), included a systematic review of articles, abstracts and manuscripts produced during or before 1997. Reviewed papers used biopsy or expert panel review as their reference standard. From three studies evaluating earlier versions of ThinPrep which provided results by biopsy, colposcopy or both for all cytological positives (at LSIL+) that TP increased the true positive rate of primary screening, on average, by 14.9% (Brown and Garber, 1999a). However, it was acknowledged that there were considerable variations in reported effectiveness of ThinPrep, and of the conventional Pap smear with which it was compared. The authors conclude that ThinPrep may lead to a much smaller increase in sensitivity in laboratories in which the Pap test is more accurate.

A review conducted by the Minnesota Health Technology Advisory Committee (1999) considered 12 studies of ThinPrep published up until March 1998. This review reports that ThinPrep offers marginal increases in sensitivity and similar specificity compared with the Pap Test, with greater (unverified) yields of abnormalities. It is concluded that whilst ThinPrep may reduce false negatives compared to the Pap test, the added value in improving women's net health outcome has not yet been demonstrated. Further research is recommended.

An Evidence Report/Technology Assessment prepared by Duke University (McCrorry et al., 1999) for AHRQ reports a comprehensive review of the effectiveness of the Pap test as well as automated devices including ThinPrep. Only eight papers with a reference standard of histology or independent panel cytology review were included. Only a few studies used histological reference standards and these found values of sensitivity and specificity to be well within the range of that reported for the Pap test. Accurate estimations of ThinPrep's specificity could not be determined from current research. The review recommends that further research is required using verification by histology rather than cytological reference standards, and involving verification of test negatives (preferably by histology) to

allow determination of test specificity.

Based on a search completed in June 1999, a comprehensive review by ECRI's Health Technology Assessment Information Service (1999) appraised 10 ThinPrep studies. The studies suggest increased yield of abnormalities although these were not necessarily verified. Two studies suggested slightly higher sensitivity for ThinPrep compared to the Pap test. However, studies reveal no evidence of improvement for ThinPrep for specificity, PPV or NPV. The reviewers conclude that limited research suggests that ThinPrep has increased sensitivity though lower specificity and therefore a higher false positive rate. However, they caution that flaws in study designs mean that test characteristics are difficult to evaluate and further research is required to address these limitations. Furthermore, there are no studies of the long-term impact of ThinPrep on women in terms of incidence or mortality of cervical cancer.

The most recent review by Nanda et al. (Nanda et al., 2000) involved some of the team who were involved in the Duke University review (McCrory et al., 1999), described earlier. This review was based on literature published through to October 1999 and was quite exclusive in its selection criteria. Only three studies of ThinPrep were appraised, two of which are in our review (Bolick and Hellman, 1998; Hutchinson et al., 1999). The review argues that the three included papers suggest slightly higher sensitivity for ThinPrep compared with the conventional Pap test, with specificity being slightly higher in one study and lower (according to indirect measures of specificity: relative false negative rates) in two studies. It is concluded that there is a lack of evidence relating to the effectiveness of ThinPrep due to deficiencies in study designs. Three main areas of limitations include: lack of prospective studies applying both tests to the same women/samples; failure to verify negative test results adequately, with verification of positives commonly limited to discordant test results only; and little evidence on specificity.

Researchers from the University of Sheffield's School of Health and Related Research (ScHARR) considered the clinical effectiveness of ThinPrep (and AutoCyte Prep) for the UK's National Institute of Clinical Effectiveness (NICE) (Payne et al., 2000). The systematic review concluded, "it is likely that the liquid-based cytology technique will reduce the number of false negative test results, reduce the number of unsatisfactory specimens and may decrease the time needed for examination of specimens by cytologists" (p. 1). However, the authors qualify these conclusions by noting that studies lacked verification of all diagnoses which make sensitivity hard to quantify. Furthermore, they note, "the specificity of the liquid-based method is largely unknown and may be worsened" (p. 37). It is concluded that other efforts to improve the sensitivity of the screening programme as a whole should be considered, including increasing uptake of screening as well as using more effective specimen devices for taking the smear.

Table 1. Secondary research appraised relevant to liquid-based slide preparation devices

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
Australian Health Technology Advisory Committee (1998) Australia	Search: 1990 - July 1997 Databases searched: Medline, CancerLit. Also queried key device users and device manufacturers, and accessed material from FDA Premarket Clinical Trials. Conference papers, information from manufacturers, Internet searches, press releases and non-peer reviewed publications were accessed as background material. Keywords: included automated cytology, AutoPap, Papnet, Pap test screening, cervical screening, cytology screening, cytopathology, automated screening, vaginal instrumentation.	Experimental studies evaluating ThinPrep or AutoCyte Prep reviewed. Critical appraisal performed where "sufficient evidence from well-designed studies" to provide estimates of sensitivity, specificity, additional cases detected and costs.	There were no randomised controlled trials of devices, and no direct-to-vial studies. ThinPrep (TP) N=13 studies. TP led to a reduction in the proportion of smears rated unsatisfactory. High concurrence between TP and PS. Reports of Se and Sp were limited, as comparisons were not made with biopsy confirmation; however, there was some evidence that TP had higher Se than PS and resulted in more LSIL being diagnosed, and a higher detection of minor non-specific changes. Split-sample use of TP suggests an increase in detection of biopsy proven cervical abnormalities by between 6% and 11%, including an increase of 5-6% in HSIL. The sampling device appears to impact on TP's performance. Screening time may be shorter for TP than PS, but extra preparatory staffing is required and a significant training period is necessary. AutoCyte Prep (ACP) N=4 studies. Reduction of unsatisfactory smears by ACP compared with PS. High concurrence between ACP and PS. Whilst some evidence that ACP leads to higher yields (and therefore lower false negatives) than PS, insufficient data to estimate sensitivity. Screening time may be shorter for ACP than PS.	Overall, evidence that slide preparation techniques reduce the proportion of slides which cannot be interpreted and improve the reading of slides. However, lack of consistency in design, initial sample selection, and measure of outcomes made interpretation of evidence difficult. Problems included differences in sample sizes between tests, lack of information about groups of women receiving each test, and a lack of community based studies. Suggest a need for local population-based studies with devices used in the context of routine clinical practice, which clearly demonstrate the device's cost effectiveness. AHTAC's expert working party did not recommend increased uptake from a public health perspective at the time of the review. However, argue that the area is in constant change and refinements to devices could improve their performance and lower their cost. (Also see Chapter 6 on costs)

Table 1. Secondary research appraised relevant to liquid-based slide preparation devices (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
Austin and Ramzy (1998) USA	Search: completed November 1997 No information on search strategy or databases searched. Published studies, "unpublished studies reported at professional meetings or in the press", and from Food and Drug Administration Premarket Clinical Trials.	Studies comparing either ThinPrep or AutoCyte Prep to conventional Pap test. <i>Split-sample designs only.</i>	TP: N=19 split sample studies. ACP: N=10 split sample studies. Yield data demonstrated overall increased detection of epithelial cell abnormalities by liquid-based slide preparation devices compared to PS. Results varied considerably from study to study and appear to have been influenced by sample collection instruments. The wooden Ayre spatula was associated with decreased detection in split-sample studies for ThinPrep compared with Papette, Cervex Brush, and cytobrush, and plastic spatula combination. For ACP, use of a cotton-tipped swab was less favorable than Cervex brush in split-sample studies.	Lack of information on the search strategy. Studies reviewed limited to split-sample studies without data provided on comparison with diagnoses with a valid reference standard, or on spectrum of disease in sample population. Authors argue for a need for direct-to-vial studies from a wide variety of clinical practice settings with biopsy confirmation of HSIL+ abnormalities. Authors also suggest that comparative studies should compare liquid-based slide preparation devices with the two-slide Pap smear, which has been reported as being very advantageous in detection of epithelial cell abnormalities.
Brown and Garber (1998); Brown and Garber (1999a) USA	Search: 1987 - December 1997 Databases searched: Medline, HealthSTAR, CINAHL, EMBASE, EconLit. Also hand searched journals and queried experts and device manufacturers. Included articles, abstracts and manuscripts. Key words: various cervical cytological tests including ThinPrep with cervical cancer or neoplasia and sensitivity and specificity.	Included studies (including abstracts) of ThinPrep's accuracy which had adequate reporting, and which were written in English. Appraised studies reporting the number and cytological results from all slides, reported FDA approved use of the technology, used biopsy of expert panel review as reference standards, and included slides with a validated diagnosis of LSIL+.	N=23 papers (21 studies): 20 split sample, 3 direct-to-vial (2 using contemporaneous, and one historical, controls). Only 7 papers concerned ThinPrep 2000. Majority (19) of papers suggest that TP is more sensitive for detection of LSIL+ than PS. No studies of ThinPrep 2000 reported histological verification of all positives. Concluded from three studies of earlier versions of ThinPrep which provided results by biopsy, colposcopy or both for all positive cytological diagnoses (at LSIL+) that TP increased TPR of primary screening, on average, by 14.9%. Results were less consistent for (the few) studies of ThinPrep2000 where some of the cytological positives are verified by histology.	Not restricted to published literature. Authors argue that the results have several caveats. Sensitivity reported for the PS in many studies is unusually low, and TP may lead to a much smaller increase in sensitivity in laboratories in which the Pap test is more accurate. Study results are likely to vary as a function of how well trained cytologists were in using ThinPrep. Considerable variation in reported effectiveness of ThinPrep. Also different experimental designs which makes it difficult to compare technologies. (Also see Chapter 6 on costs)

Table 1. Secondary research appraised relevant to liquid-based slide preparation devices (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
Minnesota Health Technology Advisory Committee (HTAC) (1999) USA	Search: 1993 - March 20 1998 Databases searched: Medline, Current Contents, FDCR, and the National Cancer Institute Patient Directed Query. Key words: Pap, various cervical cytological tests including ThinPrep, vaginal smears and diagnosis, mass screening, vaginal smears, cytodiagnosis, and cytology.	Included studies evaluating ThinPrep. Two were not discussed due to "poor study methodology or outdated research".	N=12 studies, 5 DTV and 7 split-sample including two comparing ThinPrep with biopsy confirmed diagnosis and ten comparing TP with PS reviewed. Concluded that ThinPrep offers marginal improvement in sensitivity, and similar specificity as compared with conventional Pap smear preparation. Generally found that TP detected greater yield of low grade and more severe lesions, and fewer ASCUS. Improved specimen adequacy for TP. Less time in slide review suggested by several studies.	Concludes that whilst ThinPrep capable of reducing false negatives, the added value of new devices (including ThinPrep) in improving the net health outcome of women and preventing cervical cancer has not been determined. Recommends further encouragement of regular Pap testing, and research into the safety, clinical effectiveness and cost effectiveness of the new devices.
McCrory et al., (1999) USA	Search: to January 1999 Databases searched: Medline, CancerLit, Health STAR, CINAHL, EMBASE, EconLit. Also hand searched journals, bibliographies of included papers, and queried experts and device manufacturers. Key words: various cervical cytological tests including ThinPrep with cervical cancer or neoplasia and sensitivity and specificity.	Included studies (including abstracts) of ThinPrep's accuracy reported in English, with reference standard of cytology or histology. Papers reporting cost and health outcomes needed to assess the effect of screening on life expectancy or quality, number of cervical cases avoided, or total health care costs. Papers assigned quality scores according to predetermined methodological criteria.	N=8 studies, 3 DTV and 5 split-sample Studies that employed a cytological reference standard found significant improvements in sensitivity of ThinPrep over conventional Pap testing. However, the few studies that used histological or colposcopic reference standards found values of Se and Sp to be well within the range of Se and Sp reported for the Pap test. Existing information fails to provide accurate estimates for specificity of thin-layer cytology. Only one ThinPrep study verified test negatives with colposcopy or histology.	Suggests that future research should include verification of test-negative subjects to allow estimation of specificity, preferably by colposcopy/histology. Also suggests a need for research that verification of diagnoses use histological reference standards rather than cytological reference standards. (Also see Chapter 6 on costs)

Table 1. Secondary research appraised relevant to liquid-based slide preparation devices (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
ECRI Health Technology Assessment Information Service (1999) USA	<p>Search: completed June 1999 (however, some databases searched until January 1999 only)</p> <p>Databases searched: Medline, CancerLit, HealthSTAR, Cochrane library, Current Contents, Health Services Research Project-in-Progress, and various ECRI databases.</p> <p>Web sites: US FDA web site, Health Care Financing Administration Web site. Also bibliographies of retrieved papers and non-peer reviewed "grey literature".</p> <p>Key words: ThinPrep, extensive list of terms relating to vaginal smears, cervical diseases, screening, cytology, and sample preparation techniques.</p>	<p>Studies comparing ThinPrep 2000 to conventional Pap testing. (Excluded studies evaluating AutoCyte Prep as results only available on pre-FDA approved versions).</p> <p>Studies excluded for the following reasons: pre-FDA approved versions (including ThinPrep Alpha and Beta), lack of a consistent reference standard (and study notes the possibility of selection bias), focus on use of ThinPrep for HPV testing, and insufficient quantitative information for analysis.</p> <p>Meeting abstracts also excluded.</p> <p>Note that these excluded studies were also excluded from the present review.</p>	<p>N=10 studies</p> <p>All case series studies. Four studies were split-sample, six DTV.</p> <p>Some additional statistical tests of significance were run on some papers' data.</p> <p>ASCUS/AGUS detection yields were similar for TP and PS. Most studies reported that use of ThinPrep result in a greater number of diagnoses of abnormalities; however, limited reference standards were used. Two studies demonstrate higher Se for TP compared with PS. Evidence regarding PPV generally suggested no improvement. Findings on specificity and negative predictive value suggest no improvement.</p> <p>More recent DTV studies suggest that lack of endo-cervical component (ECC) in split-sample studies was an artifact of that study design and may not be replicated in routine practice.</p>	<p>Concludes from limited research that ThinPrep is as efficacious as Pap smear preparation with increased sensitivity, but also low specificity leading to a higher false positive rate.</p> <p>No studies have examined the long-term impact of this device on the incidence of cervical neoplasia and the prevention of cancer.</p> <p>Study designs used have been problematic with respect to rare verification of negatives, and use of single pathologist review as a reference standard in several studies. Conclude that the test characteristics of this device are therefore difficult to evaluate due to these biases. Suggests further research that addresses these limitations.</p>

Table 1. Secondary research appraised relevant to liquid-based slide preparation devices (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
<p>Nanda et al., (2000)</p> <p>USA</p> <p>Note: this review uses the same search strategy as McCrory et al., (1999), but uses tighter selection criteria.</p>	<p>Search: through October 1999</p> <p>Databases searched: Medline, CancerLit, HealthStar, CINAHL, EMBASE, EconLit. Also hand searched journals, bibliographies of included papers and recent systematic reviews, and queried experts and device manufacturers.</p> <p>Key words: various cervical cytological tests including ThinPrep with cervical cancer or neoplasia and sensitivity and specificity.</p>	<p>Included studies (including abstracts) of ThinPrep's accuracy reported in English, which met following criteria:</p> <p>(1) The study must prospectively compare screening test and reference standard on the same set of patients or slides; (2) if using a cytological reference standard, the study must use an adjudicated independent panel review of discordant negatives; (3) at least 50% of patients testing positive for HSIL must be verified by an histology/colposcopy reference standard; (4) must allow separate analyses of Se (or relative true positive rate) and Sp (or relative true negative rates).</p> <p>Papers assigned quality scores according to predetermined methodological criteria.</p>	<p>N=3 studies, 1 DTV and 2 split sample. Included two studies (Bolick and Hellman, 1998; Hutchinson et al., 1999) reviewed in the current review.</p> <p>Reports on results of three studies meeting criteria with Se and Sp available for only one (DTV) study (Bolick and Hellman, 1998), and relative true positive and negative rates available for the other two. These papers suggest higher sensitivity for ThinPrep compared with the conventional Pap test. Compared to the Pap test, specificity was higher for ThinPrep in one study (Bolick and Hellman, 1998) and lower in the two studies reporting relative false negative rates.</p> <p>Note that thresholds of ASCUS and LSIL were used for cytology in some reported results in this review.</p>	<p>The authors argue that deficiencies in study designs relate to the lack of evidence for ThinPrep (and new devices generally). Three main areas of limitations include: lack of prospective studies applying both tests to the same women/samples; failure to verify negative test results adequately, with verification of positives commonly limited to discordant test results only; little evidence on specificity.</p> <p>The restrictive criteria used in this study for inclusion for review has led to fewer papers being appraised. The necessity for split-sample designs also excluded direct-to-vial studies completely.</p>

Table 1. Secondary research appraised relevant to liquid-based slide preparation devices (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
Payne et al., (2000) UK	Search: not specified but likely to be late 1999. Databases searched: Medline, HealthStar, EMBASE, Science Citation Index, Cochrane Library, NHS Centre for Reviews and Dissemination, National Research Register. Also web pages of INAHTA members and HTA organizations, citation search of AHTAC review (1998), and accepted submissions from Industry. 3 searches: clinical, and cost, effectiveness, and modeling. Key words: cervical cancer, various cytology techniques, vaginal smears; or various liquid-based screening tests; or sensitivity, specificity, diagnosis, pathology. Added economics, cost analysis, fees and charges, budgets, models, or Markov chains for other searches.	Primary studies included if they compared LBS with conventional Pap smears with outcomes including Se, Sp, proportion of inadequate or unsatisfactory smears, and included clear tabulation of numerical results.	N=35 including 26 split sample studies (15 of ThinPrep, 10 of AutoCyte Prep, and one of both devices combined), and 9 cohort (DTV) studies (7 ThinPrep and 2 ACP). In 11 studies reporting Se and Sp, the Se was higher or the same for LBS compared with PS, though in several studies results were not statistically significant. Three systematic reviews were also reviewed. For LBS compared with PS, found: <ul style="list-style-type: none"> a decrease in the proportion of inadequate smears, though the proportions were wide and overlapping for all tests. an improvement in Se “although this is hard to quantify with the data available in the published literature” (p. 36). a “probable decrease” in specimen interpretation times (from relatively few studies). 	“It is likely that the liquid-based cytology technique will reduce the number of false negative test results, reduce the number of unsatisfactory specimens and may decrease the time needed for examination of specimens by cytologists” (p. 1). Noted study limitations including that no studies had verification of all diagnoses. Also observed, “the specificity of the liquid-based method is largely unknown and may be worsened” (p. 37). Argue that the sensitivity of the programme as a whole needs to be considered and “further efforts to target an improvement in uptake may be more cost-effective than an improvement in test sensitivity” (p. 38). Also suggest that the use of more effective specimen taking devices are also important ways to reduce the burden from cervical cancer. (Also see Chapter 6 on costs)

Key:

PS: Conventional Papanicolaou Smear Test

TP: ThinPrep test

Ref Stand: Reference standard

CP: Cytopathologist

LSIL: low-grade squamous intraepithelial lesion

Se: Sensitivity(Pap threshold/Ref Stand threshold)

PPV: Positive Predictive Value

LBS: liquid-based screening

ACP: AutoCyte PREP test

CT: Cytotechnologist

ASCUS: atypical squamous cells of undetermined significance

HSIL: high-grade squamous intraepithelial lesion

Sp: Specificity(Pap threshold/Ref Stand threshold)

DTV: direct-to-vial

Primary research: Study designs and quality assessments

The search identified 10 eligible studies with adequate reference standards comparing the conventional Pap test with the ThinPrep slide preparation process. Full details of these papers, including methods, key results and assessments of quality are provided in Table 2 (p. 42).

All studies appraised included histology as their reference standard. Studies are ordered in the evidence tables according to the hierarchy of evidence with more valid designs presented first. The first four report within-subjects comparisons using split-sample studies and the remaining six tested ThinPrep as a direct-to-vial (DTV) process and report between-subjects comparisons. Of these six DTV studies, one compared ThinPrep and Pap test concurrently, and five employed an historical control for the Pap test, including one, which also included an overlapping period of sample collection.

Study quality was broadly at a similar level across ThinPrep studies. Recruitment of women was rarely described and assumed to be non-random in all but two studies (Hutchinson et al., 1999; Weintraub and Morabia, 2000). In DTV studies, allocation to test condition was commonly dependent on physician or patient choice. Smear taking (sampling) instruments also frequently varied or were not reported. Therefore one cannot assume equal prevalence of disease in the samples of women compared for each test and outcomes relating to yield of abnormalities and positive predictive value are therefore of little value. Histological verification was not reported as performed by independent laboratory/staff except in a small sub-group of one paper (Inhorn et al., 1998). Partial industry support was indicated for six studies.

Primary research: Study results

Results including comparisons of yield of abnormalities and positive predictive value PPV between tests are not meaningful if the women compared reflect a different spectrum of disease. Such is the case for the six between-groups studies where ThinPrep is used in direct-to-vial mode. A lack of randomisation in allocation of women to test condition in all of these studies means that one is not confident that the samples compared have equivalent prevalence of abnormalities. This is likely to be the case for the five of six studies, which report using historical control groups where the Pap test was applied at an earlier period. Biases relating to this design concern test allocation, sampling instruments and assessment of slides (discussed fully in Sections 2.1, 2.3, and 2.4).

With split-sample studies where one is assured of equivalent prevalence between compared samples, two studies report comparisons of yield of abnormalities detected (Hutchinson et al., 1999; Wang et al., 1999). Both studies report on women who are at very high risk of carrying abnormalities and therefore may not be relevant to New Zealand's population based screening programme. Wang et al.'s (1999) study included Taiwanese women recruited from gynaecology clinics and women with suspected abnormalities. ThinPrep slides detected significantly more abnormalities at the threshold of LSIL and HSIL, although only a small sample was considered (n=972). The study by Hutchinson et al., (1999) was conducted in rural Costa Rica and involved an unscreened high-risk sample of 8,636 women. The study found a yield of ASCUS which was over four times higher for ThinPrep slides (n=651) compared to conventional Pap tests (n=159).

This greatly higher reporting rate would have significant cost implications to a screening programme in relation to repeat smears and raised anxiety of women screened. The higher rate of ASCUS detected by ThinPrep compared to the Pap test may relate to differential use of diagnostic criteria in the two centres. This is possible because the ThinPrep smears were read by an expert pathologist (Dr Hutchinson) in the USA whereas the Pap tests were read by recently trained cytologists in Costa Rica with expert assistance (Hutchinson, 2000). A much higher rate of reactive or inflammatory changes would have been exhibited in this Costa Rican sample than one would expect in USA and therefore these (ThinPrep) smears may have been more frequently read as ASCUS than the (conventional Pap) smears read by local cytologists (Koss, 2000). Notably, the study found no difference in yield of abnormalities at HSIL or higher. Yields are not useful in understanding the accuracy of a test (unless verified against a reference standard) (see Section 2.6 and Section 3.5); however, they are important to consider in terms of cost. A greater yield without increases in true detection of abnormalities may lead to unnecessary costs of false positives to the woman (unnecessarily) investigated and the health system (see Chapter 6 for further discussion of these issues).

The study by Hutchinson et al., (1999) reports a prospective population-based case series trial in Costa Rica where randomly recruited women received both the conventional Pap test slide preparation followed by the ThinPrep method (in a split-sample method). Relative TPR was 1.04 for ThinPrep compared to the Pap test approach, and relative FPR was .83 for ThinPrep (or 1.21 for the Pap test compared to ThinPrep methods). Whilst these estimates are only indirect measures of sensitivity and specificity and are likely to be inflated by only considering verification of positives, they suggest equivalent sensitivity and slightly higher specificity for ThinPrep method compared to conventional slide preparation. As the populations were the same in this split-sample study one can also consider the positive predictive value (PPV) of both tests. These reveal equivalent prediction of abnormalities at HSIL or above. As discussed above, the study found a four-fold increase in smears read as ASCUS by ThinPrep compared to conventionally prepared smears. This study is the only one that provides any information on test accuracy at a threshold of HSIL.

Three studies employed “case control” designs nested within “case series” (see Glossary under “nested case control”). That is, women with diagnosed abnormalities of cancer or adeno-carcinoma in situ were identified from groups of women receiving both the pap test and ThinPrep as a split sample (Inhorn et al., 1998; Roberts et al., 1999), or receiving either the Pap test or ThinPrep direct to vial (Ashfaq et al., 1999). From these women (cases), the predictive value of their earlier cytology results were then examined. Such designs have limitations (as discussed in Section 2.2), and are particularly questionable when the period between confirmation of diagnosis and the cervical screen is longer than a few months. These studies found similar degrees of prediction of cancer for ThinPrep and the Pap test (Inhorn et al., 1998), and variable results for the two studies considering glandular lesions. However, numbers of cases were small in all studies.

The split sample study by Inhorn et al. (1998) is particularly dubious. It summed the results from two trials (nested case control within case series) involving very small numbers of cases (7 and 2, respectively) with two “validation studies” involving 22 and 16 cases, respectively. In the validation studies specimens were selected from known cases of cervical cancer and slides were prepared from each specimen for ThinPrep and the Pap test as split samples. The ability for cytologists to detect cancerous cells was then recorded and compared for each pair of slides. Known cases were not mixed with normal slides and the cytologists knew that the cells were taken from women with suspected carcinoma. Apart from these biases, this design only considers screening performance for a very small, subset of abnormalities. The cervical smear is primarily deigned to detect pre-cancerous cells, which can then be treated before cancer arises.

Unfortunately, there is a dearth of results that provide any meaningful indicators of test accuracy. As discussed in Section 2.8, only results relating to a threshold of HSIL+ for cytology and the reference standard are reported here. As the reference standard was usually used to verify cytological positives only, sensitivity and specificity were not able to be determined for any study at this threshold. In the absence of verification of test negatives, the relative true positive rate (TPR) and relative false positive rate (FPR) at a threshold of HSIL was able to be determined, though only in one study (Hutchinson et al., 1999).

Conclusions

The paucity of high quality research has meant that the test characteristics of ThinPrep cannot be reliably determined from the current evidence base.

This conclusion concurs with the recent systematic reviews in this area. As recommended by AHTAC in 1998, large, prospective population-based studies would be required which address some of the design inadequacies of much research to date. These aspects of study quality are discussed at length in Chapter 2; however, the key areas relate primarily to inadequate verification of diagnoses against a valid reference standard. Many studies were excluded from our review because they did not use verification by histology or an independent panel of cytologists. In those appraised, all but one failed to verify negative test results adequately, with verification of positives commonly limited to discordant test results only. Difficult-to-read smears read falsely as negative by both tests are therefore not verified against a reference standard and are assumed to be negative. This leads to a lack of evidence on specificity, and inflated estimates of sensitivity due to “work-up” bias. The only study that did verify a small fraction of negatives was that by Bolick et al. (1998) (using a threshold of LSIL on

cytology). In this study, the small proportion (less than 25%) of negatives investigated histologically are likely to have exhibited borderline level abnormalities in order to require histological follow-up, thus specificity may have been underestimated.

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Hutchinson et al., (1999)	Level of evidence: 2 Prospective, population-based case series split-sample trial	Pap test vs. ThinPrep Beta (cervex brush)	Costa Rica Dates of recruitment not reported.	8,636 Population-based study with high incidence of cervical carcinoma (30/10,000 per year, which is approximately five times that in most USA populations). Randomly ascertained, voluntary participants who have provided informed consent.	Negatives not verified therefore only relative true positive and false positive rates can be determined. Positive test results at threshold of LSIL+. Referred for colposcopy by positive cytology at ASCUS+ or suspicious cervigram. Negatives included women with negative cytology (PS, TP and Papnet), and negative colposcopies. Verification of LSIL+ by review based on biopsy or cytology confirmation by 2 or more cytological methods. Histological verification of 93% of HSILs and 100% invasive carcinomas. Results also compared with detection of cancer-linked HPV DNA type	Higher yields of ASCUS by TP (650 or 7.5%) compared with PS (159 or 1.8%) which led to substantial increases in referrals for colposcopy. *Determined from table Rel TPR (HSIL)* (TP/PS)= 92/88 =1.04 Rel FPR (HSIL)* (TP/PS)= 58/70 =.83 Rel TPR (LSIL)* (TP/PS)= 257/210 =1.22 TP Rel FPR (LSIL)* (TP/PS)= 188/210 =.90 PPV (HSIL)* for TP (92/150=61.3%) PPV (HSIL) for PS (88/158=55.7%) PPV (LSIL)* for TP (257/445=57.8%) PPV (LSIL) for PS (195/420=46.4%) Higher HPV detection in slides positive for TP and not for PS than visa versa (p<.001). Found no abnormality in random sample of 150 women referred for colposcopy after negative PS and TP tests.	<i>Recruitment:</i> random <i>Blind verification:</i> no <i>Ref Stand:</i> histology/cytology <i>Verification:</i> positives only <i>Industry:</i> partial support TP read by expert pathologist in USA, whereas PS read by experienced CTs in Costa Rica recently trained in the Bethesda System and with expert assistance (see letter by Hutchinson replying to Koss). Hutchinson argues that this led to exceptionally good performance of the Pap test compared with TP. Higher rate of ASCUS yield in TP compared with PS may relate to differential use of diagnostic criteria in two centres (esp. given higher rate of reactive or inflammatory changes than expected in the USA in this rural Costa Rican population) (Koss, 2000). Threshold of ASCUS used for test positives; however, data is reported at a threshold of LSIL and HSIL here.

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Wang et al., (1999)	Level of evidence: 2 Split-sample study	Pap test vs. ThinPrep 2000 (cytobrush & plastic spatula)	Taipei, Taiwan Attending over an 8 week period in 1998	972, 18% menopausal High risk population (Taiwan) including 90% from general gynaecological clinics, and 10% from women who had "suspected abnormalities".	Yield of abnormal diagnoses for ThinPrep and conventional smear. Pilot sample of available biopsy follow-up for 49 ASCUS+ cases including 27 HSIL+	ThinPrep more likely to detect slides as abnormal (LSIL+, HSIL+) than conventional smear (p<.001, p<.01). Study reports higher Se and lower Sp for TP compared with PS but thresholds not given and as only positives verified, direct Se and Sp cannot be calculated. TP (HSIL) NPV = 92% higher than PS NPV = 74%	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> no support No information on time-lag between biopsy and test. NPV reported in text without raw data to verify calculations.
Inhorn et al., (1998)	Level of evidence: 2 Two FDA clinical trials (one of ThinPrep Beta, one of ThinPrep 2000) with nested case-control within a split-sample case series. Also two "validation" studies (one of ThinPrep Beta, one of ThinPrep 2000) where samples were made from tumours/hysterectomy specimens from known cases of cervical carcinoma.	Pap test vs. ThinPrep 2000/Beta (broom-type instrument for trials, instrument not reported for studies)	USA Pre 1996	47 cervical cancer patients No information on populations for FDA trials. TP Beta validation study from University setting, TP 2000 validation study based on gynaecology and oncology clinics from 5 institutions.	Ability of ThinPrep and conventional smear to detect cancerous cells. Cannot determine specificity as no negatives are verified. All cases confirmed by biopsy.	No significant difference in detection rates of cancerous slides for TP (45/47) and PS (44/47). TP Beta Trial: TP 7/7, PS 7/7 TP 2000 Trial: TP 1/2, PS 2/2 TP Beta validation Study: TP 21/22, PS 19/22 TP 2000 validation study: TP 16/16, PS 16/16 For the validation studies, the cytologists reading the slides knew that the specimens came from patients with suspected cancer.	<i>Recruitment:</i> not random <i>Blind verification:</i> in TP2000 trial only <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> partial support Note that in Beta study, two cancer slides were not identified by PS as they were unsatisfactory. Sawaya and Grimes (Sawaya and Grimes, 1999) argue that unsatisfactory smears are usually repeated and therefore should not have been included with smears judged as normal. Small sample from more valid nested case control trials.

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Roberts et al., (1999)	Level of evidence: 2 Split-sample study (nested case-control within a split sample case series)	Pap test vs. ThinPrep 2000 or Beta (cervex brush alone in half, endocervical brush and cervex brush in half) Choice of test in TP cohort made by physician Physicians asked to use cervex brush.	Sydney, Australia Dates of study not provided. Women aged 22-56 years, median 33. No other information on disease spectrum of sample.	30 samples available for 26 women with histologically confirmed cases of AIS in cohort where TP smear included as split-sample.	Ability of ThinPrep and conventional smear to predict AIS. Cannot determine specificity as no negatives are verified. Threshold of test for AIS positive is "suspicious for AIS" (where cone biopsy is advised) All cases confirmed by biopsy identified from available histological follow-up data. (Additional TP slides were made and reviewed by panel if differing from histology, but these results are not reported here)	Considering original diagnoses of TP and PS, ThinPrep detected 14/30 (47%) AIS slides and conventional smear detected 20/30 (67%).	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> partial support
Bolick and Hellman (1998)	Level of evidence: 3a Concurrently recruited direct-to-vial study	Pap test (broom type or combination endocervical brush & plastic spatula) vs. ThinPrep Choice of test made by physician	Utah, USA 1996-97	39,408 PS 10,694 TP "Wide variety of clinical practices throughout the western United States". Age distribution was "similar" for both groups.	Sensitivity and specificity Positive test results at threshold of LSIL. Verification by limited biopsy of 54 TP cases and 89 PS cases at unspecified degree of abnormality (though Nanda et al. (2000) suggest it is HSIL).	Se and Sp calculated for threshold of LSIL for cytology and HSIL for histology.	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> no support Small sample Most negative test results were not verified with histological examination. No information on time-lag between biopsy and test

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Weintraub and Morabia (2000)	Level of evidence: 3b Historical cohort for Pap test (study period overlapping with TP), direct-to-vial study.	Pap test (one or two slides using either Ayre spatula (63%) or Ayre spatula and brush combination (37%)) vs. ThinPrep (cervex brush, then later using cervibrush) Choice of test in TP cohort made by physician "based on the information they had received". Comparisons indicate that TP smears were taken by physicians who were more likely to be female, and older, than those taking PS.	Geneva, Switzerland 11/96 – 12/97 for TP 1/95 – 12/97 for PS	130,381 PS 39,864 TP Sub-sample analysed which excluded those receiving surveillance examinations, that is, smears repeated within 6 months of a positive smear, with the same result. 129,619 PS 39,455 TP From 150 outpatient clinics consulted by low-risk population of women receiving regular screening compared with historical smears from the same providers. CTs and CPs same for TP and PS readings.	Yield of abnormal diagnoses for ThinPrep and conventional smear Most significant diagnosis from reading of all slides used (as more than one slides sometimes made) Available, reasonably concurrent biopsy follow-up (within 2 months of cytology). Biopsy time-lag for PS is not specified. (Single pathologist review of nine cases where HSIL on cytology was negative on histology but not a valid reference standard).	TP resulted in significantly more abnormal diagnoses (ASCUS, LSIL, and HSIL) than PS (p values all <.001 (from sub-sample) To investigate whether sampling tool influenced results, found that yield of abnormal is lowest for PS with Ayre spatula (1.75%), compared with PS by spatula-brush (3.6%), though both PS techniques had lower yields than by TP (5.5%) (p<.001). Also found increased yield for TP when comparing same physician who used spatula-brush for CS and TP. From biopsy follow-up of HSIL in sub-sample, PPV (HSIL) for ThinPrep (130/150=87%) was not significantly different than PPV (HSIL) for PS (118/144=82%). PPV (LSIL)* for TP (982/186=44%) PPV (LSIL) for PS (90/221=41%)	<i>Recruitment:</i> consecutive <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only Industry: none Increase in yields with TP was largest when compared to yield after PS using Ayre spatula than after the spatula-brush. Differences may also be due to selection bias, and TP training received prior to TP being introduced (after which there was a substantial drop in PS use) Whilst surveillance smears were generally excluded, those where a different results were found by the repeat smear were still included. It is possible that TP may have been used preferentially on these surveillance smears where there would be a greater yield of abnormal expected. As different populations sampled and prevalence may differ between tests, PPV comparison may not be valid.

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Carpenter and Davey (1999)	Level of evidence: 3b Historical cohort for Pap test, direct-to-vial study.	Pap test vs. ThinPrep Sampling instruments varied during course of study. Choice of test made by physician	Kentucky, USA TP: 2/97-7/97; PS: a year earlier, by same physicians	4,660 PS 2,727 TP University hospital setting with a high rate of abnormalities.	Yield of abnormal diagnoses for ThinPrep and conventional smear. Available biopsy follow-up over 9-12 months post-test for selected LSIL+ smears.	ThinPrep resulted in more abnormal diagnoses (LSIL+) than conventional smears. TP gave fewer ASCUS diagnoses ($p < .01$), more LSIL ($p < .01$), and no difference in HSIL detection rate. PPV* (HSIL) for TP (38/54=70.4%) PPV (HSIL) for PS (72/87=82.7%). PPV* (LSIL)** for TP (132/156=84.6%) PPV (LSIL) for PS (170/182=93.4%) * determined from tables ** includes ungraded SIL	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> partial support Non-concurrent histology. As different populations sampled and prevalence may differ between tests, PPV comparison may not be valid.

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Papillo et al., (1998)	Level of evidence: 3b Historical cohort for Pap test, direct-to-vial study.	Pap test(variable sampling instrument) vs. ThinPrep (majority using broom-type, some with cytobrush and plastic spatula, by provider choice). Choice of test in TP cohort made by physician.	Vermont, USA 4/97 – 9/97 for TP. Historical data from same cohort from 1996 for PS.	18,569 PS 8,574 TP (from cohort of 16,314 smears) 12 practice groups that primarily use ThinPrep (89% of patients) including two high-risk population clinics.	Yield of abnormal diagnoses for ThinPrep and conventional smear. Available biopsy follow-up for LSIL+ for TP 1-7 months) and historically for PS (9-21 months).	ThinPrep resulted in more abnormal diagnoses (LSIL+) than conventional smears, and fewer AGUS cases. PPV (HSIL) for TP (41/44=93.2%) PPV (HSIL) for PS (52/66=78.8%) PPV (LSIL) for TP (113/140=80.7%) PPV (LSIL) for PS (153/212=72.2%)	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> partial support Authors argue that choice of test depended on cost or reimbursement rather than risk factors, though these factors may be related to risk factors. Non-concurrent histology with greater follow-up delay for the conventional smear group. As different populations sampled and prevalence may differ between tests, PPV comparison may not be valid.

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Diaz-Rosario and Kabawat (1999)	Level of evidence: 3b Historical cohort for Pap test, direct-to-vial study.	Pap test (variable sampling instrument) vs. ThinPrep (instrument not reported) Choice of test in TP cohort made by physician	Massachusetts, USA 2/98 – 8/98 for TP. Historical data from 2/97 – 8/97 for PS, from same physicians.	74,573 PS 56,095 TP 151 outpatient medical practices including general practitioners and gynaecologists. Difference in sample sizes due to gradual conversion from the PS to TP. Same staff reading slides for PS and TP.	Yield of abnormal diagnoses for ThinPrep and conventional smear, PPV for limited TP only. Available biopsy follow-up for LSIL+.	ThinPrep resulted in significantly more abnormal diagnoses (LSIL, and HSIL) than conventional smears. PPV* (HSIL) for TP (96/158=60.7%). PPV (HSIL) for PS (216/292=74.0%). PPV* (LSIL) for TP (630/848=74.3%). PPV (LSIL) for PS (1040/1311=79.3%). * Determined from table To test for selection bias in the uptake of TP, compared detection rates for practices with complete conversion to TP (n=8937) with data from the same practices a year before using PS. The study found similar increases in diagnosis of LSIL+ in TP compared with PS, though no biopsy correlation reported for this data.	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> no support Whilst found increases in LSIL+ detection in 100% converted practices compared with PS a year before, may be reasons why this sub-sample converted 100% which reflect a different spectrum of disease in their women, or practitioner characteristics. A better test of selection bias would have involved comparing characteristics of women offered TP and those offered PS by the same practitioner in the same time-period. As different populations sampled and prevalence may differ between tests, PPV comparison may not be valid. Assume that longer time-lag between cytology and histology for PS group.

Table 2. Evidence table of primary research studies investigating liquid-based slide preparation devices - ThinPrep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Ashfaq et al., (1999)	Level of evidence: 3b Historical cohort for Pap test, direct-to-vial study (with nested case-control).	Pap test (variable sampling instrument) vs. ThinPrep (instrument not reported) Choice of test in TP cohort made by physician.	Dallas, USA 5/98 – 8/98 for TP. Historical data from 1/97 – 12/97 for PS, from same clinics, though different physicians can change. Also, all biopsy confirmed glandular cases from cohort compared to cytology diagnoses.	43,289 PS 25,783 TP Large inner-city teaching hospital with high rates of glandular abnormality.	Positive test results included AGUS, AIS and adenocarcinoma. (i) Available biopsy follow-up for cases of AGUS, adenocarcinoma in situ (AIS), and adenocarcinoma. (ii) Available cytology retrieved for patients with known AIS or adenocarcinoma in the sample.	(i) Comparing glandular cytology diagnoses with available biopsy results, there was no significant difference in detection rates of glandular cytology between PS and TP. A higher detection of biopsy-confirmed glandular lesions were (mis)diagnosed as squamous rather than glandular by PS (14 cases) compared with TP (4 cases, $p < .05$). (ii) Comparing known biopsy-confirmed glandular cases of AIS or adenocarcinoma with available cytology, there was no difference in PPV detection of glandular abnormalities (AIS, adenocarcinoma, carcinoma) between TP (10/26=35%) and PS (9/39=23%). However, a lower false negative rate was detected for TP (4/26=15.4%) than PS (17/39=43.6%, $p < .02$).	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> partial support Reports that the false negative rate for PS is consistent with other reports. Suggests that upon review of the cytology samples for the false negative results, atypical glandular cells could not be seen in 10 of PS and 2 of TP smears suggesting that the false negatives tended to be related to sampling/preparation errors in the Pap tests rather than reading errors. Authors conclude that TP allowed for lower rates of non-specific misreading of squamous lesions as glandular, and reduced false negative rates compared with PS. As longer follow-up for PS, may expect greater regression of lesions and higher false negative rate than for more recently followed-up TP smears.

Key:

PS: Conventional Papanicolaou Smear test

ASCUS: atypical squamous cells of undetermined significance

HSIL: high-grade squamous intraepithelial lesion

PPV: Positive Predictive Value

CT: Cytotechnologist

TP: ThinPrep test

AGUS: atypical glandular cells of undetermined significance

Se: Sensitivity(Pap threshold/Ref Stand threshold)

Rel TPR: relative true positive rate

CP: Cytopathologist

Ref Stand: Reference standard

LSIL: low-grade squamous intraepithelial lesion

Sp: Specificity(Pap threshold/Ref Stand threshold)

Rel FPR: relative false positive rate

Level of evidence

(1a) Different tests done on randomly allocated individuals (direct-to-vial, randomised controlled trial)

(1b) Different tests done on randomly allocated groups (direct-to-vial)

(2) All tests done on each person (split sample)

(3a) Different tests done on different individuals, not randomly allocated, and recruited concurrently (direct-to-vial)

(3b) Different tests done on different individuals, not randomly allocated, with historical cohort (direct-to-vial)

Recruitment: How was the study sample collected? random, consecutive, not random (neither consecutive nor random)*Blind verification:* Was the test and reference standard measured independently (blind to each other)? yes (e.g. biopsy interpreted by different laboratory that read smears, or panel members different to those making initial diagnosis and blind to these results) no (independence not stated)*Ref Standard:* Was the test compared with a valid reference standard? (i.e. independent, and within 3 months of test) histology (biopsy/negative colposcopy); panel review (cytological review by independent panel)*Verification:* Was the decision to perform the reference standard independent of the test results? all positives and negatives; positives and random fraction of negatives; positives and selected sample of negatives, positives only; none*Industry:* What was the industry's relationship to the study? No support (not done or funded by industry); partial support (some support from industry); total support (totally funded by industry)

4.2 AUTOCYTE PREP

Secondary research

The review of the Australian Health Technology Advisory Committee (AHTAC) (1998) described in Table 1 also included AutoCyte Prep. From research published up until July 1997, the review found that there were fewer studies of AutoCyte Prep than ThinPrep. These revealed some evidence of higher yields of abnormal test results for AutoCyte Prep compared to the Pap test; however, without adequate verification there was insufficient data to estimate sensitivity, or specificity. Review conclusions described for ThinPrep also applied to AutoCyte Prep. That is, the interpretation of evidence of AutoCyte Prep's accuracy is difficult given limitations of study designs and implementation, and further research is required. The increased uptake of AutoCyte Prep at that time was not recommended.

Only two systematic reviews in addition to the AHTAC review (Australian Health Technology Advisory Committee, 1998) considered AutoCyte Prep's effectiveness. Reviewing data from split-sample studies published no later than 1997, Austin and Ramzy (1998) concluded that there was an increased yield of abnormalities by liquid-based slide preparation devices including AutoCyte Prep compared to the Pap test. As discussed above (See Table 1, p. 33), there was a lack of information about the review's search strategy and aspects of study quality, including whether a reference standard was used. The UK based research group, School of Health and Related Research (SchARR), considered AutoCyte Prep as well as ThinPrep (Payne et al., 2000). As discussed earlier, the authors concluded that it was likely that the liquid-based screening technique reduces the number of false negative test results (See Table 1, p. 33). However, Payne et al. also note that study designs were limited in terms of lack of verification of all cytological diagnoses making estimates of sensitivity hard to quantify and specificity unknown for the new devices. It is suggested that strategies such as improving uptake of screening and use of more effective sampling instruments are other strategies that may be more cost-effective than an improvement in test sensitivity.

Primary research: Study designs and quality assessments

The search identified only three eligible studies with adequate reference standards comparing the conventional Pap test with AutoCyte Prep (See Table 3, p. 53). All studies appraised included histology by biopsy as their reference standard to verify cytological positives only. Two studies allowed within-subjects comparisons using split-sample designs and the third study compared AutoCyte Prep direct-to-vial (DTV) with the Pap test concurrently. Recruitment of women was assumed to be non-random in all studies. Histological verification was reported as performed by independent laboratory/staff in only one study (Bishop et al., 1998). Industry support for the research was total for one study (Bishop et al., 1998), partial in another, and not indicated for the remaining study.

Primary research: Study results

Considering results at the threshold of HSIL only, no studies enabled determination of sensitivity, specificity, relative true positive rates or relative false positive rates. Whilst positive predictive value (PPV) was reported in a between-subjects direct-to-vial study by Vassilakos et al. (2000), various biases are likely to have led to samples being compared having different prevalence of disease. Therefore comparisons of PPV between AutoCyte Prep and the conventionally prepared smears are not valid. Two split-sample studies allow within-subjects comparisons of yield of abnormalities. These revealed higher rates of smears read as LSIL by AutoCyte Prep than by the Pap test but equivalent rates of reading of smears as HSIL or higher.

Conclusions

Only three studies were identified as eligible for review and none allowed valid determination of AutoCyte Prep's accuracy compared with conventional Pap tests. Differences in yield of lower grade

abnormalities have cost implications that may be unjustified in the absence of evidence of increased detection of HSIL abnormalities. Therefore, as for ThinPrep, the clinical effectiveness of AutoCyte Prep awaits further research employing better design elements and quality controls, including appropriate verification. Recommendations for verification strategies are presented in Chapter 7.

Table 3. Evidence table of primary research studies investigating liquid-based slide preparation devices – AutoCyte Prep

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Bishop et al., (1998)	Level of evidence: 2 Prospective, multi-centre, split-sample trial.	Pap test vs. AutoCyte Prep (cervex brush)	USA Dates of recruitment not reported	8,983 women in eight site multi-centre trial. Sites included three reference laboratories accessing high-risk and referral patients from Kenya and Vietnam, two low-prevalence screening laboratories, and three hospital centres. Excluded women aged below 16 years. Age range was 16-87. LSIL+ prevalence ranged from 1.4% to 11% across sites.	Yield of abnormal diagnoses for AutoCyte Prep and Pap test. Negatives not verified therefore only relative rates can be determined. Positive test results at threshold of LSIL. Verification by limited biopsy follow-up (LSIL+). Different CT's read smears from each matched pair. Only those CT's trained in ACP read those smears.	In order to minimise the effect of inter- and intra-observer variability in cytology diagnosis, report comparisons where one test found a 2 grade plus difference in diagnostic category. ACP produced greater detection of LSIL+ two-grades above diagnosis by PS, than vice versa ($p < .001$). At threshold of LSIL, sample size was very small and only looked at discrepant diagnoses for calculation of Rel TPR. Rel TPR(LSIL) (ACP/PS) = 20/10=2 Rel FPR(LSIL) (ACP/PS) = 16/24 =.67	<i>Recruitment:</i> not random <i>Blind verification:</i> yes <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> total support Possible bias in that CTs trained in ACP may have improved skills or may have been selected for training based on higher skills than those not trained. Rel TPR and Rel FPR could not be determined at threshold of HSIL.

Table 3. Evidence table of primary research studies investigating liquid-based slide preparation devices – AutoCyte Prep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Minge et al., (2000)	Level of evidence: 2 Split sample study	Pap test vs. AutoCyte Prep (cervex brush)	Omaha, USA Dates not given	2,156 Three obstetric-gynaecological specialty practices (high risk population) with estimate prevalence of SIL of 6.5% Women aged 15-57.	Yield of abnormal diagnoses for AutoCyte Prep and Pap test. Negatives not verified therefore only relative rates can be determined. Smears read by separate CTs. Non-blinded review of discordant pairs by CP led to the more abnormal diagnosis being retained. Verification by limited biopsy follow-up of cases (LSIL).	AutoCyte Prep more likely to detect LSIL than PS ($p < .05$), with no difference in detection of HSIL. At threshold of LSIL, biopsy follow-up results (n=81) allowed determination of Rel TPR only from reported data. Rel TPR(LSIL) (ACP/PS) = 38/41 = .93	<i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> partial support Rel TPR and Rel FPR could not be determined at threshold of HSIL. No data on negative biopsies reported, therefore Rel FPR could not be determined.

Table 3. Evidence table of primary research studies investigating liquid-based slide preparation devices – AutoCyte Prep (continued)

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
Vassilakos et al., (2000)	<p>Level of evidence: 3a*</p> <p>Concurrent cohort for Pap test, direct-to-vial study.</p> <p>* However, ACP cohort continued beyond PS cohort.</p>	<p>Pap test (Ayre spatula, Ayre spatula with cytobrush, cotton-tipped applicator, or cervex brush) vs. CytoRich {AutoCyte Prep} (Cytoprep brush, Cytobrush or cervex brush in combination with Ayre spatula)**</p> <p>** information from studies (Vassilakos et al., 1998; Vassilakos et al., 1999) which present earlier data from the same study.</p>	<p>Geneva and Zurich, Switzerland**</p> <p>Smears performed concurrently in both groups over six months in 1998 and then for ACP alone for a further five months**.</p>	<p>PS: 19,923 ACP: 81,120</p> <p>Patient groups derived as a function of physician consulted that routinely used either ACP or PS throughout the study period. Of 97 clinicians, 19 performed PS and 78 used ACP. Practices selected from geographically and socio-demographically similar population. Similar age in both groups.**</p>	<p>Available biopsy follow-up for ACP cases: 9% ASCUS, 21% LSIL, and 80% HSIL+.</p> <p>Available biopsy follow-up was "more thorough" for less recent PS cases: 7% ASCUS, 61% LSIL, and 88% HSIL+.</p>	<p>PPV (HSIL) for ACP (320/357=89.6%). PPV (HSIL) for PS (46/62=74.2%).</p> <p>PPV (LSIL) for ACP (690/857=80.5%). PPV (LSIL) for PS (124/173=71.7%).</p>	<p><i>Recruitment:</i> not random <i>Blind verification:</i> no <i>Ref Stand:</i> histology <i>Verification:</i> positives only <i>Industry:</i> no support</p> <p>Variation in sampling instruments between tests. Whilst spectrum of disease in the "two study populations was assumed to have been stable during the study period" there is a possibility of biases in which physicians had decided to routinely use ACP instead of PS. Only information available for comparing groups was age. As different populations sampled and prevalence may differ between tests, PPV comparison may not be valid.</p> <p>CytoRich uses manual pipetting of the sample and manual staining compared with ACP.</p>

Key:

PS: Conventional Papanicolaou Smear test

ASCUS: atypical squamous cells of undetermined significance

HSIL: high-grade squamous intraepithelial lesion

PPV(Pap threshold/RS threshold): Positive Predictive Value

CT: Cytotechnologist

ACP: AutoCyte Prep Test

AGUS: atypical glandular cells of undetermined significance

Se: Sensitivity(Pap threshold/Ref Stand threshold)

Relative TPR = relative true positive rate

CP: Cytopathologist

Ref Stand: Reference standard

LSIL: low-grade squamous intraepithelial lesion

Sp: Specificity(Pap threshold/Ref Stand threshold)

Relative FPR: relative false positive rate

Level of evidence

(1a) Different tests done on randomly allocated individuals (direct-to-vial, randomised controlled trial)

(1b) Different tests done on randomly allocated groups (direct-to-vial)

(2) All tests done on each person (split sample)

(3a) Different tests done on different individuals, not randomly allocated, and recruited concurrently (direct-to-vial)

(3b) Different tests done on different individuals, not randomly allocated, with historical cohort (direct-to-vial)

Recruitment: How was the study sample collected? random, consecutive, not random (neither consecutive nor random)*Blind verification:* Was the test and reference standard measured independently (blind to each other)? yes (e.g. biopsy interpreted by different laboratory that read smears, or panel members different to those making initial diagnosis and blind to these results) no (independence not stated)*Ref Standard:* Was the test compared with a valid reference standard? (i.e. independent, and within 3 months of test) histology (biopsy/negative colposcopy); panel review (cytological review by independent panel)*Verification:* Was the decision to perform the reference standard independent of the test results? all positives and negatives; positives and random fraction of negatives; positives and selected sample of negatives, positives only; none*Industry* What was the industry's relationship to the study? No support (not done or funded by industry); partial support (some support from industry); total support (totally funded by industry)

Chapter 5: Effectiveness of automated devices for primary screening and re-screening

5.1 AUTOPAP

The search strategy identified five systematic reviews in addition to the AHTAC review (Australian Health Technology Advisory Committee, 1998), and one meta-analysis, which considered AutoPap's effectiveness. These papers may not employ the same selection criteria as used in this review (See Table 4, p. 60).

Secondary research

The review of the Australian Health Technology Advisory Committee (AHTAC, 1998), which the present report updates, was based on research published up until July 1997. The AHTAC review found that there were few peer-reviewed clinical studies of AutoPap in rescreening mode (i.e. AutoPap 300 QC) and no studies evaluating the AutoPap System used as a primary screener. Studies evaluating AutoPap suggest an increased detection of abnormalities relative to 10% random rescreening (which is mandated by CLIA in the USA, but is not common practice in New Zealand). However, studies rarely provided information about participating laboratories, or provided biopsy confirmation of positives. Therefore some increases may reflect false positives. The review did not recommend increased uptake of AutoPap, or other automated devices, in a national screening programme from a public health perspective at the time of the review.

There have been several other systematic reviews since the AHTAC report. A study of cost effectiveness by Brown and Garber (Brown and Garber, 1998) for the Blue Cross & Blue Shield Association (adapted for a later publication in JAMA (Brown and Garber, 1999a)), included a systematic literature review of papers published prior to 1998 which used biopsy or expert panel review as their reference standard. Four studies were appraised which compared AutoPap 300 QC as a rescreener of up to 20% of the most abnormal cytologically negative slides (i.e. 20% review rate) in comparison with 10% random manual rescreening. From one of these studies, 10/13 (77%) LSIL+ false negatives were detected at a 20% review rate (this represents *rescreening* sensitivity and refers to sensitivity to false negative abnormalities not detected by primary screening). However, Brown and Garber comment that AutoPap's rescreening sensitivity may be less than alternative means of manual rescreening such as targeted review of slides based on risk status. However, such high-risk slides are not intended to be read by AutoPap¹⁸. It concludes that rescreening sensitivity of AutoPap is reliant on the performance of the cytologists in rescreening selected slides, and that there is little information about specificity.

A systematic review from Minnesota (Minnesota Health Technology Advisory Committee (HTAC), 1999) of papers published to March 1998 included four studies, three on rescreening and one of AutoPap in primary screening mode (Wilbur et al., 1998) which is also reviewed here. Reported sensitivity estimates ranged widely due to different review rates used by AutoPap in rescreening mode. Studies were criticised as using enriched samples not relevant to intended use and read by cytologists with heightened vigilance under trial conditions. Whilst in rescreening mode, detection of false negatives is modestly increased compared to random 10% rescreening, the abnormalities tend to be lower grade abnormalities (LSIL, or ASCUS). Furthermore, false positives are increased and the longer-term patient and societal costs of these have not been determined (see Chapter 6 for a

¹⁸ The FDA's Premarket approval order statement states that the AutoPap System is "not intended to be used on slides designated by the laboratory as high risk" (FDA PMA Number P950009, Supplement Number S003, 19 January 1999). Whilst high-risk slides may be processed by the AutoPap System, slides designated for no further review must be rescreened manually.

discussion of these issues). The review recommends further encouragement of regular Pap testing and research into the safety, clinical effectiveness and cost effectiveness of this device.

A meta-analysis of AutoPap 300QC of studies published to October 1998 (Abulafia and Sherer, 1999) included 14 studies, 13 of which were prepared by various members of the same core group of researchers suggesting that they are not truly independent. Evidence was limited to four studies of AutoPap as a primary screener (reported sensitivity to all abnormalities ranging between 85%-100%) and five studies reporting on AutoPap as a rescreeener at 10% review (rescreening sensitivity to false negatives alone estimated as 37%; 95% CI, 34-40). However, the review does not consider aspects of study quality, such as use of the type of reference standard used, in its selection of studies for data synthesis. Studies tended to use inadequate reference standards, such as solo pathologist review, or enriched samples. There is some concern about the appropriateness of a meta-analysis given the heterogeneity of the studies, and the questionable study quality of some.

Duke University's review (McCrory et al., 1999) for AHRQ included six studies which estimated sensitivity of AutoPap as a rescreeener, and as a primary screener (note that in two of these studies a 20% review QC approach was used on initial slide readings rather than on slides screened as negative). There was little information on specificity. The review suggested that further research was needed which allowed the estimation of specificity, and used an histological reference standard. This review team completed a more recent review (Nanda et al., 2000) based on literature published through to October 1999. This review was restrictive and found that no studies met the selection criteria for inclusion relating to AutoPap. Some studies were excluded because, as rescreening is conditional on a negative initial screen, the two screening tests are not applied independently. The lack of studies of sufficient quality related primarily to three areas of design limitations. These included: lack of prospective studies applying both tests to the same women, failure to verify negative test results adequately, and little evidence on specificity.

A recent update review of AutoPap by ECRI's Health Technology Assessment Information Service appraised papers no later than December 1999 (ECRI Health Technology Assessment Information Service, 2000). Only a prospective trial evaluating the AutoPap Primary Screening System conducted as part of FDA PMA submission (reported in two papers) was regarded as of adequate quality (Wilbur et al., 1998; Wilbur et al., 1999)¹⁹. AutoPap was described as having improved sensitivity compared with manual Pap testing at thresholds of ASCUS+ and LSIL+, with equivalent specificity at a threshold of LSIL+. However, such test characteristics could not have been directly determined as concordant diagnoses were not verified. An additional study (Bibbo and Hawthorne, 1999; Bibbo et al., 1999) was criticised for not reflecting intended use (e.g. screened known-abnormal slides or may have screened the slides with the device several times) and using an inadequate reference standard of single pathologist review. The review concluded that there was limited evidence supporting the use of the AutoPap System for both primary and rescreening of normal-risk slides. However, the review recommends that further large-scale studies under normal laboratory conditions be conducted to replicate these results, using reliable reference standards to verify all diagnoses.

To summarise, the systematic reviews above generally identified few peer-reviewed clinical studies evaluating AutoPap, particularly the currently available AutoPap System. Reported estimates of sensitivity varied depending on review rates employed and the quality of studies included; the two most recent reviews reported very limited evidence as valid (ECRI Health Technology Assessment Information Service, 2000; Nanda et al., 2000). Whilst some studies suggested a modest increase in sensitivity of AutoPap as a rescreeener compared to 10% random rescreening mandated in the United States, it has been argued that there may be less advantage when AutoPap is compared to alternative means of manual rescreening such as targeted review. Moreover, the abnormalities have tended to be lower grade or borderline abnormalities, possibly at the cost of reduced specificity, though there was very limited information about this test characteristic. A prospective trial of AutoPap System as a primary screener was claimed as demonstrating improved sensitivity and equivalent specificity compared to manual screening at a threshold of LSIL+ (Wilbur et al., 1998). This study is appraised in this chapter.

Given the limited information on AutoPap's test characteristics, studies have not generally recommended increased uptake in population-based screening programmes at this stage, and more

¹⁹ The same trial is appraised in this Chapter (p. 64)

rigorous research is recommended. One review reports limited evidence of the AutoPap System.

Table 4. Secondary research appraised relevant to AutoPap

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
Australian Health Technology Advisory Committee (1998) Australia	Search: 1990 - July 1997 Databases searched: Medline, CancerLit. Also queried key device users and device manufacturers, and accessed material from Food and Drug Administration Premarket Clinical Trials. Conference papers, information from manufacturers, internet searches, press releases and non-peer reviewed publications were accessed as background material. Keywords included automated cytology, AutoPap, Papnet, Pap test screening, cervical screening, cytology screening, cytopathology, automated screening, vaginal instrumentation.	Experimental studies evaluating AutoPap as a <i>rescreeener</i> were reviewed. Critical appraisal performed where "sufficient evidence from well-designed studies" to provide estimates of sensitivity, specificity, additional cases detected and costs.	There were few peer-reviewed clinical studies of AutoPap and no randomised controlled trials. AutoPap in primary screening mode was not evaluated. Results evaluating AutoPap for QC rescreening suggest an improvement in detection of abnormalities relative to 10% random rescreening. However, random rescreening is not mandated or common QC practice in Australia or New Zealand, due to the poor potential yield of false negative cases achieved by this approach. Studies rarely provided information about participating laboratories, or false positives. Biopsy confirmation of positives was generally not used. There appeared to be higher rates of rescreened smears as abnormal than one would expect (i.e. higher than 1%).	Lack of consistency in design, initial sample selection, and measurement of outcomes made interpretation of evidence difficult. Suggest a need for local population-based studies with devices used in the context of routine clinical practice which clearly demonstrate the device's cost effectiveness. AHTAC's expert working party did not recommend increased uptake from a public health perspective at the time of the review. However, argued that the area is in constant change and refinements to devices could improve their performance and lower their cost. (Also see Chapter 6 on costs)
Brown and Garber (1998); Brown and Garber (1999a) USA	Search: 1987 - December 1997 Databases searched: Medline, HealthStar, CINAHL, EMBASE, EconLit. Also hand searched journals and queried experts and device manufacturers. Included articles, abstracts and manuscripts. Key words: various cervical cytological tests including AutoPap with cervical cancer or neoplasia and sensitivity and specificity.	Included studies (including abstracts) of AutoPap QC300's accuracy which had adequate reporting, and which were written in English. Included studies (including abstracts) of AutoPap's accuracy which had adequate reporting, and which were written in English. Appraised studies reporting the number and cytological results from all slides, reported FDA approved use of the technology, used biopsy of expert panel review as reference standards, and included slides with a validated diagnosis of LSIL+.	N=4 studies all evaluating AutoPap 300 QC in the rescreening mode. Estimate Se from one study (Colgan et al., 1995) as 77% at 20% review based on detection of 10/13 LSIL+ false negatives. With 80% Se at primary manual screening this leads to an Se estimate of 95.4%. However, AutoPap may not be more sensitive than targeted selection of slides based on risk-status. Also, a substantial proportion of slides may not be amenable to review by AutoPap. Such slides are excluded from studies, which could greatly underestimate the rate of false negatives reported.	Not restricted to published literature. Argues that the system is only as sensitive as the CTs performing manual review. There is little information about specificity. (Also see Chapter 6 on costs)

Table 4. Secondary research appraised relevant to AutoPap (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
Minnesota Health Technology Advisory Committee (HTAC) (1999) Minnesota, USA	Search: 1993 - March 20 1998 Databases searched: Medline, Current Contents, FDCR, and the National Cancer Institute PDQ. Key words: Pap, various cervical cytological tests including AutoPap, vaginal smears and diagnosis, mass screening, vaginal smears, cytodiagnosis, and cytology.	Included 4 clinical trials evaluating AutoPap.	N=4 studies, three in QC rescreening mode and one on primary screening mode (Wilbur et al., 1998). Se reported ranged from 35-77, with variation due to differing selected review rates tested. Study on primary screening demonstrated improved detection rates without loss of specificity (Wilbur et al., 1998). Biases in designs include using "enriched" samples that do not represent the routine practice or intended use for AutoPap. Heightened vigilance of CTs in trials also expected to affect results. Computer assisted devices (including AutoPap) reduce modestly the number of false positives smears compared to random manual rescreening. However, majority of false negatives are low-grade lesions. Also, additional slides identified for manual review include false positives.	No studies investigated effects of false positives on patient outcomes (including anxiety) and health costs (including additional testing and clinical procedures). The added value of new devices (including AutoPap) in improving the net health outcome of women and preventing cervical cancer has also not been determined. Recommends further encouragement of regular Pap testing, and research into the safety, clinical effectiveness and cost effectiveness of the new devices.
Abulafia and Sherer (1999) USA	Search: to October 1998 Databases searched: Medline, and references of papers. Key words: various cervical cytological tests with cervical cancer or neoplasia and sensitivity and specificity.	Included studies of AutoPap 300 QC's accuracy reported in English. Meta-analyses performed to estimate the over-all false negative rate in primary screening and rescreening modes. Studies included in meta-analyses if providing complete data about the number of abnormal slides, review rate, and number of slides selected.	N=14 Note that in 13 of 14 studies, the same basic group of researchers are involved with different subsets authoring papers. This suggests that these studies are not truly independent and that systematic interests/biases may exist and results of the meta-analysis should be interpreted with caution. Estimate of Se for 4 studies relating to primary screening range from 85%-100%. Estimate of Se for 5 studies relating to re-screening at 10% review = 37% (95% CI, 34-40).	The authors conclude that there is a paucity of data regarding the application of AutoPap 300 QC, and the majority have been performed by various members of the same core group of researchers. Limited studies suggest Se estimates reported here. Suggests that independent studies are required to confirm these findings. Note that studies compared included some that do not use reference standards or used varying reference standards, which would influence the Se estimates. Given these variations and the lack of control over study designs included here, one must be cautious about interpreting results.

Table 4. Secondary research appraised relevant to AutoPap (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
McCrory et al., (1999) USA	<p><u>Search:</u> to January 1999</p> <p><u>Databases searched:</u> Medline, CancerLit, Health STAR, CINAHL, EMBASE, EconLit. Also hand searched journals, bibliographies of included papers, and queried experts and device manufacturers.</p> <p><u>Key words:</u> various cervical cytological tests including AutoPap with cervical cancer or neoplasia and sensitivity and specificity.</p>	<p>Included studies (including abstracts) of AutoPap's accuracy reported in English, with reference standard of cytology or histology. Papers reporting cost and health outcomes needed to assess the effect of screening on life expectancy or quality, number of cervical cases avoided, or total health care costs.</p> <p>Papers assigned quality scores according to predetermined methodological criteria.</p>	<p>N= 6 studies allowing estimates of Se.</p> <p>AutoPap in QC mode: Se (ASCUS+) 20% review = .43, .51, .67. Se (LSIL+) at 20% review = .66, .77</p> <p>Primary screening mode: Se (ASCUS+) 20% review = .86, .86, .92 Se (LSIL+) 15% review = .92</p> <p>Little information permitting determination of Sp. One study found Sp= .98 based on primary screening.</p> <p>Note that some studies used AutoPap 300 QC which uses a different algorithm to AutoPap System.</p>	<p>Suggests that future research should include verification of test-negative subjects to allow estimation of specificity, preferably by colposcopy. Also suggests a need for research where verification of diagnoses use histological reference standards rather than cytological reference standards.</p> <p>(Also see Chapter 6 on costs)</p>
Nanda et al., (2000) USA <u>Note:</u> this review uses the same search strategy as McCrory et al. (1999), but uses tighter selection criteria.	<p><u>Search:</u> through October 1999</p> <p><u>Databases searched:</u> Medline, CancerLit, HealthStar, CINAHL, EMBASE, EconLit. Also hand searched journals, bibliographies of included papers and recent systematic reviews, and queried experts and device manufacturers.</p> <p><u>Key words:</u> various cervical cytological tests including AutoPap with cervical cancer or neoplasia and sensitivity and specificity.</p>	<p>Included studies (including abstracts) of AutoPap's accuracy reported in English, which met following criteria:</p> <p>(1) Study must prospectively compare screening test and reference standard on the same set of patients or slides; (2) if using a cytological reference standard, must use an adjudicated independent panel review of discordant negatives; (3) at least 50% of patients testing positive for HSIL must be verified by an histology/colposcopic reference standard; (4) must allow separate analyses of Se (or relative true positive rate) <u>and</u> Sp (or relative true negative rates).</p> <p>Papers assigned quality scores according to predetermined methodological criteria.</p>	<p>No studies met the selection criteria.</p> <p>Studies that applied manual screening followed by computerized rescreening could not be evaluated because rescreening is conditional on a negative initial screen; thus, the two tests are not applied independently.</p>	<p>The authors argue that deficiencies in study designs relate to the complete lack of evidence for AutoPap (and indeed, a lack of papers relating to new devices generally).</p> <p>Three main areas of design limitations include: lack of prospective studies applying both tests to the same women/samples; failure to verify negative test results adequately; little evidence on specificity.</p>

Table 4. Secondary research appraised relevant to AutoPap (continued)

Source	Search method	Criteria for inclusion/exclusion	Results	Comments
ECRI Health Technology Assessment Information Service (2000)	Search: completed December 1999 (however, some databases searched until September 1999 only)	Studies comparing AutoPap System to primary manual reading of smears with the USA mandated random manual rescreening of 10% of negative smears.	N=1 study (two papers)	Concludes that limited evidence supports the use of AutoPap system for both primary and rescreening of normal-risk slides.
USA	Databases searched: CancerLit, Cochrane databases, CINAHL, CRISP, Current Contents, ECRI library, EMBASE, Health Device Alerts, Health Devices Sourcebase, Health Care Financing Administration, Healthcare standards, HSRProj, HealthStar, HSTAT, IHTA, Medline, National Guideline Clearing House, TARGET, USA FDA web site. Also conference proceedings, bibliographies of retrieved papers, and non-peer reviewed "grey literature". Key words: AutoPap, extensive list of terms relating to Pap testing, cervical diseases, screening, cytology, and automated testing.	Considered use of AutoPap Primary Screening System in studies where it was used in an "intended-use" manner. That is, not used for smears from known high-risk women.	Appraises a prospective trial included in this review (Wilbur et al., 1998) which evaluates the AutoPap Primary Screening System as part of the FDA PMA submission. Results indicate that AutoPap has improved Se compared with manual Pap test screening alone, at thresholds of ASCUS+ and LSIL+. Sp was high both for AutoPap and manual screening and equivalent at threshold of LSIL+, though higher for AutoPap at a threshold of ASCUS+. An additional study reported by two papers (Bibbo and Hawthorne, 1999; Bibbo et al., 1999) did not reflect intended use and used an inadequate reference standard of single pathologist review.	However, the review suggests that there is some question about whether these results represent everyday laboratory practice. In the reviewed study, Se of the Pap test screening itself appeared to be inflated compared to previous research. Recommends that further large-scale studies be conducted to replicate these results, using reliable gold standards (such as panel review) to render final diagnoses on all positives and negative slides. Also suggests that normal laboratory practice needs to be established in studies. Notes that patient outcomes are not currently reported in the published literature.

Key:

PS: Conventional Papanicolaou Smear test

CT: Cytotechnologist

LSIL: low-grade squamous intraepithelial lesion

Sp: Specificity(Pap threshold/Ref Stand Threshold)

AP: AutoPap test

CP: Cytopathologist

HSIL: high-grade squamous intraepithelial lesion

PPV: Positive Predictive Value

Ref Stand: Reference standard

ASCUS: atypical squamous cells of undetermined significance

Se: Sensitivity(Pap threshold/Ref Stand threshold)

Primary research: Study results

Excluded studies which evaluated AutoPap included those with no or an inadequate reference standard (e.g. single pathologist review), those without a comparison group of manual screening, and those which included high-risk slides for which the device was not intended.

Only one study reported as two papers (Wilbur et al., 1998; Wilbur et al., 1999) was eligible for appraisal (See Table 5, p. 67). This was a prospective, two-armed, intended-use trial of the AutoPap System (which adds a new algorithm to the AutoPap 300 QC that was not rendered Y2K compliant and is no longer commercially available). The trial is totally financed by industry, and contributed to the FDA PMA submission for the AutoPap System.

The trial compared “current practice” of manual primary screening and 10% random rescreening with AutoPap System in combined primary and rescreening modes. As described in chapter 1, the AutoPap System designates slides as either for *No Further Review* (up to 25% of slides processed) or *Review*. The *Review* slides are ranked according to their likelihood of being abnormal and screened manually by cytologists (at least 75% of slides processed). Of the *Review* slides subsequently read as within normal limits, at least 15% ranked by AutoPap as most likely to be abnormal are selected for manual QC rescreening or *QC Review*. It should be noted that some study slides which failed processing by AutoPap due to physical characteristics or insufficient cellularity (*Process Review* and *ReRun* slides) were excluded from both arms of this trial and would normally be read manually. The number of these was not provided.

The trial involved five commercial laboratories in the USA which each supplied 150 consecutively collected slides per day over an unspecified period. Unlike some other studies (Bibbo and Hawthorne, 1999; Bibbo et al., 1999) slides were excluded if sampled from high-risk women such as those with abnormal presentation, current complaint or medical risk factor, in order to approximate intended use conditions.

Features of study quality included consecutive recruitment, blind interpretation and verification, and expert panel cytology review with limited biopsy follow-up. A three member external expert panel considered discrepant diagnoses (only) between tests. However, consensus diagnoses were only sought where the panel did not achieve majority agreement (i.e. by two of three members). This approach contrasts with the guidelines recommended by the Intersociety Working Group for Cytology Technologies (ISWG) for primary screening (Intersociety Working Group for Cytology Technologies, 1997) which suggest that consensus must be achieved for all included slides. Verification is also limited as the panel verifies only discordant results; concordant positives as well as concordant negatives are considered the truth. This is problematic. One could imagine that the more obvious abnormalities are likely to be read as positive by both, and difficult to interpret abnormalities may lead to false negative diagnoses by both. Such a design can be expected to lead to an over-estimate of both sensitivity and specificity (Miller, 1998). In this situation, results should be reported as the number of additional true positives found by one test over another (Chock et al., 1997). Results at a threshold of HSIL reveal no statistical difference in the number of true positives detected by AutoPap compared with conventional practice.

The ISWG (Intersociety Working Group for Cytology Technologies, 1997) recommend that in determination of PPV, a reference standard of biopsy follow-up be used of a statistically significant subset of patients with a positive cytological diagnosis (that different reference standards are suggested for Se and PPV is queried by Brown and Garber, (1998)). In this trial, histological follow-up was only available for 29 of 70 HSIL+ slides (Wilbur et al., 1999). ECRI excluded this study from their review (ECRI Health Technology Assessment Information Service, 2000). There were concerns that the small proportion may represent a high-risk group with recurrent disease followed up more aggressively, which would lead to an inflated positive predictive value (Karyn Tappe, Senior Research Analyst, ECRI, June 2000, *personal communication*). Whilst this trial actually excluded high-risk women, there are concerns about the influence of work-up bias. For similar reasons, the review by Nanda et al. (2000) only considered studies where at least 50% of smears read as HSIL are followed up histologically. Whilst results of biopsy follow-up for HSIL+ slides are reported, PPV could only be calculated for AutoPap System ($21/27 = 78\%$) for this small sample of slides. Without comparison

data for conventional screening, PPV is not a useful outcome given lack of information about the prevalence in the population screened.

It should be noted, as observed of studies generally in this area (Brown and Garber, 1998), that a substantial proportion of slides are not be amenable to processing by AutoPap and are excluded from studies' results. Whilst this may improve with experience, the removal of such unprocessable slides could greatly underestimate the rate of false negatives reported.

The FDA trial reported Pap smear Se of 85.6% at a threshold of LSIL and Sp of 99.6% (though direct estimates were not possible). This level of accuracy for conventional screening is high compared to results of a recent review of studies evaluating the Pap test (Nanda et al., 2000). This suggests that researchers took extraordinary care in screening slides in both arms of the trial, and that sensitivity for the Pap test and AutoPap are likely to be inflated.

Finally, there has been confusion in the interpretation of this trial concerning whether it investigated use of AutoPap System as a combined primary screener and rescreener. The study's aims and methods suggest that both modes were employed and that the comparison group involved manual primary screening followed by 10% random manual rescreening. However, some reviews have reported the results as reflecting the outcome of primary screening only (ECRI Health Technology Assessment Information Service, 2000; McCrory et al., 1999). Following correspondence with ECRI about this interpretation, ECRI contacted TriPath Imaging for clarification. It was confirmed that the diagnoses after both primary and QC rescreening were reported in the trial's results (Karyn Tappe, ECRI, June 2000, *personal communication*). In their reply to ECRI, TriPath Imaging comment, "Final diagnoses documented for each arm of the study were based on the highest level of screening for each slide...if a slide in the Review population was diagnosed negative on first manual screen and this same slide was found to be designated as QC Review by the AP (AutoPap) System the final diagnosis documented for the study would be the QC diagnosis....The overall AP system of ranking used in combination with the enriched QC population contributed to the final results that were reported in the product insert and other publications. In summary, the QC population was not analysed separately" (TriPath Imaging, June 2000, *personal communication*).

Therefore, from the FDA trial of Wilbur and colleagues it is not possible to determine whether any improvements in accuracy of AutoPap compared to conventional screening relate to AutoPap's use in primary or rescreening modes. Findings from this study are therefore only relevant where comparing the AutoPap System to conventional screening followed by 10% random rescreening. It should be noted that as an additional 5% of slides were selected by AutoPap for QC review in the trial compared to 10% selected randomly for review in conventional screening, one would expect some improved detection of false negatives by chance alone. The same results would not be found when using alternative rescreening strategies including targeted review of slides from high risk populations (such slides are not FDA approved for processing by AutoPap), rapid review or full review of all within normal limits slides.

Conclusions

Since the AHTAC review (that did not find sufficient evidence to recommend employing AutoPap from a public health perspective), only one study eligible for appraisal was identified evaluating this device. This concerned AutoPap System applied as a primary screener and rescreener combined. This prospective trial of reasonably good quality but did not permit direct estimates of test sensitivity and specificity due to limited verification. Comparing true positives verified histologically from discrepant test diagnoses for AutoPap and conventional screening, the study concluded that there was no difference in detection between the tests at a threshold of HSIL. Other systematic reviews have also commented on the lack of high quality evidence in this area. From very limited evidence, some have suggested that there is a modest increase in sensitivity of AutoPap as a rescreener compared to 10% random rescreening. However, these results may have less relevance to New Zealand where other more effective manual rescreening strategies are more commonly employed than 10% random review. Importantly, any increase in sensitivity has tended to be associated with the identification of lower grade or borderline abnormalities. There is no reliable evidence of the specificity of AutoPap.

Further research into the clinical effectiveness and cost effectiveness of this device is required. Large-scale prospective trials performed under normal laboratory conditions are recommended, using reliable gold standards to verify all diagnoses. Recommended approaches for verification are discussed in Chapter 7.

Table 5. Evidence table of primary research investigating AutoPap

Source	Study design	Comparison interventions	Location & dates of testing	Sample	Outcomes and verification	Results	Quality
<p>Wilbur et al., (1998)</p> <p>Wilbur et al., (1999) for results relevant to HSIL+ and biopsy follow-up</p>	<p>Level of evidence: 1</p> <p>Prospective, two-armed intended use trial.</p>	<p>Manual screening and 10% random QC rescreening (MS) vs. AutoPap System - Primary Screener (AP).</p> <p>The AutoPap designates each slide as:</p> <ul style="list-style-type: none"> - <i>No Further Review</i>: within normal limits (WNL) requiring no further review (up to 25%) - <i>Review</i> where slides are ranked into quintiles of the most to least likely to be normal and receive manual CT screening (at least 75%). - <i>QC review</i> Of those read as within normal limits (negative), at least 15% (of all slides processed) ranked as most likely to be abnormal are then re-screened manually. - <i>Process review</i> or <i>Rerun</i> slides that failed processing were excluded. 	<p>USA</p> <p>Time period not specified.</p> <p>5 commercial laboratories (patient populations at each site were "analyzed to ensure that slides represented a valid cross-section of the population", though no data was reported).</p>	<p>25,124 slides eligible from total of 31,507.</p> <p>150 slides per day from each lab collected consecutively.</p> <p><i>Inclusion criteria</i>: the smear was taken from the correct anatomical site, with a correct sampling device, and physically acceptable for processing*.</p> <p><i>Exclusion criteria</i>: slides "with "physical limitations, those with missing slide documentation, and "high risk" slides (from patient's records, e.g. abnormal presentation; a current complaint; or medical risk factor).</p>	<p>TPR</p> <p>Positive test results at threshold of ASCUS.</p> <p>Verification by independent, masked review by panel of 3 CPs of discrepant positive smears by majority agreement, or failing this, by adjudication at a multi-headed microscope of panel members.</p> <p>Different CT's read smears from each matched pair, and were blind to results the other test.</p> <p>Available biopsy follow-up of 27 HSIL slides (Wilbur et al., 1999).</p>	<p>Note that as concordant negatives and concordant positives were not verified, Se and Sp cannot be determined directly. Test diagnoses were only reported for those slides read as positive by panel cytology review.</p> <p>AutoPap detected 23 more true positives at threshold of LSIL than PS; 321 compared with 298 (Wilbur et al., 1998) (significant).</p> <p>AutoPap detected 3 more true positives at threshold of HSIL than PS; 68 compared with 65 (Wilbur et al., 1999) (no significant difference).</p> <p>From biopsy follow-up (Wilbur et al., 1999): PPV (HSIL) for AutoPap only = 21/27 = 78%</p> <p>No biopsy follow-up data reported for manual screening.</p>	<p><i>Recruitment</i>: consecutive</p> <p><i>Blind verification</i>: yes</p> <p><i>Ref Stand</i>: panel review, and limited biopsy</p> <p><i>Verification</i>: positives only</p> <p><i>Industry</i>: total support (FDA trial)</p> <p>* Slides which failed AutoPap processing (Rerun or Process Review) were excluded from the trial (and were manually reviewed).</p>

Key:

PS: Conventional Papanicolaou Smear test

CT: Cytotechnologist

LSIL: low-grade squamous intraepithelial lesion

Sp: Specificity(Pap threshold/Ref Stand Threshold)

AP: AutoPap test

CP: Cytopathologist

HSIL: high-grade squamous intraepithelial lesion

PPV: Positive Predictive Value

Ref Stand: Reference standard

ASCUS: atypical squamous cells of undetermined significance

Se: Sensitivity(Pap threshold/Ref Stand threshold)

Level of evidence

(1) All tests done on each person/slide

(2a) Different tests done on randomly allocated individuals (randomised controlled trial)

(2b) Different tests done on randomly allocated groups

(3a) Different tests done on different individuals, not randomly allocated, and recruited concurrently

(3b) Different tests done on different individuals, not randomly allocated, with historical cohort

Recruitment: How was the study sample collected? random, consecutive, not random (neither consecutive nor random)*Blind verification:* Was the test and reference standard measured independently (blind to each other)? yes (e.g. biopsy interpreted by different laboratory that read smears, or panel members different to those making initial diagnosis and blind to these results) no (independence not stated)*Ref Standard:* Was the test compared with a valid reference standard? (i.e. independent, and within 3 months of test) histology (biopsy/negative colposcopy); panel review (cytological review by independent panel)*Verification:* Was the decision to perform the reference standard independent of the test results? all positives and negatives; positives and random fraction of negatives; positives and selected sample of negatives, positives only; none*Industry:* What was the industry's relationship to the study? no support (not done or funded by industry); partial support (some support from industry); total support (totally funded by industry)

Chapter 6: Economic evaluations of cervical screening devices

6.1 INTRODUCTION

The search strategy identified four economic studies in addition to the AHTAC review (Australian Health Technology Advisory Committee, 1998) that investigated the cost effectiveness of introducing a “new” cervical screening device (i.e. a semi-automated or automated device) into a screening programme. All appraised studies are listed in **Table 6** below.

Table 6. Economic evaluations appraised

Authors	Year	Country	Primary funders	Device
Australian Health Technology Advisory Committee (AHTAC)	1998	Australia	AHTAC (for the Australian Federal Government)	Generic (liquid-based and automated rescreening)
Brown and Garber	1998, 1999	USA	Technology Evaluation Centre of the Blue Cross-Blue Shield Association health (insurance) plan	ThinPrep, AutoPap 300 QC, Papnet
McCrary et al. Duke University	1999	USA	Agency for Healthcare Research and Quality (AHRQ)	Generic (liquid-based and 100% automated rescreening)
Smith et al.	1999	USA	Industry (NeoPath)	AutoPap System
Payne et al. Sheffield University	2000	UK	National Institute for Clinical Effectiveness (for the NHS)	Liquid-based screening

Study designs

There were no randomised controlled trials or field studies identified assessing the economic impact of introducing new cervical screening devices compared with the conventional Pap test. Therefore we do not know the impact of introducing a new device on major clinical outcomes (e.g. incidence of invasive cancer, mortality, life-years gained) or comprehensive economic outcomes (which may include direct and indirect financial, social and psychological costs to the health system, women screened and society generally).

Markov models

The economic studies were all *disease state transition models* or *Markov models*. These track movement of a cohort of women between health states (representing progression of disease) during a fixed time interval (Drummond et al., 1997). For models appraised here of cervical screening, a hypothetical cohort of women are tracked over repeated screening intervals until death, from cervical cancer or other cause. (An exception is the AHTAC model, which is applied to a single screening cycle). The models are based on transition probabilities between the different health states for each time period of the model (e.g. a year). For example, a well-accepted pathologic model includes the following health states: healthy, low-grade disease, high-grade disease, stage I/II cancer, stage III/IV cancer, death from cervical cancer or other causes (Eddy, 1990). Markov models assume that the transition probabilities depend only upon the health state the patient is in and not on how long they have been in that health state or how they got there (Drummond et al., 1997). The transition probabilities are based on available epidemiological data on the natural history of cervical cancer precursors, and include incidence, regression and progression rates of cervical abnormality, success of treatment, and mortality rates.

Assumptions

Baseline or *base* models are generated by Markov models given assumptions about a range of major factors likely to affect outcomes. Models are developed for conventional Pap testing and then compared with models of the impact of screening with a new device. In the latter, key variables expected to change with the introduction of a new device are altered, including test characteristics (screening sensitivity and specificity), and costs of screening. The models are repeated for different screening intervals, although outcomes here are reported for those where three yearly screening is applied, the cycle used in New Zealand's national cervical screening programme (NCSP).

As cohorts are followed up over a lifetime, some costs and benefits will be incurred in the future. For example, cancer cases prevented, and reduced costs of treating these cases, would not be realised for many years (especially if the cohort assumes that all women are entered at a young age). Future dollar costs and benefit streams are reduced or "discounted" by a percentage to reflect the fact that money spent or saved in the future should not weigh as heavily in programme decisions as dollars spent or saved now (Drummond et al., 1997). Discount rates used varied widely between studies, which makes them difficult to compare. Cost effectiveness is very sensitive to variations in discounting assumptions, especially concerning the discounting of life-years gained as most benefits are distant in the future compared with incurring screening costs (Payne et al., 2000). The US Public Health Service Panel on Cost-Effectiveness in Health and Medicine (Gold et al., 1996) argue that 3% would be the most appropriate real discount rate for economic evaluations. However, it suggests it is useful to also use 5%, given the large pool of studies using this rate.

It is important to note that studies from which the estimates of test characteristics are based use "conventional" screening as their comparison group. However, conventional testing in these studies may not reflect conventional practice in New Zealand. An important known variation is that New Zealand laboratories generally use more effective rescreening practices than the US mandated 10% random review. As the counterfactual (i.e. conventional screening) compared with new devices in cost effectiveness studies may be less clinically effective than normal practice in New Zealand, the potential for benefit of introducing a new device aimed at improving detection may be over-estimated for New Zealand.

Sensitivity analyses

It is recognised that there are degrees of uncertainty in many of these estimates. *Sensitivity analyses* are conducted to see whether changes to key variables across a range of values have a significant impact on outcomes. A limitation of several models is that usually only one variable is altered at a time (in univariate sensitivity analyses) which assumes that variables are independent.

Outcomes

Outcomes for models appraised usually include cost effectiveness (CE) ratios. These represent the ratio of cost estimates over effectiveness estimates such as the cost per health outcome gained; for example, cost per year of life saved, or cost per cancer prevented. In most cases these outcomes are relative to the cost effectiveness ratio of the conventional Pap test and are *additional* costs, also termed *incremental* cost-effectiveness ratios. Higher cost effectiveness ratios represent greater costs per desired outcome, and therefore are less desirable (or suggest that the intervention is less cost effective) than lower CE ratios. An alternative is *dominated* by another alternative if it is both less effective and more costly than the first alternative.

It is recognised that in order to achieve greater levels of desired outcomes, higher costs may be incurred. Some researchers in the United States suggest a threshold of US\$50,000 per life-year saved as acceptable CE ratios compared with other health interventions (Brown and Garber, 1998; Brown and Garber, 1999a; McCrory et al., 1999). Another outcome used was days-of-life saved *per woman screened*, although this is not an informative measure of effectiveness as real gains for a small minority of women may be trivialised by averaging them over the whole population of screened women.

Health system perspective

All models appraised were from a *health system perspective*, covering costs of services to the health system. Such costs include costs of screening (e.g. laboratory costs per smear, device equipment), investigation/diagnosis (e.g. clinician charges for colposcopies, and treatment). Data sources include laboratory surveys, industry quotes, insurance claim data, fee schedules, and hospital payments. These costs may differ between countries with different health systems. More desirable would be a *societal perspective* where other costs of screening and investigation of cancer precursors are included such as time off work, inconvenience, potential discomfort and psychological distress. Such costs associated with quality of life are particularly important to consider, as they are *potentially avoidable* costs when considering false positive diagnoses. These outcomes increase when test specificity decreases. The lack of information about the potential impact on quality of life of screening is a major limitation of models appraised here.

6.2 STUDY RESULTS

The main approach, findings and limitations of each economic evaluation will be summarised below. Further details of the five studies are provided in the evidence tables of **Table 7**.

AHTAC (1998)

Design and outcomes

The AHTAC review developed an original model to estimate the cost effectiveness of introducing a generic device (liquid-based screening or automated rescreening) into the Australian screening programme. The model estimated the cost per prevented case of invasive cervical cancer after a single two year screening cycle (See **Table 7, p. 82**).

Data sources

From a systematic review, the model assumed a 10% increase in LSIL abnormalities detected by the generic device and 6% increase in HSIL detected. The authors argued that their estimates of test sensitivity for the new devices were uncertain and dependent on the methodological rigour of the studies on which the test characteristics were based. Common flaws of these studies discussed included: lack of verification with a reference standard, non-routine laboratory conditions, varying thresholds for positives, and highly selected samples (e.g. of high-risk women).

Assumptions

Based on the systematic review of clinical effectiveness (reviewed in Chapters 4 and 5), the model assumed a 10% increase in LSIL detected and a 6% increase in HSIL, with progression to cancer of 1% and 12% respectively. Diagnostic and treatment costs were generally conservative. However, LSIL abnormalities were investigated by colposcopy and biopsy, which is not recommended practice in New Zealand. The impact of introducing a new device on costs for equipment maintenance, cytologist training and labour (e.g. changes to the time required to prepare and read slides, and report results) were not considered in the model.

Limitations

- A major limitation of this study was that it only considered a “snapshot” view of one screening cycle. As many abnormalities would be identified in later screening cycles, it was suggested that this cost per cancer prevented could quadruple if the device was introduced into a regular screening programme.
- The analysis considered was based on the device being applied in *addition* to conventional screening (currently in Australia, liquid-based screening is offered as a split sample adjunct to the Pap test). However, it was acknowledged that devices used as *replacements* such as direct-to-vial liquid-based screening, or automated primary screening, may provide offset savings. A limitation

of this study was that it did not incorporate these potential savings, although they were considered in sensitivity analyses.

- The study does not report discounting costs and benefits at all, which may be expected to favour the new devices.

Key results

The study estimated an additional cost of AU\$240,000 per cancer prevented from adding a new device into a single screening cycle of the national screening programme.

Sensitivity analyses

The key parameter in the cost effectiveness estimates was the device's test characteristics, which had a large influence on CE ratios in sensitivity analyses.

Conclusions

The authors suggested that increases in abnormalities detected using the new device compared to conventional screening would largely depend on the efficiency of a particular laboratory and its quality assurance procedures. Assuming total coverage in a two yearly screening of at-risk women, Australia's detection rates were regarded as potentially very high with an estimated 92.5% of squamous cell invasive cancers potentially prevented. As a consequence, the authors argued that new devices are limited in their ability to make a cost-effective difference to the screening programme.

The AHTAC study concluded that introduction of the new devices would not be cost effective from a public health perspective. To clarify issues further, an Australian study was recommended to provide more accurate estimates of the relative costs of conventional Pap screening and semi-automated and automated screening devices.

Brown and Garber (1998, 1999)

Design

The economic analysis of Brown and Garber (1998; 1999a) was funded by the Technology Evaluation Centre of the Blue Cross-Blue Shield Association. It applied a nine state transition model (1990) to calculate the costs and health effects associated with different screening modalities (including ThinPrep and AutoPap 300QC) in a cohort of same-aged women screened regularly from the age of 20 to 65 (See Table 7, p. 82).

Data sources

- The estimate of test sensitivity (Se) for the conventional Pap test was estimated as being 80%, based on NIH's consensus panel, informed by Eddy's model (1990). The cumulative Se was 81.6% taking into account 10% random manual rescreening.
- The cumulative Se for ThinPrep was 91.9%.
- Sensitivity of AutoPap300QC²⁰ rescreening using 20% selected rescreening rate from all negative (WNL) slides was 77% which, when combined with initial manual primary screening (with sensitivity of 80%) provided an estimate of cumulative Se of 95.4%²¹.
- As there is little evidence available on specificity of new devices, it was assumed to be the same as for conventional testing.

²⁰ The estimate of Autopap's Se was not based on the currently available AutoPap System which offers algorithmically assisted selection in primary screening mode combined with 15% rescreening of WNL slides, rather than 20% review alone. One may expect higher sensitivity for a device that aims to enhance both primary and rescreening.

²¹ Cumulative sensitivity of primary manual screening (*pscreening*) with Se of 80% and AutoPap rescreening (*pre-screening*) with Se of 77% was calculated as follows: $pscreening + \{(1 - pscreening) \times pre-screening\} = .80 + (.20 \times .77) = .954 = 95.4\%$ cumulative sensitivity.

- Costs and benefits were discounted at 3%, which is acceptable.

Limitations

Estimates of new devices' sensitivity were based on a systematic literature review of selected studies using a reference standard of biopsy (particularly for ThinPrep) or independent panel cytology review (particularly for AutoPap). These studies varied in the sample collection instruments used (e.g. Ayre spatula), disease prevalence, and laboratory standards, and resulted in widely varying estimates of test sensitivity. Studies of ThinPrep investigated older versions of the device (ThinPrep Beta) as no studies of ThinPrep 2000 met Brown and Garber's inclusion criteria²². The estimate of Se for AutoPap 300 QC was based on one study from a single laboratory involving a small sample of 3,487 slides initially screened as within normal limits (Colgan et al., 1995).

Estimates of cumulative sensitivity used for ThinPrep and conventional screening are likely to be inflated as they assume rescreening sensitivity to be the same as for manual primary screening (80%). However, one would expect that slides missed in primary screening would also be likely to be missed in rescreening of negatives.

Costs of screening were based in large part upon the pricing of disposable items or services used with each slide and excluded capital and training costs. Costs of investigations were also included.

Key results

- At three yearly screening, employing ThinPrep compared with conventional Pap testing in primary screening (both followed by 10% random rescreening) leads to a cost effectiveness ratio of about US\$37,000 per years of life saved (1996 costs).
- At three yearly screening, AutoPap (applied as a rescreener at a 20% review rate after conventional primary screening) dominated ThinPrep; that is, AutoPap produced greater health benefit at less cost than ThinPrep.
- At three yearly screening, the cost effectiveness ratio for AutoPap applied as a rescreener compared with Pap testing was an incremental US\$16,000 per life-year saved. This is lower than for many commonly used health interventions.
- Devices were much less cost effective when combined with more frequent screening (of two or fewer years).
- Effects of these devices on life expectancy per woman screened were modest (less than a day for AutoPap).

Sensitivity analyses

Cost effectiveness ranking was reasonably insensitive to variations in most variables using univariate sensitivity analyses. The CE ratios improved (became lower) with (i) increases in prevalence of disease, (ii) decreases in sensitivity of conventional Pap testing, and (iii) increases in the improvement in sensitivity produced by the device.

The authors conclude (Brown and Garber, 1998) that the new devices are likely to have most impact if introduced in laboratories with high false negative rates (FNR) due to high prevalence of cervical abnormalities in the screened population or poor Pap testing sensitivity. However, they suggest that evidence of high FNR should prompt an evaluation of laboratory performance rather than the introduction of a new automated device. Comparable increases in sensitivity may be achieved by directing slides to laboratories with higher baseline sensitivities than by introducing new devices.

In additional modelling subsequent to Brown and Garber's report (1999a), Adalsteinn Brown determined that when specificity is assumed to affect quality of life it takes only a small decrease in specificity to drastically reduce the cost effectiveness of screening devices (Adalsteinn Brown, June 2000, *personal communication*).

²² Subsequent analyses of results for a proportion of subjects where ThinPrep 2000 was used produced a lower estimate of sensitivity than reported in the review (Adalsteinn Brown, June 2000, *personal communication*).

Conclusions

The study's sensitivity to clinical effectiveness variations in the new devices is important given the doubt surrounding these estimates. "Our estimates (of test sensitivity) are subject to uncertainty because the literature on the effectiveness... is incomplete and sometimes contradictory"... "the highest quality studies suggest that the devices increase the TPR (true positive rate) by a modest amount, especially in a laboratory that is already highly accurate" (p. 351 Brown and Garber, 1999a).

The review concludes that the devices can be cost effective when used as part of three yearly screening, but caution that if their high cost deters participation, the devices may do little to reduce incidence and mortality of the disease.

McCrorry et al. (1999)

Design

The team from Duke University's review for AHRQ (McCrorry et al., 1999) designed a 20 state Markov model which was informed by a critical review of cost effectiveness literature (Brown and Garber, 1998; Eddy, 1990; Fahs et al., 1992). The model was constructed to determine the cost effectiveness of introducing two generic devices into a screening cohort of women aged 15-85 years (**See Table 7, p. 82**). The devices were improved primary screening (which could include liquid-based screening), and 100% automated rescreening (i.e. applied to all WNL slides after conventional primary screening).

Data sources

A meta-analysis was conducted to determine test characteristics for the Pap test and new devices.

- Estimated Pap test sensitivity of 51%, focusing on three studies performed in low prevalence populations and unaffected by verification bias.
- The cumulative Se for improved primary screening was 84%, assuming a 60% reduction in the false negative rate and 10% random manual rescreening.
- The cumulative Se for automated 100% rescreening was 80%, assuming a 60% reduction in the false negative rate of rescreening following conventional primary screening.
- The Pap test and generic new devices were assumed to have the same specificity of 97%.
- As the clinical significance of the unsatisfactory smear is widely variable, the authors concluded that the effect on the need for repeat specimens is primarily a cost issue. Therefore, the per-screen costs of the new devices took into account an expected cost saving due to an estimated reduction of unsatisfactory smears, and therefore repeat smears.

As test characteristics were uncertain, the authors undertook a threshold type analysis to determine the thresholds of sensitivity and specificity at which improved screening would produce cost effectiveness ratios of US\$50,000 per life-year or less (suggested as an acceptable CE ratio for health interventions).

Costs of diagnosis and treatment were calculated from episodes of care and included related services and complications rather than average procedure-related costs alone (as is common in other models, such as Brown and Garber's). Costs and benefits were discounted at 3%.

Limitations

Management of LSIL in the base case involved colposcopy, which is not routine practice in New Zealand. As for Brown and Garber above, the analysis did not consider the AutoPap System and may have under-estimated potential increases in sensitivity. The rescreening device considered was assumed to apply to all WNL slides, which is not routine use for the device and may have over-estimated potential increases in sensitivity.

Key results

As a result of threshold analyses, cost effectiveness ratios of US\$50,000 per life-year or less were found when the new devices were assumed to lead to reductions in false negative rates described below. At three yearly screening compared to conventional Pap primary screening followed by 10% random manual rescreening:

- *improved primary screening* (assuming a 60% reduction in the false negative rate) led to a CE ratio of US\$22,010 per additional life-year saved and a gain of 2.2 days-of-life per woman screened;
- *100% automated rescreening* (assuming a 60% reduction of false negative rate) subsequent to conventional Pap primary screening led to a CE ratio of US\$30,507 per life-year saved with a gain of 2.01 days-of-life. (Note: this was dominated by improved primary screening assuming the same reduction in FNR of 60%).

Conventional screening and device-assisted screening were not cost effective for less frequent screening (one or two years).

Sensitivity analyses

Both sensitivity and specificity were influential variables in the sensitivity analyses. For devices that improve sensitivity, cost effectiveness ratios increase rapidly (i.e. the new devices become less cost effective) as specificity decreases. This point is discussed further toward the end of this chapter.

When the sensitivity and specificity of conventional Pap testing were given similar values to those used by Brown and Garber (1998), no strategy in the model provided CE ratios below the US\$50,000 threshold. According to McCrory and colleagues, their model found lower gains in life expectancy than others using the same test characteristics. They argue that this may be related to their source of mortality estimates from other-causes, different distribution of cases within stages, the inclusion of additional health states such as HPV infection, and higher costs of diagnosis and treatment.

Conclusions

The authors conclude, “under favourable assumptions, the use of devices that improve primary screening sensitivity or rescreening sensitivity can have acceptable cost effectiveness compared with conventional Pap smear screening at a frequency of every three years” (p. 140). However, as most abnormalities found are low-grade, McCrory and colleagues suggest that there is little benefit in terms of reduced cervical cancer incidence or life-years saved.

The authors note that there is substantial uncertainty about the sensitivity and specificity of new devices. They observed that the range of values that were reported (from the few studies that used histological or colposcopic reference standards) was well within that reported for the conventional Pap test.

Further research is recommended which enables reliable estimates of test characteristics to be made. Assessment of the impact of new devices on quality of life was also highlighted as requiring intensive research attention.

Smith, Lee, Leader and Wertlake (1999)

Design

In a study funded by a grant from the manufacturers of AutoPap, Smith et al. (1999) performed the only cost effectiveness analysis, which considered the AutoPap (Primary Screening) System, which acts as a primary screener, as well as a selected rescreener (at 15% review). A seven state Markovian analysis, based on Eddy (1990), compared the cost effectiveness of the AutoPap System to conventional primary screening with 15% random rescreening, in a cohort of routinely screened women entered at age 18 (See Table 7, p. 82).

Data sources

No systematic review was performed to determine estimates of test characteristics.

- Cumulative sensitivity for the Pap test was 81.6% at a threshold of LSIL, taken from Brown and Garber's work (1999a).
- Cumulative sensitivity for AutoPap System was 91%, which is closest to the Se of 92.2% reported at a threshold of LSIL in the FDA trial (Wilbur et al., 1998) from which the estimate was said to be sourced.
- Cumulative specificity for the Pap test was 94% at a threshold of ASCUS, based on that found in the industry sponsored FDA trial of Wilbur et al. (1998), reviewed in Chapter 5.
- Cumulative specificity for AutoPap System was 95% at a threshold of ASCUS, also based on the FDA trial (1998).

However, there are concerns with the derivation of test characteristics.

- Different thresholds for sensitivity compared to specificity were used.
- It was not explained why Pap test sensitivity was based on Brown and Garber's work. It may have been more appropriate to use the trial estimates for Pap test sensitivity, as was done for estimating Pap test specificity.
- The estimates of device test accuracy may have been inflated in the trial. First, the sensitivity of the Pap test was reportedly higher in the trial (Se=85.6 at LSIL+) than sourced from Brown and Garber (81.6%) which may have reflected heightened vigilance in the trial for both arms including AutoPap. Second, as discussed in Chapter 5, test characteristics in the trial represent indirect estimates and can be expected to lead to an over-estimate of both sensitivity and specificity (Miller, 1998).

In response to similar concerns raised by Lonky (2000), Smith et al. (2000) argued:

“The absolute magnitude of missed disease for both methods {AutoPap and the Pap test} may be greater than assumed, as suggested, but it would not have an effect on the relative change in sensitivity and specificity between the two methods” (p. 84). And further, “it is one of the reasons that we used sensitivity and specificity values within a trial for comparison”.

However, estimates of test sensitivity for the Pap test were sourced from Brown and Garber in the base case. If results were inflated, they would have over-estimated AutoPap's sensitivity alone (and both tests' specificity). Marginal savings determined from individual trial site when trial derived estimates of Pap test accuracy were used may better reflect the comparative performance of these tests, as discussed below.

Limitations

AutoPap's test characteristics were determined only for those smears suitable for processing and for women who were not at high risk, as is the intended use of the device. However, this approach does not take into account the fact that manually reading the subset of unprocessed and high-risk slides using the Pap test would alter overall costs and effectiveness of screening if AutoPap were introduced into a screening programme.

Costs and benefits were discounted at 3% per annum. The base model assumed that some women diagnosed with LSIL would be investigated aggressively with colposcopy/biopsy, whilst others would be investigated with 6-monthly Pap smears, and a return to the usual screening interval after 3 negative smears. The latter is similar to that recommended in New Zealand.

Key results

- At three yearly screening, CE ratio for the AutoPap System compared to conventional Pap testing is a *reduction* of US\$975 per life-year saved, and a gain of 13.1 marginal days-of-life per woman screened. This cost reduction is in strong contrast to other economic analyses that have found *additional* costs per life-year saved for new devices compared to conventional screening.

- The base model was sensitive to variances in the cost of AutoPap testing, and disease prevalence. AutoPap's cost increased and it became less cost effective compared to conventional screening if the marginal cost of AutoPap was increased beyond US\$7, or disease prevalence was increased.

Key variables in the sensitivity analyses were test characteristics. This was illustrated when the test characteristics for AutoPap and the Pap test were varied to reflect those reported for each of the individual five trial sites that contributed to the Wilbur et al. trial (1998). In contrast to the base model, this approach used estimates of Pap test accuracy from the same trial (not from another source) which was conducted under usual laboratory conditions, and reported results at the same threshold (ASCUS). Results reveal how the cost effectiveness estimates were likely to vary widely between different laboratories, which had greatly varying test characteristics. Whilst all sites reported nominally higher sensitivity values for AutoPap compared to the Pap test, tests with greater specificity values resulted in marginal cost savings compared to the other test, as was the case in two sites²³.

Conclusions

The authors conclude, “the superior performance and marginally less costly AutoPap Primary Screening System appears to be a valuable asset ... (and) warrants serious consideration” (p. 527). However, given the reservations described about the test characteristics, as well as the anomalies of sourcing estimates and varying thresholds used, one has to be cautious about the results of this industry funded analysis of AutoPap, particularly given the potential for conflict of interest in its source of funding.

Payne, Chilcott & McGoogan (2000)

Design

Researchers from the University of Sheffield's School of Health and Related Research (ScHARR) conducted an economic analysis on behalf of the UK's National Institute of Clinical Effectiveness (NICE) (Payne et al., 2000). A state transition model was applied to estimate the cost effectiveness of introducing generic liquid-based screening (LBS) in screening a cohort of 100,000 women screened routinely from 18 to 64 years (**See Table 7, p. 82**). In contrast to other models, Payne et al. assumed only 85% screening coverage, with 15% of women in the cohort never screened.

Data sources

Unlike most economic analyses reviewed (with the exception of AHTAC's), test characteristics varied with different grades of abnormality.

- Estimated Pap test sensitivity (Se) were as follows: Se(borderline/CINI)=43%, Se(borderline/CINII)=37%, Se(moderate/CINIII)=50%, Se(moderate/invasive cancer)=40%, based on the modelling of Sherlaw-Johnson (1994).
- Marginal increases in sensitivity for LBS compared with conventional screening was estimated as 15% at CINI and CINII, and 2% for CINIII+, based on a systematic review of the literature.
- The Pap test and LBS were assumed to have the same specificity of 98%.
- The proportion of inadequate (also called, “unsatisfactory”) smears was estimated as being 3% for LBS, and 9% for conventional Pap smears (using UK data).
- Costs per LBS test took into account consumables and capital equipment. The costs of adjuvant radiation therapy, convalescence, palliative care and long term support were not included. This would lead to an underestimate of treatment costs, which would bias CE ratios in favour of conventional screening .
- In contrast to other models, costs were discounted at 6% p.a. whilst benefits (life-years) were discounted at 1.5% p.a. This is consistent with recommendations for NICE by the UK's NHS but

²³ Compared to the Pap test, in two of the five sites AutoPap led to marginal savings (US\$1,375 and \$2,719) whereas in three sites, AutoPap led to marginal costs (US\$773, \$1,717, and \$3,560).

does not reflect standard practice internationally. However, discount rates were varied in sensitivity analyses.

Limitations

Disease incidence and progression were not age dependent; as evidence suggests that incidence of the more common lower grade abnormalities is higher in younger years (due to HPV infection); this would underestimate the effects of screening generally.

As in other analyses, cost of new devices were likely to be underestimated (as training, storage and transportation, for example, were not addressed) which would therefore favour the new devices in CE analyses. The relative impact of these costs, however, is not likely to be large. However, Payne et al. argue that indirect benefits from screening with LBS such as reducing false positive results (i.e. through increasing specificity) and reducing repeat screens (i.e. through increasing slide adequacy) would balance the underestimate of direct costs such that total estimates were not likely to be underestimated. The authors also argue that specificity for the conventional Pap test may be lower than estimated which may give “room for improvement” for LBS techniques in relative specificity.

These arguments are speculative and are not supported by evidence presented in Payne et al.’s systematic review of clinical effectiveness (reviewed in Chapter 4).

- i. With respect to test characteristics, the authors concluded that “the specificity of the liquid-based method is largely unknown and may be worsened” (p. 37).
- ii. Whilst there was evidence supporting a decrease in the proportion of inadequate (unsatisfactory) specimens by LBS, the authors comment that the “literature reveals a wide and overlapping range in this proportion with both conventional smears and liquid-based methods” (p. 36). The model compared LBS’s inadequacy rates to the inadequacy rate for conventional screening in the UK of 9% (ranging from 7-11% in sensitivity analyses). This is relatively high compared to rates cited in research from other countries, though definitions of inadequate/unsatisfactory slides vary internationally, particularly with respect to the proportion of the slides that has to have squamous cells (Payne et al., 2000).

Key results

- At three yearly screening, liquid-based cytology compared with conventional Pap testing had the potential to slightly reduce incidence of invasive cancer, the proportion of women with invasive cancer, and the proportion of all deaths from cancer.
- There was also a very slight gain of 0.33 life days per woman screened.
- The introduction of three yearly LBS was predicted to reduce the lifetime number of smear tests by 6% whilst increasing the number of colposcopies undertaken by 5%. The reduction in smears arose primarily from the estimated reduction in inadequate slides and consequent reduction in the necessity for repeat screening.
- At three yearly screening, the introduction of LBS would lead to an incremental cost of £2,723 per invasive cancer avoided, and incremental cost of £2,522 per life-year gained.
- As with other reviews, lower CE ratios were predicted for less frequent (five year) screening intervals.

Sensitivity analyses

In multi-way sensitivity analyses, there was a high probability of the incremental cost effectiveness estimate of LBS compared with Pap testing being under £10,000 per life-year gained. However, estimates depended greatly on cost/benefit discounting rates, incremental costs of LBS slides, rate of inadequate smears, and marginal specificity. For example, when cost and life-years were discounted at 3% (as for McCrory et al.’s and Brown and Garber’s analyses appraised above) and incremental per-screen costs increased to UK£7 for LBS, the CE for LBS was over £25,000 per life-year gained. A small change in marginal specificity was also predicted to have a marked effect on cost effectiveness.

Conclusions

The authors found that the introduction of liquid-based screening at three year screening intervals “may be within an acceptable range of cost-effectiveness” (p. 76). However, the authors urge caution in interpreting results: “due to the small differences in health benefit and the large uncertainties in the analysis, the marginal analysis may be misleading” (p. 70).

The authors recommended an assessment of model values and assumptions to identify key areas for future research. Similar to other reviews, it was noted that improving screening uptake or using more effective specimen taking devices for conventional screening may result in liquid-based screening having a reduced impact. The authors concluded that a full cost effectiveness study based on a trial of liquid-based screening²⁴ in a low prevalence population would provide “more definitive information than is possible by modeling studies” (p. 79).

6.3 DISCUSSION

All the studies reviewed employed disease state transition models of disease natural history. The results of the models are particularly sensitive to estimates of test characteristics and costs of screening.

Costs of new devices uncertain

Results from the models were highly sensitive to changes in the cost of new devices. Generally, these costs have tended to be under-estimated, focusing mainly on consumables and capital costs only. Costs of new devices not factored into models include maintenance costs of equipment, the necessity for significant training of cytologists as well as (for LBS) smear takers, and allowance for potentially more complex storage and transportation costs. Demands on staff are important to measure because the shortage of well-trained cytologists has been one of the main factors pushing forward towards automated and semi-automated methods (Dr Gabriele Medley, Pathologist, Victorian Cytology Service, Australia, August 2000, *personal communication*). However, costs of staff time have been difficult to estimate due to inconsistent interpretations of impact on workload. For example, with liquid-based screening, a small number of studies have reported that cytologists review slides more quickly than conventional smears; however, slides have also been reported as being more tiring to review and requiring frequent breaks by staff (Payne et al., 2000). Administration time may also be increased for liquid-based slide preparation devices (Brown and Garber, 1998). Increases in costs may be balanced by decreases in capital and consumables costs if devices are taken up more widely. Unit costs may be reduced for laboratories with a larger through-put.

Uncertainty in sensitivity estimates

The key parameters used in comparing automated screening devices and conventional screening are the sensitivity and specificity of the tests. Cost effectiveness outcomes were highly sensitive to changes in test characteristics. However, these estimates were also the main source of uncertainty in the models.

As documented in our own review (chapters 4 and 5), several papers noted that estimates of test sensitivity have been limited by study designs, inadequate reference standards, and incomplete verification. A crucial aspect of estimating increases in *marginal* sensitivity of the new devices was the estimate used of the comparative sensitivity of the conventional Pap test. Such estimates are dependent on varying standards of laboratories involved and the disease prevalence of the population screened. The new devices are likely to have the most impact in laboratories with high false negative rates, although such high FNR's should be addressed in other ways than the introduction of new devices (Brown and Garber, 1998). Estimates of Pap test sensitivity were much lower in the Duke University study (McCrory et al., 1999) than the 80% reported by the analyses of other reviews (Brown and Garber, 1998; Brown and Garber, 1999a; Smith et al., 1999). Some trials of new devices using concurrent arms have found high estimates of sensitivity for both new devices and conventional testing,

²⁴ Based on this report, such a trial has been recommended by the National Institute of Clinical Excellence (NICE). The UK's NHS plans to pilot liquid-based slide preparation devices from March 2001 in sites in England and Wales (Kaminsky et al., 1997). This trial will run alongside a pilot of screening women with mild or borderline smear for human papillomavirus (Wise, 2000).

which may be due to increased vigilance for both trial arms²⁵. Using a comparable and reliable estimate of sensitivity for conventional Pap and new device testing is necessary for a meaningful cost effectiveness analysis.

A model by Raab et al. (1999) used sensitivity analyses to see what threshold of test sensitivity would be required to make new devices cost effective (this study was excluded as it considered only *annual* screening). This model assumed that for a new device to be cost effective, there must be an increased detection of HSILs, “because HSILs have a much higher probability to progress to cancer than any other SIL” (p. 261). This perspective is consistent with the approach taken in the present review, that new devices should demonstrate clinical effectiveness gains at HSIL. However, the economic models appraised in this review assumed that marginal increases in sensitivity estimates applied to *all* grades of abnormality, which is an invalid assumption²⁶. Exaggerated estimates of marginal increases for higher level abnormalities for new devices will lead to biased estimates of associated days-of-life saved overall.

As will be discussed in the next chapter, other approaches to increasing Pap test sensitivity may be more cost effective than introducing new screening devices.

Uncertainty in specificity estimates

There is little reliable evidence on specificity from the research literature, due to the lack of verification of test negatives. Because of this, the models reviewed have generally assumed that specificity is the same for conventional testing and the new device. However, this assumption is questionable until more valid studies are conducted. Whilst a new device may simultaneously raise sensitivity and specificity, this has not been conclusively demonstrated (McCrorry et al., 1999). Increases in sensitivity may come at the cost of decreased specificity.

The problem of not knowing test specificity is substantial and has profound implications for cost effectiveness because of the costs of evaluating low-grade smears that do not lead to significant cervical histology. These higher health service costs could put strains on the service and detrimentally affect women with genuine abnormalities who require investigation. There are also costs in terms of quality of life relating to increases in false positive results in terms of inconvenience, discomfort and distress arising from unnecessary investigations. When specificity is assumed to affect quality of life it takes only a small decrease in specificity to drastically reduce the cost effectiveness of screening devices (Adalsteinn Brown, June 2000, *personal communication*). This is because of the large number of women affected by a small proportional decrease in specificity (and concomitant rise in false positives) as most screen results are normal/negative.

Impact of screening on quality of life

The incremental effect of new devices on life-expectancy were universally extremely modest for women who undergo screening regularly (although costs per life saved is likely to be a more useful outcome). As most estimated additional abnormalities found are low-grade, there is little benefit in terms of reduced cervical cancer incidence or life-years saved; especially the case if there is no difference in sensitivity for high grade lesions as discussed earlier. Brown and Garber concluded, “because cervical cancer develops slowly, improvements in test sensitivity are likely to do little to change life expectancy for women who undergo screening regularly and whose smears are interpreted at laboratories with at least average sensitivity” (p. 30) (1998). It has been argued that improvements to Pap smear screening sensitivity can only be justified if a high value is placed on detecting and treating low-grade lesions (McCrorry et al., 1999). To determine this comprehensively, one would need to take into account the psychological costs of detecting and treating low-grade abnormalities and the resultant impact on quality of life.

²⁵ For example, the FDA trial of the AutoPap System, purportedly conducted under standard laboratory conditions (Wilbur et al., 1998; Wilbur et al., 1999) reported Pap test Se of 85.6% at a threshold of LSIL and Sp of 99.6% (using indirect estimates).

²⁶ For example, the FDA trial of AutoPap revealed that there was no significant difference in detection of high-grade lesions (Wilbur et al., 1999), in contrast to lower grade lesions.

Whilst quality-adjusted life-years are the preferred outcome measure, reliable estimates of the effects of new devices on quality of life have been unavailable (Brown and Garber, 1999b). However, there is evidence that positive smears have a profound impact on quality of life. In a study where women were interviewed in their homes after receiving their smear results, a positive cervical smear was found to be psychologically traumatic for a significant minority of women, with high anxiety associated with social maladjustment and negative feelings about the self (Bell et al., 1995). Anxiety may be related to misunderstanding of the meaning of smear results and the necessity for further investigations. In a retrospective study in the UK, of women told that their smear was mildly abnormal, 47% of those who were to receive an immediate-colposcopy (n=182), and 33% receiving a repeat Pap test (n=163), thought they had cancer (Jones et al., 1996). In addition to anxiety about cancer, relationship problems relating to having an HPV infection are also evident for women with mildly positive smears (Adalsteinn Brown, June 2000, *personal communication*).

In addition to psychological distress, absence from work, inconvenience, and discomfort may be associated with screening, diagnosis, and clinical management of abnormalities. As most abnormalities rarely progress to cancer, it has been argued that taken overall these outcomes “might well outweigh the negative impact of cancer itself on women’s quality of life at the population level” in analyses of new screening devices (p. 145) (McCrorry et al., 1999). The impact is most profound for false positives where investigations could be avoided. However, there could also be positive impacts on quality of life by new devices if they reduced the necessity for re-screening inadequate/unsatisfactory smears, as may be the case for liquid-based slide preparation devices (Sawaya and Grimes, 1999), though data is not conclusive. It could also be argued that an increase in psychological costs for (many) women receiving false positive results is an acceptable price to pay for saving one extra life by an increase in true positives. However, at present there is no good basis for evaluating the societal cost of a false positive and this is a major limitation on cost effectiveness models to date.

Potential negative impacts on quality of life need also to be balanced with cost savings from avoiding possible litigation for failure to diagnose abnormalities. Such medico-legal concerns have not been considered by any model to date, though may be challenged by statements from professional societies that new devices are not the “standard of care” (American College of Obstetricians & Gynecologists Committee on Gynecologic Practice, 1998).

Conclusions

Further research is required which generates valid estimates of test characteristics to be made. Prospective clinical trials comparing devices need to include careful evaluations of the respective effects of the Pap test and new devices on quality of life (Brown and Garber, 1999b). Further discussion of the implications of the above for New Zealand’s national cervical screening programme are included in the next chapter.

Table 7. Evidence table for economic evaluations

Source	Design and outcomes	Data sources	Assumptions	Limitations	Key results	Sensitivity of model	Reported conclusions
<p>Australian Health Technology Advisory Committee (1998)</p> <p>Australia</p>	<p><i>Design:</i> An original model was used to determine the additional cost and cost effectiveness of augmenting the Australian screening programme with a new generic device (either liquid-based screening or automated rescreeing) in a single 2 year screening cycle of women aged 20-69 years.</p> <p>Distribution/ progression of health states was based on data from the Victorian Cytology Service, AHTAC Working party experience, Östör (1993), and flow-charts of Braggett et al., (1993) .</p> <p><i>Outcomes:</i> Cost per potential cancer prevented (i.e. those abnormalities that would progress to cancer if not treated)</p>	<p><i>Test characteristics:</i> systematic review of clinical effectiveness used to estimate the number of additional potential cancer cases attributed to new device. Estimated 7.5% increase in positives detected by new device where 90% are LSIL (10% increase in LSIL and 6% increase in HSIL detected).</p> <p>Screening, diagnosis and treatment costs taken from current average price information (no discounting) from routinely available Australian statistics.</p> <p>Additional test costs of the new devices of AU\$25.</p>	<p><i>Model assumed:</i> (i) 2 yearly cycle; (ii) device does not replace present practice (therefore with no potential offset savings); (iii) only one device introduced.</p> <p>Estimate of increase in positives is limited (non-routine conditions, lack of reference standard, varying thresholds, highly selected samples).</p> <p>Management of LSIL involves colposcopy and biopsy (not routine in NZ).</p> <p>No discounting of costs or benefits.</p>	<p>No costs of maintenance of hardware; impact on labour/training; educating practitioners and women. Savings of reduced inadequate smears in liquid-based screening not costed.</p> <p><i>Limitations include:</i> (i) one cycle, many abnormalities would be picked up at subsequent screening cycles, (ii) costs of investigating abnormalities based on direct service fee, (iii) abnormalities may be treated earlier or later due to increased Se, (iv) new devices may have an impact on screening uptake, (v) new devices may supplement or replace conventional screening methods (vi) costs may be reduced as devices are diffused.</p>	<p>Expect additional 300 potential cancer cases detected with an estimated additional AU\$240,000 cost per cancer prevented, <i>per screening cycle</i> (however, many abnormalities would be detected and treated in subsequent screening).</p> <p>Applying the model to multiple screening cycles, expect a four-fold increase in additional cost per cancer prevented (AU\$960,000) based on Schechter (1996).</p>	<p>Cost effectiveness estimates presented for variations in four factors. (i) <i>clinical effectiveness</i> (percentage increase in total positives: 3.5% - 30%, which is dependent on the efficiency of practice in each laboratory), (ii) grade of detected positives (90, 95, or 100% LSIL), (iii) <i>unit price</i>, as it may decrease if used routinely due to economies of scale (AU\$10-30) (iv) <i>potential offset savings</i> if the new device replaced some of present screening practice (AU\$7.5-15 per smear).</p> <p>The key parameter is the device's increased accuracy. Costs of additional treatment were small compared with costs of screening with the new device.</p>	<p>"A significant reduction in present costs would be required before the devices could be considered to be cost effective from a population health perspective" (p.45).</p> <p>Suggest that an Australian study which compares costs and relative effectiveness of the new devices and of regular repeat Pap smears is needed.</p> <p>Whilst costs do not justify routine use of devices in the screening programme, laboratories may introduce them for other reasons and uptake would be a matter of individual choice. Devices used as replacements or for primary screening could provide offset savings.</p>

Table 7. Evidence table for economic evaluations (continued)

Source	Design and outcomes	Data sources	Assumptions	Limitations	Key results	Sensitivity of model	Reported conclusions
Brown and Garber (1998) Brown and Garber (1999a) USA	<p><u>Design:</u> 9 state time-varying transition health state model, based on Eddy (1990), was used to determine the cost effectiveness of introducing ThinPrep, AutoPap 300 QC (and Papnet) into a cohort of women screened from 20 – 65 years, representative of US population. Results were reported for screening intervals of 1, 2, 3 and 4 years.</p> <p><u>Outcomes:</u> - cost per incremental year of life saved - life-years gained - cost per woman screened</p>	<p><u>Test characteristics:</u> Systematic review - estimates at LSIL+. <u>Pap test:</u> TPR for primary screening (80%) plus 10% random rescreening (Cum Se=81.6%). <u>ThinPrep:</u> Increase TPR by 11%, plus 10% random rescreening (Cum Se=91.9%). Studies include earlier ThinPrep models. <u>AutoPap:</u> Primary screening TPR of 80% plus increase of 15.4% for detection of 20% WNL slides (Cum Se=95.4%; estimate from one study with a small sample). NB: Assumed no change to Sp between tests with Sp=95.8% (95.4% for AutoPap at 20% review). <u>Marginal costs per slide:</u> from articles, industry, financial reports, a survey of Californian laboratories. TP: US\$9.75, AP = US\$5. <u>Costs of screening, diagnosis, treatment:</u> Medicare data</p>	<p>Model includes factors disease incidence and progression (age dependent), regression of pre-invasive lesions, success of treatment after diagnosis (stage dependent), all cause mortality, costs and test characteristics (see data sources).</p> <p>Women begin screening at the same age.</p> <p>Se for rescreening is the same as for primary screening. Se and Sp are not differentiated between the higher lesion grades.</p> <p>AutoPap 300QC selected 20% slides for QC review.</p> <p>Screening time the same</p> <p>Updated to 1996 US\$ using the CPI.</p> <p>All costs and benefits discounted at 3% per annum.</p>	<p>Only considers AutoPap as a QC rescreener, not for primary screening.</p> <p>Caveats concerning source of Se estimates include that some studies report Se of conventional screening that is unusually low.</p> <p>Costs of devices excluded training and capital costs, arguing that these account for less than US\$0.25 for each device when equipment is used at full capacity.</p> <p>New practices may generate lower costs for treatment of lesions. Societal costs of convalescence and time off work are not included.</p> <p>Mortality has declined since Eddy's model leading to an overestimate of the health effects of the devices.</p>	<p>ThinPrep produced less health benefit at greater cost (was dominated by) strategies employing AutoPap 300QC 20% selected rescreening.</p> <p>At 3 yearly screening, CE ratio for AutoPap than conventional Pap testing is about US\$16,000 per life-year saved. This represents more life-years at lower cost compared with conventional Pap testing at 2 yearly screening. For ThinPrep, cost per life-year saved would be US\$37,000 with triennial screening compared with conventional Pap testing (but was dominated by AutoPap).</p> <p>For triennial screening, AutoPap increases life expectancy by less than one day (.96) compared to conventional screening, at an incremental cost of US\$43 more per woman screened.</p>	<p>Univariate sensitivity analyses performed.</p> <p>AutoPap dominates ThinPrep regardless of a wide range of variations in assumptions.</p> <p>The cost effectiveness (CE) ratios improved (became lower) with increases in prevalence of disease, decreases in sensitivity of conventional Pap testing, and increases in the improvement in sensitivity produced by the device.</p> <p>Of little impact on relative CE ratios were changes to the discount rate, costs of conventional Pap testing, of detection and treatment.</p>	<p>"Our estimates are subject to uncertainty because the literature on the effectiveness... is incomplete and sometimes contradictory"... "the highest quality studies suggest that the devices increase the TPR by a modest amount, especially in a laboratory that is already highly accurate" (p.351).</p> <p>"These findings may change as new evidence becomes available on the TPR of the devices, especially if they differ in their ability to classify the stage of abnormality correctly"... "The major barrier to prevention of cervical cancer is not the accuracy of the Pap test, but the failure to be screened at all...if (the devices') high value deters participation in... programs they will not reduce the toll of disease" (p. 352).</p>

Table 7. Evidence table for economic evaluations (continued)

Source	Design and outcomes	Data sources	Assumptions	Limitations	Key results	Sensitivity of model	Reported conclusions
McCrory et al., (1999) USA	<p><u>Design:</u> 20 State Markov model (referring to various papers, including (Eddy, 1990; Fahs et al., 1992)) constructed to determine the cost effectiveness of introducing a generic device in two ways to improve Se (with no decrement in Sp) for (i) improved primary screening (such as LBS), and (ii) automated rescreening of all WNL slides, into a cohort of women aged 15 - 85 years screened routinely.</p> <p>Results were reported for screening intervals of 1, 2 and 3 years.</p> <p>A threshold type analysis determined the thresholds of Se and Sp at which improved screening would produce CE ratios of US\$50,000 per life-year or less.</p> <p><u>Outcomes:</u> - cost per incremental life-year saved - cost per cancer case/death prevented</p>	<p><u>Test characteristics:</u> Systematic review of clinical and economic evidence. <i>Pap test</i> Se (51%) and Sp (98%) estimated from 3 studies in low prevalence populations unaffected by verification bias.</p> <p>From threshold analyses: <u>Improved primary screening:</u> Reduction of FNR of 60%, plus 10% random rescreening at same Se (Cum Se=84%). <u>Automated rescreening:</u> Pap test primary screening plus 100% rescreening with a reduction of FNR of 60% (Cum Se=80%). Sp was 97% for all tests.</p> <p><u>Costs of tests:</u> Incremental cost at US\$10 per test, which reflects savings from fewer smears.</p> <p><u>Costs of screening, diagnosis, and treatment:</u> health insurance claims, fee schedules, and payments for hospital admissions.</p>	<p>Model includes HPV infection and regression, disease incidence and progression (age dependent), regression of pre-invasive lesions, success of treatment (stage dependent), all cause mortality, costs and test characteristics. Various assumptions (p. 46) favor more sensitive devices. Treatment for LSIL+ receive colposcopy, also modelled for colposcopy at ASCUS+.</p> <p>Women had no SIL or HPV infection at age 15.</p> <p>Se and Sp are not differentiated between the higher lesion grades. Specificity did not vary between tests</p> <p>Costs adjusted to 1997 US\$. Costs and benefits discounted 3% pa.</p>	<p>Only considered AutoPap 300QC (not AutoPap System), and assumed automated screening used for 100% of negatives which is not routine use.</p> <p>There is substantial uncertainty about the estimates of sensitivity and specificity of thin-layer cytology and automated rescreening devices. The uncertainty is not reflected in the point estimates for effectiveness and cost effectiveness.</p> <p>Societal costs of convalescence and time off work are not included.</p> <p>Did not consider impact of patient and provider behaviour (such as irregular or no screening, lack of follow-up, inappropriate diagnosis/treatment practices).</p> <p>Limited data on impact of screening and treatment on quality of life.</p>	<p>Conventional screening and device-assisted screening were not cost effective for less frequent screening (one or two years).</p> <p>At 3 yearly screening, compared to Pap primary screening followed by 10% random manual rescreening:</p> <p>- <i>Improved primary screening</i> (assuming 60% reduction in the FNR) led to a CE ratio of US\$22,010 per additional life-year saved and a gain of 2.2 days-of-life per woman screened;</p> <p>- <i>100% automated rescreening</i> (assuming a 60% reduction of FNR) subsequent to Pap primary screening led to a CE ratio of US\$30,507 per life-year saved with a gain of 2.01 days-of-life. (NB: dominated by improved primary screening at same assumed reduction of FNR of 60%)</p>	<p>Extensive univariate sensitivity analyses performed.</p> <p>Model was relatively insensitive to assumptions about cervical cancer incidence, the cost of devices, diagnostic strategies for abnormal screening results, age at onset of screening, or most other variables tested.</p> <p>In the absence of reliable estimates of test characteristics, study presented CE under a wide range of assumptions. "It is clear from our sensitivity analyses that both sensitivity and specificity are important in determining cost-effectiveness" (p. 140).</p> <p>The model was validated against epidemiological data and previously published models of cervical screening.</p>	<p>"there is substantial uncertainty about the estimates of sensitivity and specificity of the new technologies" (p. 140).</p> <p>"Although it is clear that both thin-layer cytology and automated rescreening devices provide an improvement in effectiveness at higher cost, the imprecision in estimates of effectiveness makes drawing conclusions about the relative cost-effectiveness .. problematic" (p. 140).</p> <p><u>Need future research:</u> (i) into the specificity of new devices (ii) using histological reference standards (iii) into the effect of pre-malignant and malignant cancers, and treatments on quality of life for comprehensive assessment of impact of devices for screening.</p>

Table 7. Evidence table for economic evaluations (continued)

Source	Design and outcomes	Data sources	Assumptions	Limitations	Key results	Sensitivity of model	Reported conclusions
Smith et al., (1999) USA	<p><u>Design:</u> Conducted a 7 state Markovian analysis (based on Eddy (1990) and other published sources) of the marginal cost effectiveness of the AutoPap System for primary screening (and selected 15% rescreening) as compared to manual cervical screening and 10% random rescreening. This was performed in a hypothetical cohort of 100,000 women entered at age 18 (followed until death), screened routinely. Results were reported for screening intervals of 1, 2, 3 and 4 years.</p> <p><u>Outcomes:</u> Marginal number of years saved by using conventional Pap test or AutoPap.</p>	<p><u>Test characteristics:</u> No systematic review performed. <u>Pap test:</u> Cum Se=81.6% at LSIL+ from Brown and Garber, above (Brown and Garber, 1999a). Sp=94% at ASCUS+ from the FDA clinical trial of Wilbur et al. (Wilbur et al., 1998).</p> <p><u>AutoPap:</u> From the FDA trial (Wilbur et al., 1998): Se=91% (closest to the trial's Se at LSIL of 92.2%); Sp=95% at ASCUS+ (Wilbur et al., 1998).</p> <p><u>Costs of tests:</u> US\$20 per Pap smear (mid-point of range used). Cost of US\$4.50 per AutoPap (from industry).</p> <p><u>Costs of screening, diagnosis, and treatment:</u> US Congress, Office of Technology Assessment background paper on costs of screening elderly women, and health insurance claim data.</p>	<p><u>Model includes:</u> disease incidence and progression (age dependent), regression of pre-invasive lesions, success of treatment after diagnosis (stage dependent), all cause mortality, costs and test characteristics. Comparative model rather than a comprehensive model. Mid-point of value ranges used.</p> <p>LSIL investigated by colposcopy/biopsy, or 6 monthly Pap smears (return to programme after 3 negative smears). Follow-up after cancer was annual until death.</p> <p>Se and Sp are not differentiated between the higher lesion grades.</p> <p>Costs adjusted to 1997 US\$. Costs and benefits discounted 3% pa.</p>	<p>AutoPap only used for women who are not at high risk.</p> <p>Estimates use inconsistent thresholds, and varying sources.</p> <p>Does not consider women not routinely screened, and does not comprehensively assess the impact of the tests on workload, quality of life, or potential litigation avoided.</p> <p>Only considers effect of AutoPap, not liquid-based screening or other devices.</p> <p>Probably under-estimates the true costs of screening.</p>	<p>AutoPap dominated manual screening; i.e. is less costly and more effective than manual screening alone (in the long run).</p> <p>Advantage comes from (i) reducing number of slides requiring manual review, and (ii) increasing proportion of cervical abnormalities detected at the preinvasive stage (reducing costs of treatment).</p> <p>Three year screening resulted in a reduction in cost for AutoPap compared with PS of US\$35 per person screened yielding an additional 13.1 days-of-life saved. This is equivalent to a reduction in a cost per life-year saved of US\$975.</p>	<p>Univariate sensitivity analyses performed.</p> <p>The majority of variables did not heavily affect the cumulative outcome of the analysis.</p> <p>The model was sensitive to: AutoPap cost of US\$27 (i.e. marginal cost of US\$7 instead of US\$4.50). Test Se (increase associated with cost increase). Test Sp (increase associated with cost decrease). Disease Prevalence (as prevalence increase, AutoPap becomes more expensive than the Pap test)</p> <p>Used Wilbur et al. trial data from 5 sites. These revealed varying Se and Sp estimates for Pap test and AutoPap. In 3 of the 5 sites, AutoPap led to higher costs and in 2 cost savings compared to the Pap test.</p>	<p>"In terms of seeking to meet the goal of eliminating cervical cancer while controlling the costs of screening, the superior performance and marginally less costly AutoPap Primary Screening System appears to be a valuable asset. This new device warrants serious consideration because of its potential to advance the health and well-being of women" (p. 527).</p> <p><u>Note:</u> Funded by a research grant from the manufacturers of AutoPap.</p>

Table 7. Evidence table for economic evaluations (continued)

Source	Design and outcomes	Data sources	Assumptions	Limitations	Key results	Sensitivity of model	Reported conclusions
	(ii) Resource: number of smear tests, colposcopies undertaken (iii) Economic: Cost per life-year saved, cost per invasive cancer avoided	<u>Costs of tests:</u> Marginal cost of £3.60 per LBS smear which takes into account consumables and capital equipment, assuming labour costs unchanged. <u>Direct costs of screening, diagnosis, and treatment:</u> UK NCSP costs (1994).	Inadequate slides immediately rescreened. Treatment 100% accurate. Se varied between lesion grades. Costs adjusted to 1999£. Costs discounted at 6% p.a.; life-years at 1.5% p.a.	Insufficient quality of life information was available to estimate cost per quality adjusted life-year saved. Did not quantify reduced anxiety and travel costs associated with fewer repeat screens due to improved slide adequacy, or litigation costs (which will still occur for LBS false negatives)	(vi) cost (incrementally) £2,723 per invasive cancer avoided, and cost (incrementally) of £2,522 per life-year gained (with high probability of it being under £10k in multiway sensitivity analyses). Lower CE ratios for longer screening intervals.	The model was validated against UK incidence data and other CE models.	A full CE study based on a trial of LBS's introduction in a low prevalence population would provide "more definitive information than is possible by modelling studies" (p. 79). Recommended an assessment of model values and assumptions to identify key areas for future research.

NOTE: Costs are reported in their original currency. Conversion from one Australian, United States, and English currencies to New Zealand dollars is approximately as follows: 1 AUD = 1.3 NZD; 1 USD = 2.5 NZD; 1 GBP = 3.5 NZD (as at 22 September 2000)

Key:

PS: Conventional Papanicolaou Smear Test

AP: AutoPap

ASCUS: atypical squamous cells of undetermined significance

Se: Sensitivity

CE: Clinical effectiveness

TP: ThinPrep test

LBS: Liquid-based screening

LSIL: low-grade squamous intraepithelial lesion

Sp: Specificity

WNL: Within Normal Limits

ACP: AutoCyte Prep Test

CIN: cervical intraepithelial neoplasia

HSIL: high-grade squamous intraepithelial lesion

Cum Se: Cumulative Sensitivity (after primary screening and rescreening)

FNR: False Negative Rate

Chapter 7: Discussion

7.1 SUMMARY OF EVIDENCE

This report systematically reviewed the international evidence for clinical effectiveness and cost effectiveness of replacing conventional screening with automated and semi-automated devices in New Zealand's population-based screening programme, updating the Australian review by AHTAC of articles published to July 1997 (Australian Health Technology Advisory Committee, 1998).

Over 700 articles were identified by the search strategy and supplementary search updates. From reading abstracts/titles, the reviewer (MB) identified 58 articles as potentially eligible for inclusion. These were retrieved as full text and the inclusion and exclusion criteria applied to identify a final group of 26 papers. These included 20 papers reporting primary research: 15 on clinical effectiveness of new devices (10 ThinPrep, three AutoCyte Prep, two AutoPap) and five on cost effectiveness; and six additional systematic reviews/meta analyses²⁷.

Clinical effectiveness of new devices

Many studies were excluded because of design limitations, mainly relating to a lack of verification by either histology or cytology by adjudicated panel cytology review.

Liquid-based slide preparation devices

Six systematic reviews and/or meta-analyses were appraised in addition to the AHTAC review. Thirteen studies reporting original data were appraised, which compared screening involving liquid-based slide preparation devices with screening by conventional Pap test; 10 evaluated ThinPrep and three concerned AutoCyte Prep. All studies used histology (biopsy) as their reference standard for verification of cytological diagnoses. All but one failed to verify negative test results adequately, with verification of positives commonly limited to test results which were discordant (i.e. one screening test gave a positive result and the other gave a negative result). This results in a lack of evidence on specificity, and in the latter cases, inflated estimates of sensitivity. The only study that did verify negatives did so for a small fraction of negatives at a threshold of LSIL which were likely to have exhibited borderline abnormalities in order to have required histological follow-up, thus specificity may have been underestimated.

Inadequate verification means that the sensitivity and specificity of LBS cannot be reliably determined. In the absence of direct estimates of test characteristics, yield of abnormalities can be compared to provide indirect (though inflated) estimates of sensitivity and specificity (or more correctly, relative true positive rates and relative false positive rates). Such outcomes are only useful when prevalence of disease can be assumed to be the same in women screened by the device and the Pap test.

Six studies allowed within-subjects comparisons using split-sample designs where one is assured of equivalent prevalence of disease between samples compared. A Taiwanese study (Wang et al., 1999) reported higher yield of abnormalities at a threshold of HSIL by ThinPrep compared with the Pap test, however only a small sample was considered. A Costa Rica study (Hutchinson et al., 1999), reporting indirect estimates, suggested equivalent sensitivity and slightly higher specificity for ThinPrep compared to conventional slide preparation in detection of abnormalities at HSIL or higher. However, the yield of ASCUS was over four times higher for ThinPrep slides compared to conventional Pap tests. Both of these studies involved samples of high-risk women which may not be generalisable to a population based screening programme. Two split sample studies employed case control designs nested within case series, though numbers were small in both. One study found similar degrees of prediction of cancer for ThinPrep and the Pap test (Inhorn et al., 1998), and another investigating glandular lesions reported greater prediction of adenoma in situ by the Pap test (Roberts et al., 1999).

²⁷ Some systematic reviews were included in the count of economic studies that included review components.

Two studies of AutoCyte Prep (Bishop et al., 1998; Minge et al., 2000) revealed higher detection of smears read as at least low-grade abnormality by LBS compared with conventional slide preparation, but equivalent rates of detection for smears read as at least high-grade abnormality.

The other seven studies compared the liquid-based screening device with the Pap test collected from different cohorts of women, two collected concurrently and five employed an historical control for the Pap test. These samples are open to many biases that make it possible that the prevalence of disease varies in the groups compared, making comparison of detection rates meaningless.

Given these and other limitations in study quality, the clinical effectiveness of ThinPrep and AutoCyte Prep for detection of high-grade abnormalities cannot be reliably determined from the current evidence base. Moreover, it is not possible to say whether one device has advantages over another in terms of considered outcomes. Valid estimates of test sensitivity and specificity of these devices await further research employing better designs. Given the paucity of high quality research, the latest International Academy of Cytology (IAC) Task Force (No 3) on sampling issues concluded, “conventional Pap test should remain the international standard of care for the diagnosis of cervical cancer precursors in cervical cancer screening programmes” (draft May 2000, Dr Ulrik Baandrup, *personal communication*).

Semi-automated devices for primary screening and re-screening

In addition to the AHTAC review, six systematic reviews and meta-analyses were appraised that considered AutoPap. Only one study reporting original data was eligible for appraisal; it evaluated the AutoPap System device as a combined primary screener and rescreener. This prospective trial involved limited verification of cytological diagnoses and therefore did not permit direct estimates of test sensitivity and specificity. Using indirect estimates verified by panel cytology review of discordant test results (leading to an overestimation of diagnostic performance), there was no significant difference demonstrated in detection of abnormalities that were at least of high grade between the AutoPap System and conventional screening (the Pap test followed by 10% random rescreening. Histological follow-up was only reported for a minority of cytological diagnoses for AutoPap only, allowing no comparison with conventional screening. There was inadequate evidence concerning the specificity of AutoPap.

Whilst there may potentially be increases in detection of low grade abnormalities for AutoPap compared with conventional screening with 10% random rescreening, there is no evidence to suggest an increase in detection of high grade abnormalities which are of clinical importance in a population based screening programme.

Conclusions

There were no randomised controlled trials using an outcome of invasive cancer incidence or mortality. Whilst there was some evidence that new devices may marginally increase detection of low grade abnormalities, estimates of test sensitivity and specificity could not be reliably determined from the current evidence base. Studies were severely limited by design, inadequate reference standards, and incomplete verification of cytological diagnoses. There was no reliable evidence for improved detection of high-grade abnormalities by semi-automated and automated devices for cervical screening.

Recently, empirical evidence has shown that methodological shortcomings in studies may overestimate the accuracy of a diagnostic test being evaluated (Gold et al., 1996). Research studies applying appropriate reference standards for verification of cytological diagnoses are required. To be relevant to New Zealand’s conventional screening (where 10% random rescreening is not used), comparisons with alternative rescreening strategies to 10% random review are also recommended.

Cost effectiveness of new devices

There were no randomised-controlled trials or field studies identified assessing the economic impact of introducing new cervical screening devices compared with the conventional Pap test. Instead, all the studies reviewed employed disease state transition models.

Cost effectiveness outcomes were highly sensitive to changes in test characteristics of sensitivity and specificity. As found in the present review and those of other researchers, these estimates were also the main source of uncertainty in the models. In the absence of reliable data, models assumed that there was increased sensitivity of new devices compared to conventional screening, and no difference in specificity between tests. Under these conditions, the incremental effect of new devices on life expectancy was extremely modest for women who are screened regularly. Increases in sensitivity may come at the cost of decreased specificity that would lead to increases in false positive results (where slides are read as abnormal when they are in truth, normal).

Conclusions

Cost effectiveness models were severely limited by the uncertainty surrounding estimates for improved sensitivity and the lack of information on changes to specificity that may occur with the introduction of new devices into screening programmes. When improved detection at all grades of abnormality was assumed, the impact of new devices on days-of-life saved was extremely small for women screened at three yearly intervals. As most (assumed) additional abnormalities found are low-grade, these are likely to regress, or if they persist are very likely to be detected at the next regular screen. Given the very slow growth of cervical cancer from pre-cancerous abnormalities which progress, women screened regularly at laboratories meeting minimal quality standards will, in the vast majority of circumstances, have any abnormalities missed at one screen detected at a subsequent screen and potentially treated before cancer develops.

The possibility that new devices may decrease specificity and therefore increase false positive diagnoses has not been comprehensively evaluated in cost effectiveness models. False positives are likely to have a significant impact on both health service costs, as well as quality of life in terms of a screened woman's inconvenience, discomfort and distress, arising from unnecessary investigations. As will be discussed in Section 7.3 below, this would drastically reduce the cost effectiveness of screening devices that reduce specificity.

Further research is required which generates valid estimates of test sensitivity and specificity. Comparing the cost-effectiveness of conventional screening versus screening with new devices should ideally be from a societal perspective. This approach will require careful investigation of the impact of screening and consequential investigation and treatment of detected abnormalities, on quality of life.

7.2 FUTURE RESEARCH

Further research should address limitations in study design demonstrated in this review. Key features are described below, followed by a brief list of emerging areas of research.

Clinical effectiveness research

- Large-scale prospective multi-centre trials performed under normal laboratory conditions and involving low risk populations are required.
- Discrepancy studies, where only discordant tests (one reporting positive, the other negative, diagnoses) are verified, should be avoided.
- Debate continues surrounding the issue of verification of negatives. Ideally, all negatives would be verified in the determination of sensitivity and specificity. However, there are ethical and practical concerns associated with colposcopic/histological follow-up of negatives in a controlled trial. It has been argued that a random fraction of cytology negative women referred for colposcopy is more feasible and would permit statistical correction from work-up bias and estimation of test specificity (McCrory et al., 1999). However, it could be argued that what is unethical for all negatives may also be unethical for a random sub-sample.

Given the ethical and logistic problems with these approaches, an acceptable, though arguably inferior (McCrory et al., 1999), surrogate reference standard may be independent expert panel cytology review of all discrepant, and concordant positive slides, and all or a random fraction of concordant negative slides (see Section 2.6). The panel should include three experienced cytologists without commercial interests in the devices evaluated who were not associated with the

original slide readings. Panel members should be blind to (unaware of) the diagnoses assigned to slides reviewed and, where possible, blind to the test being used (e.g. Pap test, AutoPap). Consensus should be sought on diagnosis for final verification of truth. Where this cannot be reached, an additional cytologist should be brought in and consensus sought at a multi-headed microscope.

- When using expert panel cytology review as the reference standard for verification of diagnoses, a majority (preferably all) HSIL+ slides should be followed up by colposcopy/biopsy and this diagnosis should be used as final truth determination (i.e. verification) for these slides (McCrory et al., 1999).
- It should be recognised that both reference standards (histology and cytological panel review) have their own inherent biases.
- Automated methods for rescreening should be compared with a range of different methods of manual rescreening, such as rapid rescreening and targeted rescreening more commonly used in New Zealand. By contrast, most studies have compared automated rescreening with Clinical Laboratory Improvement Amendments (CLIA) mandated 10% random rescreening used in the USA which is known to be broadly ineffective in reducing false negatives (McCrory et al., 1999; Wain, 1997).

Cost effectiveness research

- Cost effectiveness modeling studies will be more meaningful when we have valid estimates of test characteristics, and comprehensive measurement of costs of screening.
- According to the Methods of Cost/Benefit Evaluation Task Force of the International Academy of Cytology (IAC) (1997), a societal perspective should be utilized in cost effectiveness analyses and presented as marginal benefits and costs (Solomon et al., 1998). This perspective should take into account the impact of screening on the quality of life of women screened, which is an area in great need of careful evaluation (Brown and Garber, 1999b).

Future developments

There are many developments in cervical screening technologies beyond the devices considered in this review. Excluded technologies identified from our search strategy that may emerge in the future are described in Section 3.1. In addition, other developments on the horizon are highlighted below. It is recommended that the conclusions of this report be reconsidered in light of these developments in 12 months time.

- TriPath is currently seeking a Supplement to the AutoPap (Primary Screening) System's FDA pre-market approval to allow it to read liquid-based slides prepared by AutoCyte Prep (Supplement application submitted October 6, 1999). TriPath is also developing a new automated screening system (currently known as "AutoCyte SCREEN 2") which will incorporate features of AutoPap, AutoCyte SCREEN and Papnet. This is expected to be released in about two years (Fiona Diversi, Cytology Specialist for Dade Behring, Australasian distributors for TriPath Imaging, May 2000, *personal communication*).
- Cytoc Corporation is developing a computer-assisted screener, which may be used with liquid-based slide preparation devices. It will be submitted for FDA PMA later this year with judgement expected in 2001 (Dr Bill Mackey, Medical Director of Biotek, New Zealand's sole distributors of ThinPrep devices, May 2000, *personal communication*).
- Human Papilloma Virus (HPV) DNA testing identifies whether women carry "high risk" types of HPV that are associated with high grade cervical abnormalities and cancer. This can allow triaging of women so as to distinguish women who have no detectable HPV or only "low risk types" who require diminished surveillance (McMeekin et al., 1997). The UK is currently piloting the use of HPV testing (using Digene Corporation's "Hybrid Capture" device) as part of the NHS's Cervical Screening Programme. The pilot will run for one year beginning in the (Northern Hemisphere) summer of 2000 at three mainland sites in England and Wales. The pilot study will test women for HPV if they have a mild or borderline Pap smear result (i.e. LSIL or ASCUS). If these women test positive for high risk HPV types, they will be referred immediately for further treatment. Those who are HPV negative will be offered a second HPV test along with a repeat Pap

smear after six months. If the second HPV test is negative and the cervical abnormality has not progressed to high-grade disease, the woman would be returned to the normal screening cycle. The recommendations for this pilot were made following publication of a health technology assessment report of September 1999 produced by Prof. Jack Cuzick's team based at the Imperial Cancer Research Fund in London (Cuzick et al., 1999). Results of the trial will not be available before August 2001.

- Dr Tony Hanselaar, of the University Medical Centre Nijmegen, The Netherlands, and his team are in the process of performing a systematic review of LBS, automated screening and HPV-testing to inform decision making for their national population screening program. Findings will be submitted for plenary consensus discussion with involved parties in October 2000, with results published subsequently in Dutch, and then translated into English.
- A group at the UK's ARIF (Aggressive Research Intelligence Facility) led by Dr Chris Hyde hold a NHS Health Technology Assessment grant to conduct a systematic review and undertake economic modelling relating to the same devices as reviewed in the present report (*personal communication*, July 2000). The project commenced in August 2000 and results are anticipated by October 2001.

7.3 IMPLICATIONS OF RESULTS FOR THE NATIONAL CERVICAL SCREENING PROGRAMME

In this section, the effectiveness of conventional screening is discussed, followed by a discussion of the key implications of the evidence reviewed in this report for a population-based cervical screening programme. Finally, alternative interventions to the introduction of new devices are highlighted for a national cervical screening programme.

Conventional screening can be highly effective

The Pap smear is a relatively simple, easy to perform, well tolerated, and inexpensive screening test. It has been the most successful cancer screening technique of the 20th century, making cervical cancer largely a disease of women who have not been regularly screened (Solomon et al., 1998). In countries where cervical screening programmes have been established, the Pap test has led to a decrease in cervical cancer mortality and morbidity (Free et al., 1991; Mitchell and Giles, 1996). There have also been significant decreases in these rates in New Zealand (Members of the Working Party on Cervical Screening, New Zealand, 1998; New Zealand Health Information Service, 1999). These decreases are particularly impressive given that large increases would have been expected to occur over this period in the absence of screening due to changes in the demographic profile (and other factors) of New Zealand²⁸ (Cox and Skegg, 1992).

Whilst regular screening is potentially extremely effective, errors in single screens can and do occur, as with any screening test. Whilst false negatives are inevitable, they translate as an extremely low error rate with one or two serious abnormalities missed per 1000 women (DeMay, 1997). Moreover, the vast majority of those abnormalities will be detected at subsequent screens for women who are routinely screened appropriately, assuming acceptable levels of smear taking and laboratory performance with appropriate quality assurance procedures in place. This is because of the slow progression rates of abnormalities toward cancer, which provide a lengthy time for the disease to be identified through regular screening. Good quality, three-yearly conventional screening can prevent 93% of squamous cell invasive cancers, assuming total coverage of women eligible for screening²⁹ (Miller, 1996). The potential for improvement in effectiveness using new devices is therefore limited. Significant improvement in cost effectiveness is therefore only likely to be achieved by reductions in the cost of screening.

²⁸ Changes include the ageing of generations with increased risks of cervical cancer and the greater than average number of women in these generations as a result of the increase in fertility after the Second World War (Cox and Skegg, 1992)

²⁹ It should be noted that the Pap smear is less successful at detecting glandular lesions than squamous cell lesions. However, no new device has demonstrated advantages over the Pap test in detecting glandular abnormalities and therefore preventing adenocarcinoma.

Impact of introducing new devices into a population based cervical screening programme

Test sensitivity

From the current limited evidence, it is possible that new devices may marginally increase detection of low-grade abnormalities, though estimates of test sensitivity are uncertain. There is only a small margin for improvement in cervical cancer prevention for women screened regularly in the “conventional” manner, and improvement may be even less in New Zealand where more effective rescreening practices are employed. Therefore, new devices assumed to improve detection of abnormalities are not likely to have much of an impact on cervical cancer incidence or mortality. This has been borne out in cost effectiveness models that have assumed significant increases in the screening sensitivity of new devices. Moreover, our review of the literature indicates no evidence for improved detection of high-grade abnormalities by semi-automated and automated devices for cervical screening. New devices should demonstrate clinical effectiveness gains at HSIL because HSILs have a much higher probability to progress to cancer than lower grade abnormalities that are more likely to persist or regress.

Test specificity

There is minimal evidence relating to the test specificity of new devices for cervical screening from the literature reviewed to date. As efforts to increase test sensitivity are commonly associated with decreases in test specificity, new devices may decrease specificity and therefore increase false positive diagnoses, although data is inadequate at this time to determine this conclusively.

False positive diagnoses have received little attention in research evaluating new devices (Sawaya and Grimes, 1999). New devices have focussed on addressing false negatives, which have received a higher public profile compared with the much more common but less dramatic impact of false positives. As the vast majority of smears read are negative a small decrease in specificity would represent a far greater number of false positives compared with the impact of a similar decrease in sensitivity on numbers of false negatives. Costs of false positives include the health sector costs of unnecessary diagnosis, treatment and follow-up. These may add pressure for services that may impede women with genuine abnormalities receiving the care they require. False positives are also likely to result in social and psychological costs for women screened in terms of inconvenience, discomfort and distress. These societal costs are likely to be unacceptable from a population perspective. The potential for increases in the number of false positives from screening with new devices, if realised, would therefore have a profound impact on quality of life and the related cost effectiveness of the devices, as supported by the economic models reviewed in this report. Such costs associated with (potentially many) false positive results need to be compared with the benefits that may arise from any (potentially small) increase in the early detection of true positives.

Slide adequacy

There is some evidence to suggest that slides prepared using liquid-based slide preparation devices may lead to fewer “inadequate” or, as termed in New Zealand, “unsatisfactory” smears, although wide and overlapping ranges of the proportion of such smears have been reported for conventional smears and liquid-based methods (Payne et al., 2000). Smear adequacy rates between studies and countries are difficult to compare because they vary widely depending on the criteria used (which are poorly defined), the rates of infection in the population (Dr Gabriele Medley, Pathologist, Victorian Cytology Service, Australia, August 2000, *personal communication*). As the clinical significance of the inadequate smear is widely variable (McCrorry et al., 1999), slide adequacy was not reported for studies appraised in this review. Instead, the potential effect of a reduction of inadequate smears has been largely regarded as a cost issue, economic and psychological, in terms of the avoidance of the need for repeat smears.

Such benefits, if demonstrated conclusively, may help women who have persistent inadequate/unsatisfactory smears, or for communities with high rates of unsatisfactory smears (Shield et al., 1999). However, marginal benefits for such sub-populations do not warrant introduction of devices into a population-based screening programme, which was the focus of this review. A technology applied to a whole population might result in unacceptable costs to the detriment of other

aspects of screening programmes or health services generally, as well as other repercussions such as potential increases in false positive diagnoses discussed earlier.

Screening uptake

In New Zealand, ThinPrep has been marketed directly to women through posters in surgeries, pamphlets, and as part of full-page newspaper advertisements. There has been concern that women are being “wooded” into paying for ThinPrep smears by “market forces” before national policy is in place (Ponter, 1998). Pamphlets promoting ThinPrep in New Zealand describe it as “proven to give superior cervical smear results” and as developed to “improve the detection of cervical cancer or early stages of cancer”. The additional charge for the screening test (NZ\$15 - NZ\$20, on average) is justified as: “well worth the peace of mind that comes from knowing you have received the most reliable screening test available today” (Biotek, Undated). Describing ThinPrep, another pamphlet states “this advanced test gives you a new level of confidence in the reliability of your results” (Medlab South, Undated). Such emotionally charged and potentially misleading claims relating to new devices may have two effects on a national screening programme.

There are fears that promotion of new devices may undermine confidence in conventional Pap testing (Wain, 1997). Exaggerating the benefit of new devices may make women who do not take them up feel that they are receiving sub-standard care with conventional screening (Sawaya and Grimes, 1999). Moreover, women who are not currently participating in screening programmes may be discouraged further from having Pap smears, believing that they are inaccurate, unreliable and not worthy of their confidence. If these factors deter participation in regular cervical screening, new devices could have a significant impact on the ability of screening to reduce cervical cancer incidence (Brown and Garber, 1998; Brown and Garber, 1999a).

As devices such as ThinPrep are already being marketed to women and health providers in New Zealand, it is important that promotional information be balanced by material for the lay reader and health professionals (including smear takers) based on key findings of independent evidence such as found in this report. Such information could reassure women that the “single most important step a woman can take in cervical cancer prevention is to ensure that she has regular Pap tests with adequate follow-up” (van Deth, 1998).

Laboratories

This report focuses on whether new devices should be introduced into a population-based screening programme. However, uptake is a matter of individual choice, and laboratories may introduce them for reasons other than clinical effectiveness. For example, laboratories may attempt to increase their market share for health providers’ business through promotion of a new device, using this as an “edge” over competing laboratories (Ponter, 1998). There are other possible advantages for laboratories. Additional tests can be done with liquid-based slide preparation devices that may make it more cost effective for the laboratory. For example, liquid-based vials can be used for other pathology tests (including HPV DNA testing) and the device can also be used for taking non-gynaecological specimens. New devices may also be advantageous in ensuring greater uniformity across laboratories and assist in quality assurance protocols (McCrorry et al., 1999). There is also a suggestion that semi-automated computerised screening devices may be beneficial for laboratories in relieving pressure from the shortage of trained cytologists (McCrorry et al., 1999; McGoogan, 1997). Whilst few studies have considered specimen interpretation time, there is some evidence that LBS is associated with shorter times per slide. However, whether LBS devices are time-saving for cytologists overall is unclear as screening by LBS requires additional preparation and training time, and may require more intense concentration and the need for more breaks (Payne et al., 2000).

Alternative ways of improving effectiveness of cervical screening

Given the evidence base relating to new devices reviewed here, introduction of semi-automated and automated devices for cervical screening cannot be recommended for the New Zealand national screening programme at this time. However, whilst cervical screening through the Pap test has been remarkably successful, there is room for improvement. From a population perspective, it is important to consider whether resources that would be required to introduce new devices into the national

screening programme would lead to better outcomes for women screened if dedicated to other ways of improving the programme.

The effectiveness of a screening programme as a whole needs to be considered in trying to improve prevention of cervical cancer. Opportunities for improving cervical screening can occur at various steps along the screening pathway. These include:

- identifying eligible women,
- encouraging regular Pap tests,
- collecting the sample,
- fixing the sample,
- labelling and transport of the sample to the laboratory,
- screening and reporting of smears,
- rescreening strategies for quality control,
- clinical interpretation,
- treatment and follow-up.

Strategies for improving screening can be introduced at several of these stages.

Increasing up-take of routine screening by eligible women is likely to have greater impact on cervical cancer incidence and mortality than improving sensitivity of screening tests (McCrorry et al., 1999). Increasing access to and uptake of screening therefore should be the primary focus of efforts to reduce cervical cancer incidence because the majority of women who develop cervical cancer have not had regular screening (Solomon et al., 1998). In New Zealand, screening uptake is poorest for older women, whereas for younger women, a problem is inappropriate short interval re-screening (Coppell et al., 2000). This occurs when women are screened more frequently than three-yearly, as recommended³⁰. Annual screening compared with three yearly screening would prevent only an additional 2% of cases of cervical cancer while increasing the cost to the cervical screening programme by 400-500% (Paul et al., 1991).

Further strategies for enhancing the Pap test include improving the skills of smear takers and the use of more effective smear taking instruments (Payne et al., 2000). Improving laboratory standards including quality assurance procedures is clearly important, as highlighted in New Zealand by the Ministerial Inquiry into the under-reporting of cervical smear abnormalities in the Gisborne region. Whilst new devices are likely to have the most impact in laboratories with poor performance, such deficiencies should be addressed in other ways (Brown and Garber, 1998). A recent study found that implementing standardised methods of Quality Assurance³¹ in a Western Australian laboratory led to a doubling of reporting of HSIL from 0.47% to 0.91% over three years, maintaining high correlations with biopsy results (PPV of around 80%) (Sparkes et al., 2000). In New Zealand, specific laboratory standards (New Zealand Health Funding Authority, 1999c) being circulated for consultation by the

³⁰ In New Zealand, it is recommended that women have a repeat smear after one year when she has her first smear, and when the interval between smears is has been greater than five years, to minimise consequences of false negative results (Coppell et al., 2000).

³¹ The QA approach included standardised terminology, increasing teaching emphasis on inconclusive and false negatives, involvement of al staff in cytologic-histologic correlation exercises, combined pathologist-senior cytologist reporting of unsatisfactory and non-specific minor changes as well as more severe changes, regular reporting to cytotechnologists of the results of their screening and the laboratory results as a whole, performance standards for laboratories monitored by the RCPA, and the availability of comprehensive follow-up data from the Cervical Cancer Registries.

Health Funding Authority relate to a laboratory needing to have a minimum number of senior, experienced staff, compulsory accreditation, and a minimum number of smears screened per year. Adherence to stringent protocols for monitoring laboratory performance is also important. Finally, encouraging patient attendance for follow-up investigations and treatment is another avenue for improving the effectiveness of screening.

A summary of potential implications of introducing new devices into a population based screening programme is presented in **Table 8 (p. 98)**.

Table 8. Possible impact of the introduction of new devices into a population based cervical screening programme

Outcome	Likelihood based on current evidence base	Implications of outcomes for population based screening programme
Increasing costs of screening	Likely	<ul style="list-style-type: none"> - if screening costs paid by the government, there would be a reduction in other services - if screening test costs passed on to women, may reduce uptake, especially in poorer women
Increasing test sensitivity	Possible, but no reliable evidence	<ul style="list-style-type: none"> - reduce false negative rate, earlier detection of abnormalities - minimal impact on incidence of cervical cancer - minimal impact on mortality from cervical cancer
Decreasing test specificity	Possible, but no reliable evidence	<ul style="list-style-type: none"> - increase false positive rate and resulting unnecessary investigations and clinical management - significant increase in health service costs associated with false positives - significant decrease in quality of life for women with false positives (e.g. distress, discomfort, time)
Reduction of unsatisfactory (inadequate) smears	Possible, but no reliable evidence	<ul style="list-style-type: none"> - clinical impact uncertain - reduce health service costs with reduction in need for repeat smears - decrease in psychological costs with reduction in need for repeat smears

CONCLUSIONS

The following conclusions are based on the current evidence available from this report's systematic reviews of literature published on the clinical and cost effectiveness of new devices for population cervical screening (i.e. semi-automated and automated devices: ThinPrep, AutoCyte Prep, and AutoPap). Note that the impact of Human Papilloma Virus (HPV) DNA testing (to distinguish a low-risk group who require diminished surveillance) has not been considered in this review.

1. Estimates of test sensitivity and test specificity for the new devices could not be reliably determined. The research reviewed here provides no evidence for improved detection of high-grade abnormalities by new devices for cervical screening. New devices should demonstrate clinical effectiveness gains in detecting higher grade abnormalities. High-grade squamous intraepithelial lesions have a much higher probability of progressing to cancer than low grade abnormalities that are likely to regress.
2. Estimates of test sensitivity and specificity were the main source of uncertainty in the economic models investigating the cost effectiveness of new devices. In economic models where improved detection from the introduction of new devices was assumed, the impact of new devices on days-of-life saved was extremely small for women screened at three yearly intervals. Cost effectiveness may be even poorer in New Zealand where more effective rescreening practices are employed for conventional screening.
3. Any increases in sensitivity resulting from the introduction of new devices may come at the cost of decreased specificity. This would lead to increases in false positive results (where slides are read as abnormal when the woman does not have a cervical abnormality). False positives lead to health sector costs of unnecessary diagnosis, treatment and follow-up that may lead to pressure on health services to the detriment of women with true abnormalities. Investigations of false positives also are associated with social and psychological costs for women including inconvenience, discomfort and distress. The potential for increases in false positives would, if realised, have a profound impact on quality of life and the related cost effectiveness of the devices.
4. Higher quality research is required to generate valid estimates of test sensitivity and specificity. Methodological limitations to address include the application of appropriate reference standards for verification of cytological diagnoses, including test negatives. Economic modelling studies will be more meaningful with more valid estimates of test characteristics, and a comprehensive measurement of costs of screening from a societal perspective, including careful investigation of the impact of screening and clinical management on quality of life.
5. It is important that promotional information for new devices is balanced by material for health professionals and for women based on key findings of independent evidence such as found in this report. Additionally, the New Zealand Health Funding Authority/Ministry of Health should investigate legal avenues to restrict advertising making unsubstantiated claims for new devices.
6. The vast majority of missed abnormalities will be detected at subsequent screens for women who are routinely screened appropriately, assuming acceptable levels of smear taking and laboratory performance. Three yearly cervical screening using the conventional Pap smear can be highly effective, preventing 93% of cervical cancer, assuming all eligible women are screened. Therefore, the Pap test should remain the standard of care in population cervical screening.
7. The introduction of new devices for cervical screening cannot be recommended for the New Zealand national cervical screening programme at this time.
8. Rather than committing resources to the introduction of new devices into the national screening programme, better outcomes may be achieved for women screened if resources are directed to other ways of improving the programme. These strategies could include the following:
 - *increasing up-take of routine screening by eligible women,*
 - *ensuring that women are screened at appropriate intervals,*
 - *implementing standards for smear taking, and ensuring the use of the most effective smear taking instruments,*
 - *implementing strict laboratory standards and quality assurance,*
 - *and ensuring adequate follow-up and treatment where required.*

9. Resources should be directed to the appropriate monitoring of the national cervical screening programme.
10. This review is based on research published to end of May 2000. It is recommended the conclusions of this report be revisited in 12 months (October 2001).

References

- Abulafia, O., & Sherer, D. M. (1999). Automated cervical cytology: meta-analyses of the performance of the AutoPap 300 QC System. *Obstetrical & Gynecological Survey*, 54, 469-76.
- Adams, J. (1991). *Cervical screening programmes: a review of the literature and its implications for New Zealand*. Wellington: Dept. of Health.
- American College of Obstetricians & Gynecologists Committee on Gynecologic Practice (1998). ACOG committee opinion. New Pap test screening techniques. Number 206, August 1998. Committee on Gynecologic Practice. American College of Obstetricians and Gynecologists. *International Journal of Gynaecology & Obstetrics*, 63, 312-4.
- Ashfaq, R., Gibbons, D., Vela, C., Saboorian, M. H., & Iliya, F. (1999). ThinPrep pap test. Accuracy for glandular disease. *Acta Cytologica*, 43, 81-5.
- Austin, R. M. (1998). Implementing liquid-based gynecologic cytology: balancing marketing, financial, and scientific issues. *Cancer*, 84, 193-6.
- Austin, R. M., & Ramzy, I. (1998). Increased detection of epithelial cell abnormalities by liquid-based gynecologic cytology preparations. A review of accumulated data. *Acta Cytologica*, 42, 178-84.
- Australian Health Technology Advisory Committee (1998). *Review of automated and semi-automated cervical screening devices*. Canberra: Commonwealth Dept. of Health and Family Services.
- Bartels, P. H., Bibbo, M., Hutchinson, M. L., Gahm, T., Grohs, H. K., Gwi-Mak, E., Kaufman, E. A. et al. (1998). Computerized screening devices and performance assessment: Development of a policy towards automation: IAC task force summary. *Acta Cytologica*, 42, 59-68.
- Bartels, P. H., & Vooijs, G. P. (1999). Automation of primary screening for cervical cancer. Sooner or later?. *Acta Cytologica*, 43, 7-12.
- Bedrossian, C., Bonfiglio, T., Davey, D., Hutchinson, M., Kaufman, E., Krieger, P., Mody, D. et al. (1998). Proposed guidelines for secondary screening (rescreening) instruments for gynecologic cytology. Intersociety Working Group for Cytology Technologies. *Acta Cytologica*, 42, 1311-4.
- Bell, S., Porter, M., Kitchener, H., Fraser, C., Fisher, P., & Mann, E. (1995). Psychological response to cervical screening. *Preventive Medicine*, 24, 610-6.
- Bibbo, M., & Hawthorne, C. (1999). Performance of the AutoPap primary screening system at Jefferson University Hospital. *Acta Cytologica*, 43, 27-9.
- Bibbo, M., Hawthorne, C., & Zimmerman, B. (1999). Does use of the AutoPap assisted primary screener improve cytologic diagnosis? *Acta Cytologica*, 43, 23-6.
- Biotek (Undated). *The ThinPrep Pap Test*. Auckland: Biotek. [promotional pamphlet]
- Bishop, J. W., Bigner, S. H., Colgan, T. J., Husain, M., Howell, L. P., McIntosh, K. M., Taylor, D. A., & Sadeghi, M. H. (1998). Multicenter masked evaluation of AutoCyte PREP thin layers with matched conventional smears. Including initial biopsy results. *Acta Cytologica*, 42, 189-97.
- Bolick, D. R., & Hellman, D. J. (1998). Laboratory implementation and efficacy assessment of the ThinPrep cervical cancer screening system. *Acta Cytologica*, 42, 209-13.

- Braggett, D., Lea, A., Carter, R. et al. (1993). *Issues in cervical cancer screening and treatment - new technologies and costs of alternative management strategies*. Canberra: Australian Institute of Health & Welfare.
- Brown, A. B., & Garber, A. M. (1998). The cost-effectiveness of three new technologies to enhance Pap testing. Chicago, ILL: Blue Cross & Blue Shield Association Technology Evaluation Center.
- Brown, A. D., & Garber, A. M. (1999a). Cost-effectiveness of 3 methods to enhance the sensitivity of Papanicolaou testing. *JAMA*, 281, 347-53.
- Brown, A. D., & Garber, A. M. (1999b). Cost-effectiveness of methods to enhance sensitivity of papanicolaou testing. *JAMA*, 282, 1420.
- Carpenter, A. B., & Davey, D. D. (1999). ThinPrep(TM) Pap test(TM): Performance and biopsy follow-up in a University Hospital. *Cancer*, 87, 105-12.
- Chamberlain, J. (1986). Reasons that some screening programmes fail to control cervical cancer. In M. Hakama, A. B. Miller & N. E. Day (Eds.), *Screening for cancer of the uterine cervix*. Lyons: International Agency for Research on Cancer.
- Chock, C., Irwig, L., Berry, G., & Glasziou, P. (1997). Comparing dichotomous screening tests when individuals negative on both tests are not verified. *Journal of Clinical Epidemiology*, 50, 1211-7.
- Coleman, D. V. (1998). Evaluation of automated systems for the analysis of cervical smears. *Cytopathology*, 9, 359-68.
- Colgan, T. J., Patten, S. F. J., & Lee, J. S. (1995). A clinical trial of the Autopap 300 QC system for quality control of cervicovaginal cytology in the clinical laboratory. *Acta Cytologica*, 39, 1191-8.
- Coppell, K., Paul, C., & Cox, B. (2000). An evaluation of the National Screening Programme Otago site. *New Zealand Medical Journal*, 113, 48-51.
- Corkill, M., Knapp, D., Martin, J., & Hutchinson, M. L. (1997). Speciman adequacy of ThinPrep sample preparations in a direct-to-vial study. *Acta Cytologica*, 41, 39-44.
- Cox, B., & Skegg, D. C. G. (1992). Projections of cervical cancer mortality and incidence in New Zealand: the possible impact of screening. *Journal of Epidemiology and Community Health*, 46, 373-7.
- Cuzick, J., Sasieni, P., Davies, P., Adams, J., Normand, C., Frater, A., van Ballegooijen, M., & van den Akker, E. (1999). A systematic review of the role of human papillomavirus testing within a cervical screening programme. *Health Technology Assessment* 3, 1-104.
- Davey, D. D. (1997). Quality and liability issues with the Papanicolaou smear: the problem of definition of errors and false-negative smears. *Archives of Pathology & Laboratory Medicine*, 121, 267-9.
- DeMay, R. M. (1997). Common problems in Papanicolaou smear interpretation. *Archives of Pathology and Laboratory Medicine*, 121, 229-38.
- Diaz-Rosario, L. A., & Kabawat, S. E. (1999). Performance of a fluid-based, thin-layer papanicolaou smear method in the clinical setting of an independent laboratory and an outpatient screening population in New England. *Archives of Pathology & Laboratory Medicine*, 123, 817-21.
- Drummond, M. F., O'Brien, B., Stoddart, G. L., & Torrance, T. W. (1997). *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press.

- Dupree, W. B., Suprun, H. Z., Beckwith, D. G., Shane, J. J., & Lucente, V. (1998). The promise and risk of a new technology: The Lehigh Valley Hospital's experience with liquid-based cervical cytology. *Cancer*, 84, 202-7.
- Dyer, C. (1999). Health authority loses cervical smear appeal. *BMJ*, 319, 1391.
- ECRI Health Technology Assessment Information Service (1999). Automated monolayer slide preparation systems for pap smear screening: ThinPrep 2000. Plymouth Meeting, PA: ECRI.
- ECRI Health Technology Assessment Information Service (2000). Automated pap smear screening technologies: the autopap system. Plymouth Meeting, PA: ECRI.
- Eddy, D. M. (1990). Screening for cervical cancer. *Annals of Internal Medicine*, 113, 214-26.
- Fahey, M., Irwig, L., & Macaskill, P. (1995). Meta-analysis of Pap test accuracy. *American Journal of Epidemiology*, 141, 680-9.
- Fahs, M. C., Mandelblatt, J., Schechter, C., & C., M. (1992). Cost effectiveness of cervical cancer screening for the elderly. *Annals of Internal Medicine*, 117, 520-7.
- Free, K., Roberts, S., Boume, R., Dickie, G., Ward, B., Wright, G., & Hill, B. (1991). Cancer of the cervix: old and young, new and then. *Gynecological Oncology*, 43, 129-36.
- Frommer, D. J., Kapparis, A., & Brown, M. K. (1988). Improved screening for colorectal cancer by immunological detection of occult blood. *British Medical Journal*, 296, 1092-4.
- Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). Cost effectiveness in health and medicine. New York: Oxford University Press.
- Halford, J. (1998). Cervical cytology - A review of evolving technologies. *Australian Journal of Medical Science*, 19, 8-19.
- Hutchinson, M. L. (1996). Assessing the costs and benefits of alternative rescreening strategies. *Acta Cytologica*, 40, 4-8.
- Hutchinson, M. L. (2000). Utility of liquid-based cytology for cervical carcinoma screening: author reply. *Cancer Cytopathology*, 90, 68-9.
- Hutchinson, M. L., Zahniser, D. J., Sherman, M. E., Herrero, R., Alfaro, M., Bratti, M. C., Hildesheim, A. et al. (1999). Utility of liquid-based cytology for cervical carcinoma screening: results of a population-based study conducted in a region of Costa Rica with a high incidence of cervical carcinoma. *Cancer*, 87, 48-55.
- Inhorn, S. L., Wilbur, D., Zahniser, D., & Linder, J. (1998). Validation of the ThinPrep Papanicolaou test for cervical cancer diagnosis. *Journal of Lower Genital Tract Disease*, 2, 208-12.
- International Academy of Cytology (1992). The revised Bethesda system for reporting cervical/vaginal cytologic diagnoses: report of the 1991 Bethesda Workshop. *Acta Cytologica*, 36, 273-6.
- Intersociety Working Group for Cytology Technologies (1997). Proposed guidelines for primary screening instruments for gynecologic cytology. Intersociety Working Group for Cytology Technologies. *Acta Cytologica*, 41, 924-9.
- Irwig, L., & Glasziou, P. (1996). The Cochrane Methods Working Group on systematic review of screening and diagnostic tests: recommended methods.
- Jones, M. H., Singer, A., & Jenkins, D. (1996). The mildly abnormal cervical smear: patient anxiety and choice of management. *Journal of the Royal Society of Medicine*, 89, 257-260.

- Jones, R. W., Best, D. V., Cox, B., Fitzgerald, N. W., Hill, M., Jennings, P., Peddie, D., & Sage, M.J. (2000). Guidelines for the management of women with abnormal cervical smears 1998. *New Zealand Medical Journal*, 113, 168-71.
- Kaminsky, F. C., Benneyan, J. C., & Mullins, D. L. (1997). Automated rescreening in cervical cytology. Mathematical models for evaluating overall process sensitivity, specificity and cost. *Acta Cytologica*, 41, 209-23.
- Koss, L. G. (2000). Utility of liquid-based cytology for cervical carcinoma screening. *Cancer Cytopathology*, 90, 67-8.
- Koss, L. G., Stewart, F., Foote, F. W., Jordan, M. J., Bader, G. M., & Day, E. (1963). Some histological aspects of behavior of epidermoid carcinoma in situ and related lesions of the uterine cervix. A long-term prospective study. *Cancer*, 16, 1160-211.
- Krieger, P. A., McGoogan, E., Voojjs, G. P., Amma, N. S., Cochand-Priollet, B., Colgan, T. J., Davey, D. D. et al. (1998). Quality assurance/control issues: IAC Task Force summary. *Acta Cytologica*, 42, 133-40.
- Last, J. M. (1995). A dictionary of epidemiology. New York: Oxford University Press.
- Lee, K. R., Ashfaq, R., Birdsong, G. G., Corkill, M. E., McIntosh, K. M., & Inhorn, S. L. (1997). Comparison of conventional Papanicolaou smears and a fluid-based, thin-layer system for cervical cancer screening. *Obstetrics & Gynecology*, 90, 278-84.
- Leiman, G. (1999). Thin-layer cytology meets viral hybrid capture. *Advances in Anatomic Pathology*, 6, 161-4.
- Linder, J. (1998). Recent advances in thin-layer cytology. *Diagnostic Cytopathology*, 18, 24-32.
- Linder, J., & Zahniser, D. (1998). ThinPrep Papanicolaou testing to reduce false-negative cervical cytology. *Archives of Pathology & Laboratory Medicine*, 122, 139-44.
- Lonky, S. A. (2000). Economic impact of automated primary screening for cervical cancer. *The Journal of Reproductive Medicine*, 45, 83-4.
- Martin-Hirsch, P., Lilford, R., Jarvis, G., & Kitchener, H. C. (1999). Efficacy of cervical-smear collection devices: a systematic review and meta-analysis. *Lancet*, 354, 1763-70.
- McCorry, D. C., Matchar, D. B., Bastian, L., Datta, S., Hasselblad, V., Hickey, J., Myers, E. et al. (1999). Evaluation of cervical cytology: Evidence Report/Technology Assessment Number 5. (Prepared by Duke University under Contract No. 290-97-0014). AHCPR Publication No. 99-E010. Rockville, MD: Agency for Health Care Policy and Research (AHCPR).
- McGoogan, E. (1997). Automation and cervical cytopathology: an overview. In E. Franco & J. Monsonego (Eds.), *New developments in cervical cancer screening and prevention* (pp. 265-273). Oxford: Blackwell Science.
- McGoogan, E., Colgan, T. J., Ramzy, I., Cochand-Priollet, B., Davey, D. D., Grohs, H. K., Gurley, A. M. et al. (1998). Cell preparation methods and criteria for sample adequacy. International Academy of Cytology Task Force summary. Diagnostic Cytology Towards the 21st Century: An International Expert Conference and Tutorial. *Acta Cytologica*, 42, 25-32.
- McMeekin, D. S., McGonigle, K. F., & Vasilev, S. A. (1997). Cervical cancer prevention: towards cost-effective screening. *Medscape Women's Health*, 2 no.12.
- Medlab South (Undated). *Now you can feel even better about your pap smear*. Christchurch: Medlab South. [promotional pamphlet]

- Melamed, M. R., Hutchinson, M. L., Kaufman, E. A., Schechter, C. B., Garner, D., Kobler, T. P., Krieger, P. A. et al. (1998). Evaluation of costs and benefits of advances in cytologic technology. International Academy of Cytology Task Force summary. Diagnostic Cytology Towards the 21st Century: An International Expert Conference and Tutorial. *Acta Cytologica*, 42, 69-75.
- Members of the Working Party on Cervical Screening (New Zealand) (1998). Recommendations for cervical screening 1997. *New Zealand Medical Journal*, 111, 94-8.
- Miller, A. (1996). Screening for cervical cancer. In S. C. Rubin & W. J. Hoskins (Eds.), *Cervical cancer and preinvasive neoplasia*. Philadelphia, PA: Lippincott Raven.
- Miller, W. C. (1998). Bias in discrepant analysis: when two wrongs don't make a right. *Journal of Clinical Epidemiology*, 51, 219-31.
- Minge, L., Fleming, M., VanGeem, T., & Bishop, J. W. (2000). AutoCyte Prep system vs conventional cervical cytology: comparison based on 2,156 cases. *Journal of Reproductive Medicine*, 45, 179-84.
- Minnesota Health Technology Advisory Committee (HTAC) (1999). New technologies for cervical cancer screening [St. Paul, MN.]: HTAC. Available from: www.health.state.mn.us/htac.
- Mitchell, H. S., & Giles, G. C. (1996). Cancer diagnosis after a report of negative cervical cytology. *Medical Journal of Australia*, 164, 270-3.
- Nanda, K., McCrory, D. C., Myers, E. R., Bastian, L. A., Hasselblad, V., Hickey, J. D., & Matchar, D. B. (2000). Accuracy of the papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Annals of Internal Medicine*, 132, 810-9.
- New Zealand Health Funding Authority (1999). *Cervical screening: guidelines for the management of women with abnormal cervical smears*. [Wellington]: Health Funding Authority.
- New Zealand Health Information Service (1999). Deaths from cancer at selected sites - numbers and rates by sex, 1995-97. *Mortality and demographic data 1997*. Wellington: New Zealand Health Information Service. Available from <http://www.nzhis.govt.nz/stats/cancerstats-p.htm>. Cited 8 August 2000.
- New Zealand Committee of Inquiry into Allegations Concerning the Treatment of Cervical Cancer at National Women's Hospital and into Other Related Matters (1988). *The report of the Committee of Inquiry into Allegations Concerning the Treatment of Cervical Cancer at national Women's Hospital and into Other Related Matters*. Auckland: The Committee.
- New Zealand Ministry of Health (1993). *Government policy for national cervical screening*. Wellington: Ministry of Health.
- New Zealand Department of Health (1991). *Government policy for cervical screening*. Wellington: Department of Health.
- New Zealand Health Funding Authority (1999a). *Evaluation and monitoring plan: National cervical screening programme: draft, version 3.0*. [Wellington]: Health Funding Authority.
- New Zealand Health Funding Authority (1999b). *National cervical screening programme: proposed national indicators*. [Wellington]: Health Funding Authority.
- New Zealand Health Funding Authority (1999c). *Policy and quality standards for the National Cervical Screening Programme: draft 1 for consultation, November 1999*. [Wellington]: Health Funding Authority.

- New Zealand Ministry of Health (1996). *National Cervical Screening Programme policy*. [Wellington]: Ministry of Health.
- New Zealand Ministry of Health (1997). *A brief narrative on Maori women and the National Cervical Screening Programme*. [Wellington]: Ministry of Health.
- Östör, A. (1993). Natural history of cervical intraepithelial neoplasia: a critical review. *International Journal of Gynecological Pathology*, 12, 186-92.
- Papillo, J. L., Zarka, M. A., & St John, T. L. (1998). Evaluation of the ThinPrep Pap test in clinical practice. A seven-month, 16,314-case experience in northern Vermont. *Acta Cytologica*, 42, 203-8.
- Paul, C., Bagshaw, S., Bonita, R., Durham, G., Fitzgerald, N. W., Jones, R. W. Marshall, B., and McAvoy, B.R. (1991). Cervical screening recommendations: a working group report. *New Zealand Medical Journal*, 104, 291-5.
- Payne, N., Chilcott, J., & McGoogan, E. (2000). Liquid-based cytology in cervical screening. A report by the School of Health and Related Research (SchHARR), the University of Sheffield, for the NCCHTA on behalf of NICE. Sheffield: SchHARR, University of Sheffield.
- Ponter, E. (1998). New technology for cervical screening: an advantage for the wealthy...? [unpublished paper], *Annual Conference of the Sociological Association of Aotearoa(NZ)*. Eastern Institute of Technology, Taradale. New Zealand, 28 November 1998.
- Raab, S. S., Zaleski, M. S., & Silverman, J. F. (1999). The cost-effectiveness of the cytology laboratory and new cytology technologies in cervical cancer prevention. *American Journal of Clinical Pathology*, 111, 259-66.
- Roberts, J. M., Thurloe, J. K., Bowditch, R. C., Humcevic, J., & Lavery, C. R. (1999). Comparison of ThinPrep and Pap smear in relation to prediction of adenocarcinoma in situ. *Acta Cytologica*, 43, 74-80.
- Rosenthal, D. L. (1998). Automation and the endangered future of the Pap test. *Journal of the National Cancer Institute*, 90, 738-49.
- Sackett, D. L., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (1997). *Evidence-based medicine*. New York: Churchill Livingstone.
- Sawaya, G. F., & Grimes, D. A. (1999). New technologies in cervical cytology screening: a word of caution. *Obstetrics & Gynecology*, 94, 307-10.
- Schechter, C. B. (1996). Cost-effectiveness of rescreening conventionally prepared cervical smears by PAPNET testing. *Acta Cytologica*, 40, 1272-82.
- Sherlaw-Johnson, C., Gallivan, S., Jenkins, D., & Jones, M. H. (1994). Cytological screening and management of abnormalities in prevention of cervical cancer: an overview with stochastic modelling. *Journal of Clinical Pathology*, 47, 430-5.
- Sherman, M. E., Schiffman, M., Herrero, R., Kelly, D., Bratti, C., Mango, L. J., Alfaro, M. et al. (1998). Performance of a semiautomated Papanicolaou smear screening system: results of a population-based study conducted in Guanacaste, Costa Rica [see comments]. *Cancer*, 84, 273-80.
- Shield, P. W., Nolan, G. R., Phillips, G. E., & Cummings, M. C. (1999). Improving cervical cytology screening in a remote, high risk population. *Medical Journal of Australia*, 170, 255-8.
- Skegg, D. C., Paul, C., R.J. S., Fitzgerald, N. W., Barham, P. M., & Clements, C. J. (1985). Recommendations for routine cervical screening. *New Zealand Medical Journal*, 98, 636-9.

- Smith, B. L., Lee, M., Leader, S., & Wertlake, P. (1999). Economic impact of automated primary screening for cervical cancer. *Journal of Reproductive Medicine*, 44, 518-28.
- Smith, B. L., Lee, M., Leader, S., & Wertlake, P. (2000). Economic impact of automated primary screening for cervical cancer. *The Journal of Reproductive Medicine*, 45, 83-84.
- Solomon, D., Davey, D., Nozawa, S., & Syrjanen, K. J. (1998). Future directions in cervical cytology. *Acta Cytologica*, 42, 1-4.
- Sparkes, J., Schoolland, M., Barrett, P., Kurinczuk, J. J., Mitchell, K. M., & Sterrett, G. F. (2000). Trends in the frequency and predictive value of reporting high grade abnormalities in cervical smears. *Cancer Cytopathology*, 90, 215-21.
- Spitzer, M. (1998). Cervical screening adjuncts: recent advances. *American Journal of Obstetrics & Gynecology*, 179, 544-56.
- Syrjanen, K., Kataja, V., Yliskoski, M., Chang, F., Syrjanen, S., & Saarikoski, S. (1992). Natural history of cervical human papilloma virus lesions does not substantiate the biological relevance of the Bethesda System. *Obstetrics & Gynecology*, 79, 675-82.
- van Deth, A. (1998). Objective assessment of new technologies is coming. *Medical Journal of Australia*, 168, 254.
- Vassilakos, P., Griffin, S., Megevand, E., & Campana, A. (1998). CytoRich liquid-based cervical cytologic test. Screening results in a routine cytopathology service. *Acta Cytologica*, 42, 198-202.
- Vassilakos, P., Saurel, J., & Rondez, R. (1999). Direct-to-vial use of the AutoCyte PREP liquid-based preparation for cervical-vaginal specimens in three European laboratories. *Acta Cytologica*, 43, 65-8.
- Vassilakos, P., Schwartz, D., de Marval, F., Yousfi, L., Broquet, G., Mathez-Loic, F., Campana, A. et al. (2000). Biospy-based comparison of liquid-based, thin-layer preparations to conventional pap smears. *The Journal of Reproductive Medicine*, 45, 11-16.
- Wain, G. V. (1997). Automation in cervical cytology: Whose cost and whose benefit? *Medical Journal of Australia*, 167, 460-1.
- Wang, T. Y., Chen, H. S., Yang, Y. C., & Tsou, M. C. (1999). Comparison of fluid-based, thin-layer processing and conventional Papanicolaou methods for uterine cervical cytology. *Journal of the Formosan Medical Association*, 98, 500-5.
- Weintraub, J., & Morabia, A. (2000). Efficacy of a liquid-based thin layer method for cervical cancer screening in a population with a low incidence of cervical cancer. *Diagnostic Cytopathology*, 22, 52-9.
- Wilbur, D. C., Prey, M. U., Miller, W. M., Pawlick, G. F., & Colgan, T. J. (1998). The AutoPap system for primary screening in cervical cytology. Comparing the results of a prospective, intended-use study with routine manual practice. *Acta Cytologica*, 42, 214-20.
- Wilbur, D. C., Prey, M. U., Miller, W. M., Pawlick, G. F., Colgan, T. J., & Taylor, D. D. (1999). Detection of high grade squamous intraepithelial lesions and tumours using the autopap system. *Cancer Cytopathology*, 87, 354-8.
- Wise, J. (2000). UK pilot scheme for HPV testing announced. *BMJ*, 320, 600.

List of abbreviations and acronyms

ACP	—	AutoCyte Prep
AIS	—	adenocarcinoma in situ
AGUS	—	atypical glandular cells of undetermined significance
AHCPR	—	Agency for Health Care Policy and Research (USA)
AHRQ	—	Agency for Healthcare Research and Quality (formerly AHCPR) (USA)
AHTAC	—	Australian Health Technology Advisory Committee
AP	—	AutoPap
ASCUS*	—	atypical squamous cells of undetermined significance
CE	—	clinical effectiveness
CIN	—	cervical intraepithelial neoplasia
CINAHL	—	Cumulative Index to Nursing and Allied Health Literature
CIS	—	carcinoma in situ
CLIA	—	Clinical Laboratory Improvement Amendments (USA)
Cum Se	—	Cumulative Sensitivity (after primary screening and rescreening)
DNA	—	deoxyribonucleic acid
FDA	—	Food and Drug Administration (USA)
FNR	—	false negative rate
FPR	—	false positive rate
HFA	—	Health Funding Authority (NZ)
HPV	—	Human Papilloma Virus
HSIL*	—	high-grade squamous intraepithelial lesion
HTA	—	health technology assessment
IAC	—	International Academy of Cytology
INAHTA	—	International Network of Agencies for Health Technology Assessment
LBS	—	liquid-based screening (sample collection and slide preparation device)

LSIL*	—	low-grade squamous intraepithelial lesion
NCSP	—	National Cervical Screening Programme (NZ)
NHS	—	National Health Service (UK)
NSI	—	Neomedical Systems
NZHTA	—	New Zealand Health Technology Assessment (The Clearing House for Health Outcomes and Health Technology Assessment)
Pap	—	Papanicolaou
PMA	—	Pre-Market Approval
PPV	—	positive predictive value
PS	—	Papanicolaou (Pap) smear
QA	—	quality assurance
QC	—	quality control
RCPA	—	Royal College of Pathologists of Australasia
RCT	—	randomised controlled trial
SCC	—	squamous cell carcinoma
Se	—	sensitivity
Sp	—	specificity
SD	—	standard deviation
SIL	—	squamous intra-epithelial lesion
TBS	—	The Bethesda System
TP	—	ThinPrep
TPR	—	True positive rate
WNL	—	within normal limits

* When grade of abnormality is reported followed by “+” (e.g. LISIL+) this refers to abnormalities at this grade as well as higher grade lesions.

Glossary

Adenocarcinoma ~ A rarer form of cervical cancer where the tumour arises from glandular tissue.

Atypical squamous cells of undetermined significance (ASCUS) ~ Cellular abnormalities that are more marked than those attributable to reactive changes, but that quantitatively or qualitatively fall short of a definitive diagnosis of squamous intraepithelial lesion (SIL). Because the cellular changes in the ASCUS category may reflect an exuberant benign change or a potentially serious lesion, which cannot be unequivocally classified, they are interpreted as being of undetermined significance

Bethesda system (TBS) ~ A system for cervical/vaginal cytologic diagnoses, developed at the National Cancer Institute-sponsored workshop in December 1988. Reports include a descriptive diagnosis and an evaluation of specimen adequacy.

Bias ~ Deviation of results or inferences from the truth, or processes leading to such deviation.

Biopsy ~ In a cervical biopsy a sample of tissue is removed to be examined under a microscope as an aid in diagnosis. A cervical biopsy is taken from the cervix which appears abnormal through visual examination.

Blinded study ~ A study in which observers and/or subjects are kept ignorant of the group to which they are assigned.

Case control study ~ An epidemiological study involving the observation of *cases* (persons with the disease, such as cervical cancer) and a suitable *control* (comparison, reference) group of persons without the disease. The relationship of an attribute (e.g. positive screening test result) to the disease is examined by comparing *retrospectively* the past history of the people in the two groups with regard to how frequently the attribute is present. See also *nested case control*.

Case series ~ A descriptive study of a subset of a defined population (i.e. a single patient or group of patients) which aims to describe the association between factors or attributes which the sample are exposed to, and the probability of occurrence of a given disease or other outcome. Case series are collections of individual case reports, which may occur within a fairly short period of time.

Cervical intraepithelial neoplasia (CIN) ~ Precancerous changes found either by cervical screening or by histology. The degree of abnormality ranges from CIN I (mild) to CIN III (severe).

Cervical screening ~ Cervical screening is used as a means to detect precursors of more serious disease in apparently well women, allowing treatment to prevent the development of that disease.

Cohort study ~ The analytic method of epidemiologic study in which subsets of a defined population can be identified who are, have been, or in the future may be exposed or not exposed in different degrees, to a factor or factors (e.g. receiving a screening test for cervical cancer) hypothesised to influence the probability of occurrence of a given disease or other outcome (e.g. positive biopsy). Studies usually involve the observation of a large population, for a prolonged period (years), or both.

Colposcopy ~ A diagnostic examination of the vagina and cervix or neck of the uterus, using a colposcope (a lighted magnifying instrument resembling a small mounted pair of binoculars), to examine the vaginal walls and cervix for visible evidence of cervical abnormality. It is often combined with cervical biopsy.

Confounder ~ A third variable that indirectly distorts the relationship between two other variables, because it is independently associated with each of the variables.

Cost effectiveness (CE) ~ Involves the relationship between costs and effects, providing information on whether a technology is being delivered to those who would benefit from it with an optimal use of resources. It is expressed as a ratio of the effects (number of lives saved, number of disability days avoided) obtained for a specific cost (expressed in dollars). For example, the numerator may be the difference in lifetime costs between one intervention and another, while the denominator may be the difference in life expectancies associated with the two interventions. Low cost effectiveness ratios are desirable.

Coverage ~ The number, percent, or proportion of eligible women reached by a programme.

Cytology ~ The study of cells using a microscope. Used in cervical screening to detect cancer or cell changes, which may be precursors of cancer.

Direct-to-vial (DTV) study ~ This study design involves only a sample. The LBC test is prepared so that the whole cervical sample is rinsed directly from the sampling instrument/s into a vial. The DTV approach is an alternative sample delivery method to the “split sample” technique when a conventional Pap smear is made first and the rest of the cells on the sampling instrument are rinsed into the liquid in the vial.

Discounting ~ In cost effectiveness studies, future dollar costs and benefit streams are reduced or “discounted” by a percentage to reflect the fact that money spent or saved in the future should not weigh as heavily in programme decisions as dollars spent or saved now.

Discrepancy study ~ A study involving the verification of only discordant diagnoses of two interventions (such as two screening tests); that is, where one test provides a positive result and the other provides a negative result.

Dominance ~ In cost effectiveness studies, an alternative is eliminated by dominance if it is both less effective and more costly than (i.e. dominated by) at least one other alternative.

Dysplasia ~ Abnormal cell growth.

Extended dominance ~ In cost effectiveness studies, an alternative is eliminated by extended dominance if it has a higher cost effectiveness ratio than a more effective option.

False negative smear ~ A smear, which is reported as negative, from a woman who has a high-grade cervical lesion, diagnosed by biopsy within two years of having that smear.

False positive smear ~ A smear which is reported to have a high-grade abnormality, in a woman in whom investigation, including biopsy, within six months of the smear, does not confirm an abnormality.

Final truth determination ~ Use of a reference standard to provide an accurate or “truth” diagnosis for verification of positive and negative diagnoses by a screening or diagnostic test (see also “reference standard”).

Grey literature ~ That which is produced by all levels of government, academics, business and industry, in print and electronic formats, but which is not controlled by commercial publishers.

High-grade lesion ~ A cytological diagnosis encompassing CIN II and CIN III (moderate dysplasia, severe dysplasia and carcinoma in situ).

High risk groups ~ Usually refers to groups of women that have been identified as having a higher than expected, or higher than average for the population as a whole, incidence of the disease in question. This group is traditionally also under-screened and in New Zealand includes women over 45 years old, Maori and Pacific Island women, and women with previously detected cervical abnormalities.

Histology ~ The microscopic study of the minute structure and composition of tissues.

Human papilloma virus (HPV) ~ A member of a group of viruses, some of which are sexually transmitted.

Inadequate smear ~ A smear that for technical reasons cannot be reported on by the laboratory (UK term for unsatisfactory smear)

Incidence ~ The number of new events (cases; e.g. of disease) occurring during a certain period, in a specified population.

Invasive cancer of the cervix ~ A condition where cancerous cells spread beyond the surface epithelium into the underlying tissues. It is diagnosed by clinical examination with biopsy. The cervical smear is not a reliable method of diagnosing cervical cancer but it may be a useful predictor of an invasive lesion.

Low-grade lesion ~ A cytological diagnosis encompassing the changes seen with HPV infection and/or CIN I (mild dysplasia).

Liquid-based screening (LBS) ~ Automated liquid-based sample collection and slide preparation system designed to provide more representative cell samples of evenly dispersed cells. The two currently available devices for LBS reviewed in this report are Thinprep™ (Cytoc Corporation) and AutoCyte Prep™ (TriPath Imaging).

Matching ~ The process of making a study group and a comparison group comparable with respect to extraneous factors.

Meta-analysis ~ Any systematic method that uses statistical analysis to integrate the data from a number of independent studies.

Morbidity ~ The ratio of sick to well persons in a given population.

Mortality ~ The number of deaths from a specified disease which are diagnosed or reported during a defined period of time in a given population.

Nested case control study ~ A case control study in which cases and controls are drawn from the population in a cohort study. That is the case control study is “nested” within the cohort study design so that the effects of some potential confounding variables are reduced or eliminated. A case control study can also be nested into a case series study. See also *case control study*, *cohort study*, and *case series study*.

Non-invasive cervical cancer ~ Abnormal tissue on the cervix, ranging from low-grade to high-grade abnormal cell growth.

Pap (Papanicolaou) smear ~ A technique used with the aim of detecting precursors of cervical cancer, or actual cancer at the earliest possible stage. The test is based on the examination of cells, which are removed from the cervix and examined under the microscope. Also called the cervical smear, the Pap smear is the conventional test commonly used as a control (comparison or reference) group in studies comparing its effectiveness with new devices for cervical screening.

Prevalence ~ The number of events in a given population at a designated time (point prevalence) or during a specified period (period prevalence)

Randomised controlled trial ~ An epidemiologic experiment in which subjects in a population are randomly allocated into groups to receive or not receive an experimental preventive or therapeutic procedure, manoeuvre or intervention. The groups are compared prospectively. RCTs are generally regarded as the most scientifically rigorous method of hypothesis testing available in epidemiology.

Reference standard ~ An independently applied test that is compared to a screening or diagnostic test being evaluated in order to verify the latter's accuracy. A reference standard therefore provides an accurate or "truth" diagnosis for verification of positive and negative diagnoses. It is sometimes described as providing "final truth determination".

Selection bias ~ Error due to systematic differences in characteristics between those who are selected for inclusion in a study and those who are not (or between those compared within a study and those who are not).

Sensitivity ~ The probability of a positive test result in the presence of abnormality (which the test is designed to detect). See Appendix 1 for calculation.

Specificity ~ The ability of a screening test to correctly identify a person who is free of abnormality. See Appendix 1 for calculation.

Split-sample studies ~ In the context of this report, split-sample studies are those in which a slide is prepared in the conventional way by Pap smear *followed by* the transfer of remaining cells on the sampling instrument to a vial of liquid which is subsequently used to prepare a liquid-based screening slide. The split-sample study therefore represents a within-subjects, "matched pair" design.

Squamous cell carcinoma ~ The most common form of cervical cancer arising from the squamous cells in the epithelium (tissues which line the vagina and the outer layer of the cervix).

Systematic review ~ Literature review reporting a systematic method to search for, identify and appraise a number of independent studies.

True negative smear ~ A smear that is reported as negative, from a woman who does not have a high-grade cervical lesion diagnosed by biopsy within two years of having that smear.

True positive smear ~ A smear that is reported to have a high-grade abnormality, in a woman in whom investigation, including biopsy, within six months of the smear, confirms an abnormality.

Unsatisfactory smear ~ A smear that for technical reasons cannot be reported on by the laboratory (also known as an inadequate smear in the UK).

This glossary was prepared with reference to Drummond et al., (1997), Adams (1991), Last (1995) and Australian Health Technology Advisory Committee (1998).

Appendix 1

CALCULATION OF TEST CHARACTERISTICS

Sensitivity is the probability of a positive test result in the presence of abnormality (which the test is designed to detect).

		Reference Standard	
		Positive	Negative
Screening Test	Positive	a	b
	Negative	c	d

Where a = true positives, b = false positives, c = false negatives, and d = true negatives.

$$\text{Sensitivity} = \frac{a}{a + c} \times 100$$

$$\text{Specificity} = \frac{d}{b + d} \times 100$$

$$\text{PPV} = \frac{a}{a + b} \times 100$$

Sensitivity is calculated as the number of true positives (a) divided by the sum of true positives and false negatives (a + c) and multiplied by 100 (see above).

Specificity is the ability of a screening test to correctly identify a person who is free of abnormality. It is calculated as the number of true negatives (diagnosed as negative by the screening test) and the reference standard (d) divided by the sum of all those diagnosed who do not have the disease - true negatives and false positives (b + d) multiplied by 100.

Appendix 2

CALCULATION OF RELATIVE TRUE POSITIVE RATES AND RELATIVE FALSE POSITIVE RATES

Sampling scheme when only test positives on either test are verified by comparison with a reference standard (Chock et al., 1997).

	Reference standard positive			Reference standard negative		
	Test1+	Test1-	Total	Test1+	Test1-	Total
Test2+	a	b	a + b	A	B	A + B
Test2-	c	{d}	{c + d}	C	{D}	{C + D}
Total	a + c	{b + d}	{n}	A + C	{B + D}	{N}

Variables in brackets { } are unknown as concordant negatives are not verified

Relative TPR (Test 2: Test 1) = $(a + b)/(a + c)$, and relative FPR (Test 2: Test 1) = $(A + B)/(A + C)$.

Appendix 3

SEARCH STRATEGIES

- Search was limited to the years 1997 onwards.
- There was no exclusion by language.
- The original searches were performed in October 1999.
- The *Current Contents* and *Science Citation Index* searches were repeated in February, April, and May 2000 to locate articles which had been published subsequently.
- Filters for study design (e.g. randomised controlled trials) were not included in the strategies as the number of references was small enough to be screened manually for relevant study designs.
- Note that Papnet and AutoCyte Screen were initially included in search strategies although these devices were excluded from the review at the study selection stage once it was determined that they were no longer commercially available.

Medline

vaginal smears/is,mt,st
 limit 1 to yr=1997-1999
 from 2 keep (SELECTED REFERENCES)
 cervix neoplasms/di
 limit 4 to yr=1997-1999
 5 not 2
 (letter or news).pt.
 6 not 7
 from 8 keep (SELECTED REFERENCES)
 cervix neoplasms/
 mass screening/ec,mt,st
 10 and 11
 limit 12 to yr=1997-1999
 2 or 8
 13 not 14
 15 not 7
 from 16 keep (SELECTED REFERENCES)
 exp cytological techniques/
 image processing, computer-assisted/
 10 and 19
 vaginal smears/
 18 and 21
 4 and 18
 20 or 22 or 23
 14 or 16
 24 not 25
 limit 26 to yr=1997-1999
 27 not 7
 from 28 keep (SELECTED REFERENCES)
 17 or 29
 3 or 9 or 17 or 29 or 30

Healthstar

vaginal smears/is,mt,st
 limit 1 to yr=1997-1999
 cervix neoplasms/di

limit 3 to yr=1997-1999
 4 not 2
 (letter or news).pt.
 5 not 6
 cervix neoplasms/
 mass screening/ec,mt,st
 8 and 9
 limit 10 to yr=1997-1999
 2 or 7
 11 not 12
 13 not 6
 exp cytological techniques/
 image processing, computer-assisted/
 8 and 16
 vaginal smears/
 15 and 18
 3 and 15
 17 or 19 or 20
 12 or 14
 21 not 22
 limit 23 to yr=1997-1999
 24 not 6
 2 or 11 or 17 or 20 or 25
 limit 26 to nonmedline
 limit 27 to yr=1997-1999
 from 28 keep (SELECTED REFERENCES)

Embase

uterine cervix cytology/
 mass screening/
 1 and 2
 vagina smear/
 2 and 4
 ((vagina: or cervi:) and (screen: or smear:)).ti,ab,sh.
 ((pap or papanicolaou or papnet) and (screen: or smear:)).ti,ab,sh.
 papanicolaou test/
 cancer screening/
 1 and 9
 4 and 9
 8 and 9
 (automat: or rapid:).ti,ab.
 7 and 13
 3 or 5 or 10 or 11 or 12 or 14
 limit 15 to yr=1997-1999
 6 or 7
 13 and 17
 method:.ti,ab.
 17 and 19
 18 or 20
 limit 21 to yr=1997-1999
 16 or 22
 letter/
 case report/
 24 or 25
 23 not 26
 (labor or fetus).ti,ab,sh.
 27 not 28
 from 29 keep (SELECTED REFERENCES)

Current Contents

- 001 cervical smear:.mp.
- 002 vaginal smear:.mp.
- 003 cervical screen:.mp.
- 004 pap smear:.mp.
- 005 papanicolaou.mp.
- 006 or/1-5
- 007 automat:.mp.
- papnet.mp.
- computer assist:.mp.
- 010 rapid.mp.
- 011 or/7-10
- 012 6 and 11
- 013 from 12 keep (SELECTED REFERENCES)

Other databases were searched using combinations of the following keywords:

thinprep, papanicolaou, autopap, papnet, autocyte prep, autocyte screen, cytorich, rescreen*, ((vagina* or cervi*) near (screen* or smear*)), cervi* near cytology, pap* near (screen* or smear*)

Follow-up searches were done on all databases for the words:

thinprep, papnet, autopap, autocyte screen, cytorich, autocyte prep, rescreen* AND (cervi* or vagina*) near (screen* or smear* or cytology)

Appendix 4

RETRIEVED STUDIES EXCLUDED FOR REVIEW

- Bibbo, M., & Hawthorne, C. (1999). Performance of the AutoPap primary screening system at Jefferson University Hospital. *Acta Cytologica*, 43, 27-9.
- Bibbo, M., Hawthorne, C., & Zimmerman, B. (1999). Does use of the AutoPap assisted primary screener improve cytologic diagnosis? *Acta Cytologica*, 43, 23-6.
- Bishop, J. W. (1997). The cost of production in cervical cytology: comparison of conventional and automated primary screening systems. *American Journal of Clinical Pathology*, 107, 445-50.
- Chevront, D. A., Elston, R. J., & Bishop, J. W. (1998). Effect of a thin-layer preparation system on workload in a cytology laboratory. *Laboratory Medicine*, 29, 174-9.
- Corkill, M., Knapp, D., & Hutchinson, M. (1998). Improved accuracy for cervical cytology with the ThinPrep method and the endocervical brush-spatula collection procedure. *Journal of Lower Genital Tract Disease*, 2, 12-6.
- Dupree, W. B., Suprun, H. Z., Beckwith, D. G., Shane, J. J., & Lucente, V. (1998). The promise and risk of a new technology: The Lehigh Valley Hospital's experience with liquid-based cervical cytology. *Cancer*, 84, 202-7.
- Fetterman, B., Pawlick, G., Koo, H., Hartinger, J., Gilbert, C., & Connell, S. (1999). Determining the utility and effectiveness of the NeoPath AutoPap 300 QC System used routinely. *Acta Cytologica*, 43, 13-22.
- Grant, C. M. (1999). Cervical screening interval: costing the options in one health authority. *Journal of Public Health Medicine*, 21, 140-4.
- Grohs, D. H. (1998). Impact of automated technology on the cervical cytologic smear. A comparison of cost. *Acta Cytologica*, 42, 165-70.
- Guidos, B. J., & Selvaggi, S. M. (1999). Use of the Thin Prep Pap Test in clinical practice. *Diagnostic Cytopathology*, 20, 70-3.
- Howell, L. P., Davis, R. L., Belk, T. I., Agdigos, R., & Lowe, J. (1998). The AutoCyte preparation system for gynecologic cytology. *Acta Cytologica*, 42, 171-7.
- Huang, T. W., Lin, T.-M., & Lee, J.-J. (1999). Sensitivity studies of AutoPap(TM) system location-guided screening of cervical-vaginal cytologic smears. *Acta Cytologica*, 43, 363-8.
- Johnson, T., Maksem, J. A., Belsheim, B. L., Roose, E. B., Klock, L. A., & Eatwell, L. (2000). Liquid-based cervical-cell collection with brushes and wooden spatulas: a comparison of 100 conventional smears from high-risk women to liquid-fixed cytocentrifuge slides, demonstrating a cost-effective, alternative monolayer slide preparation method. *Diagnostic Cytopathology*, 22, 86-91.
- Lee, J. S., Kuan, L., Oh, S., Patten, F. W., & Wilbur, D. C. (1998). A feasibility study of the AutoPap system location-guided screening. *Acta Cytologica*, 42, 221-6.
- Lee, J. S., Wilhelm, P., Kuan, L., Ellison, D. G., Lei, X., Oh, S., & Patten, S. F., Jr. (1997). AutoPap system performance in screening for low prevalence and small cell abnormalities. *Acta Cytologica*, 41, 56-64.

- Linder, J. (1997). Liquid-based cytology: comparison of ThinPrep 2000 with conventionally prepared Pap smears. In E. Franco & J. Monsonogo (Eds.), *New developments in cervical cancer screening and prevention* (pp. 284-93). Oxford: Blackwell Science.
- Marshall, C. J., Rowe, L., & Bentz, J. S. (1999). Improved quality-control detection of false-negative Pap smears using the Autopap 300 QC system. *Diagnostic Cytopathology*, 20, 170-4.
- Patten, S. F., Jr., Lee, J. S., Wilbur, D. C., Bonfiglio, T. A., Colgan, T. J., Richart, R. M., Cramer, H., & Moinuddin, S. (1997a). The AutoPap 300 QC System multicenter clinical trials for use in quality control rescreening of cervical smears: I. A prospective intended use study. *Cancer*, 81, 337-42.
- Patten, S. F., Jr., Lee, J. S., Wilbur, D. C., Bonfiglio, T. A., Colgan, T. J., Richart, R. M., Cramer, H., & Moinuddin, S. (1997b). The AutoPap 300 QC System multicenter clinical trials for use in quality control rescreening of cervical smears: II. Prospective and archival sensitivity studies. *Cancer*, 81, 343-7.
- Raab, S. S. (1997). The cost-effectiveness of cervical-vaginal rescreening. *American Journal of Clinical Pathology*, 108, 525-36.
- Radensky, P. W., & Mango, L. J. (1998). Interactive neural-network-assisted screening. An economic assessment. *Acta Cytologica*, 42, 246-52.
- Richart, R. M., Patten, S. F. J., & Lee, L. J. S. (1997). Automated screening using the AutoPap 300 device. In E. Franco & J. Monsonogo (Eds.), *New developments in cervical cancer screening and prevention* (pp. 279-83). Oxford: Blackwell Science.
- Sherman, M. E., Mendoza, M., Lee, K. R., Ashfaq, R., Birdsong, G. G., Corkill, M. E., McIntosh, K. M. et al. (1998). Performance of liquid-based, thin-layer cervical cytology: correlation with reference diagnoses and human papillomavirus testing. *Modern Pathology*, 11, 837-43.
- Sherman, M. E., Schiffman, M. H., Lorincz, A. T., Herrero, R., Hutchinson, M. L., Bratti, C., Zahniser, D. et al. (1997). Cervical specimens collected in liquid buffer are suitable for both cytologic screening and ancillary human papillomavirus testing. *Cancer*, 81, 89-97.
- Stevens, M. W., Milne, A. J., James, K. A., Brancheau, D., Ellison, D., & Kuan, L. (1997). Effectiveness of automated cervical cytology rescreening using the AutoPap 300 QC System. *Diagnostic Cytopathology*, 16, 505-12.
- Stevens, M. W., Nespolon, W. W., Milne, A. J., & Rowland, R. (1998). Evaluation of the CytoRich technique for cervical smears. *Diagnostic Cytopathology*, 18, 236-42.
- Takahashi, M., & Naito, M. (1997). Application of the CytoRich monolayer preparation system for cervical cytology. A prelude to automated primary screening. *Acta Cytologica*, 41, 1785-9.
- Vassilakos, P., Griffin, S., Megevand, E., & Campana, A. (1998). CytoRich liquid-based cervical cytologic test. Screening results in a routine cytopathology service. *Acta Cytologica*, 42, 198-202.
- Vassilakos, P., Saurel, J., & Rondez, R. (1999). Direct-to-vial use of the AutoCyte PREP liquid-based preparation for cervical-vaginal specimens in three European laboratories. *Acta Cytologica*, 43, 65-8.
- Wertlake, P. (1999). Results of AutoPap system-assisted and manual cytologic screening. A comparison. *Journal of Reproductive Medicine*, 44, 11-7.
- Wilbur, D. C., Facik, M. S., Rutkowski, M. A., Mulford, D. K., & Atkison, K. M. (1997). Clinical trials of the CytoRich specimen-preparation device for cervical cytology. Preliminary results. *Acta Cytologica*, 41, 24-9.

Appendix 5

ADDITIONAL BACKGROUND PAPERS

- Abulafia, O., & Sherer, D. M. (1999). Automated cervical cytology: meta-analyses of the performance of the PAPNET system. *Obstetrical & Gynecological Survey*, 54, 253-64.
- Anderson, J. M., Akkerman, D., Barrette, B., LaFerla, J., Williams, J., Fay, C., Charlton, A. J., & al., e. (1999). Health care guideline: cervical cancer screening (Pap smears). Bloomington, MN: Institute for Clinical Systems Improvement.
- Ashfaq, R., Saliger, F., Solares, B., Thomas, S., Liu, G., Liang, Y., & Saboorian, M. H. (1997). Evaluation of the PAPNET system for prescreening triage of cervicovaginal smears. *Acta Cytologica*, 41, 1058-64.
- Austin, R. M. (1999). Who should decide how effective cervical cancer screening will be? *Acta Cytologica*, 43, 4-6.
- Australia. National Pathology Accreditation Advisory Council (1997). *Requirements for gynaecological (cervical) cytology*. Canberra: National Pathology Accreditation Advisory Council.
- Baird, P. J. (1997). Evaluation of the PAPNET system in a general pathology service. *Medical Journal of Australia*, 167, 285.
- Bartels, P. H., & Wied, G. L. (1997). Automated screening for cervical cancer: diagnostic decision procedures. *Acta Cytologica*, 41, 6-10.
- Bishop, J. W., Chevront, D. A., & Elston, R. J. (1999a). Utility of residual AutoCyte cervical cytology samples for image analysis. *Acta Cytologica*, 43, 39-46.
- Bishop, J. W., Kaufman, R. H., & Taylor, D. A. (1999b). Multicenter comparison of manual and automated screening of AutoCyte gynecologic preparations. *Acta Cytologica*, 43, 34-8.
- Boon, M. E. (1997a). Cooperation of the image analyser and the cytologist. In E. Franco & J. Monsonego (Eds.), *New developments in cervical cancer screening and prevention* (pp. 311-6). Oxford: Blackwell Science.
- Boon, M. E. (1997b). The history of neural network technology in cytology. In E. Franco & J. Monsonego (Eds.), *New developments in cervical cancer screening and prevention* (pp. 294-305). Oxford: Blackwell Science.
- Boronow, R. C. (1998). Death of the Papanicolaou smear? A tale of three reasons. *American Journal of Obstetrics & Gynecology*, 179, 391-6.
- Brotzman, G. L., Kretzchmar, S., Ferguson, D., Gottlieb, M., & Stowe, C. (1999). Costs and outcomes of PAPNET secondary screening technology for cervical cytologic evaluation. A community hospital's experience. *Archives of Family Medicine*, 8, 52-5.
- Bryant, J., & Stevens, A. (1995). Cervical screening interval: the development and evaluation Committee Report No. 46. Southampton: Wessex Institute of Public Health Medicine Development and Evaluation Committee.
- Byrnes, E. C. (1998). A new age in Pap testing. *Advance for Nurse Practitioners*, 6, 65-6, 92.

- Cenci, M., Nagar, C., Giovagnoli, M. R., & Vecchione, A. (1997). The PAPNET system for quality control of cervical smears: validation and limits. *Anticancer Research*, 17, 4731-4.
- Coleman, D., & PRISMATIC Project Management Team (1999). Assessment of automated primary screening on PAPNET of cervical smears in the PRISMATIC trial. PRISMATIC Project Management Team. *Lancet*, 353, 1381-5.
- Cuzick, J. (2000). Human papillomavirus testing for primary cervical cancer screening. *JAMA*, 283, 108-9.
- Doornewaard, H., van der Schouw, Y. T., van der Graaf, Y., Bos, A. B., Habbema, J. D. F., & van den Tweel, J. G. (1999). The diagnostic value of computer-assisted primary cervical smear screening: a longitudinal cohort study. *Modern Pathology*, 12, 995-1000.
- Doornewaard, H., Woudt, J. M., Strubbe, P., van de Seijp, H., & van den Tweel, J. G. (1997). Evaluation of PAPNET-assisted cervical rescreening. *Cytopathology*, 8, 313-21.
- Duggan, M. A. (2000). Papnet-assisted, primary screening of cervico-vaginal smears. *European Journal of Gynaecological Oncology*, 21, 35-42.
- Duggan, M. A., & Brasher, P. (1997). Paired comparison of manual and automated Pap test screening using the PAPNET system. *Diagnostic Cytopathology*, 17, 248-54.
- Farnsworth, A., Chambers, F. M., & Goldschmidt, C. S. (1997). Evaluation of the PAPNET system in a general pathology service. *Medical Journal of Australia*, 167, 285-6.
- Follen, M., & Richards-Kortum, R. (2000). Emerging technologies and cervical cancer. *Journal of the National Cancer Institute*, 92, 363-5.
- Franco, E., & Monsonego, J. (1997). *New developments in cervical cancer screening and prevention*. Oxford: Blackwell Science.
- Franco, E. L., & Ferenczy, A. (1999). Assessing gains in diagnostic utility when human papillomavirus testing is used as an adjunct to papanicolaou smear in the triage of women with cervical cytologic abnormalities. *American Journal of Obstetrics & Gynecology*, 181, 382-6.
- Garber, A. M. (1998a). Making the most of Pap testing. *JAMA*, 279, 240-1.
- Garber, A. M. (1998b). Rescreening of cervical Papanicolaou smears using PAPNET. *JAMA*, 279, 1788.
- Ghidoni, D., Fabbris, E., Folicaldi, S., Amadori, A., Medri, M., Bucchi, L., & Bondi, A. (1998). Accuracy comparison between PAPNET diagnoses and conventional diagnoses in an Italian cervical cytology laboratory. *Diagnostic Cytopathology*, 19, 279-83.
- Gill, G. W. (1997). Pap smear risk management by process control. *Cancer*, 81, 198-211.
- Godfrey, S. E. (1999). The pap smear, automated rescreening, and negligent nondisclosure. *American Journal of Clinical Pathology*, 111, 14-7.
- Gottlieb, S. (1999). Papillomavirus DNA in smear test raises risk of cervical cancer. *BMJ*, 319, 1454.
- Greenberg, M. (1998). Rescreening of cervical Papanicolaou smears using PAPNET. *JAMA*, 279, 1785-6.
- Gurley, A. M., Roberts, J. M., Thurloe, J. K., Bowditch, R., & Laverty, C. R. A. (1998). Better screening methods are available. *Medical Journal of Australia*, 168, 254.

- Hailey, D. M., & Lea, R. (1995). Prospects for newer technologies in cervical cancer screening programmes. *Journal of Quality in Clinical Practice*, 15, 139-45.
- Halford, J. A., Wright, R. G., & Ditchmen, E. J. (1997). Quality assurance in cervical cytology screening. Comparison of rapid rescreening and the PAPNET Testing System. *Acta Cytologica*, 41, 79-81.
- Halford, J. A., Wright, R. G., & Ditchmen, E. J. (1999). Prospective study of PAPNET: review of 25 656 pap smears negative on manual screening and rapid rescreening. *Cytopathology*, 10, 317-23.
- Howell, L. P., Belk, T., Agdigos, R., Davis, R., & Lowe, J. (1999). AutoCyte Interactive Screening System. Experience at a university hospital cytology laboratory. *Acta Cytologica*, 43, 58-64.
- Husain, O. A., Kocjan, G., Butler, E. B., & McGloin, J. E. (1997). PAPNET. The human and other dimensions. *Acta Cytologica*, 41, 1439-44.
- Intersociety Working Group for Cytology Technologies (1998). A proposed methodology for evaluating secondary screening (rescreening) instruments for gynecologic cytology. Intersociety Working Group for Cytology Technologies. *Acta Cytologica*, 42, 1311-4.
- Jenny, J., Isenegger, I., Boon, M. E., & Husain, O. A. (1997). Consistency of a double PAPNET scan of cervical smears. *Acta Cytologica*, 41, 82-7.
- Kaufman, R. H., Schreiber, K., & Carter, T. (1998). Analysis of atypical squamous (glandular) cells of undetermined significance smears by neural network-directed review. *Obstetrics and Gynecology*, 91, 556-60.
- Kemp, R. A., MacAulay, C., Garner, D., & Palcic, B. (1997). Detection of malignancy associated changes in cervical cell nuclei using feed-forward neural networks. *Analytical Cellular Pathology*, 14, 31-40.
- Kok, M. R., Boon, M. E., Schreiner-Kok, P. G., & Koss, L. G. (2000). Cytological recognition of invasive squamous cancer of the uterine cervix: comparison of conventional light-microscopical screening and neural network-based screening. *Human Pathology*, 31, 23-8.
- Kok, M. R., Habers, M. A., Schreiner-Kok, P. G., & Boon, M. E. (1998). New paradigm for ASCUS diagnosis using neural networks. *Diagnostic Cytopathology*, 19, 361-6.
- Koss, L. G. (1997). Automation in cervicovaginal cytology: system requirements and benefits. In E. Franco & J. Monsonogo (Eds.), *New developments in cervical cancer screening and prevention* (pp. 274-8). Oxford: Blackwell Science.
- Koss, L. G., Sherman, M. E., Cohen, M. B., Anes, A. R., Darragh, T. M., Lemos, L. B., McClellan, B. J. et al. (1997). Significant reduction in the rate of false-negative cervical smears with neural network-based technology (PAPNET Testing System). *Human Pathology*, 28, 1196-203.
- Koss, L. J. (1998). Rescreening of cervical Papanicolaou smears using PAPNET. *JAMA*, 279, 1786.
- Krieger, P., & Bibbo, M. (1997). Our journey towards improved accuracy in cytology: the role of new technologies. *Acta Cytologica*, 41, 11-4.
- Krieger, P., & Naryshkin, S. (1997). Despite potential flaws, the false-negative proportion remains the best practical measure of the accuracy of cervical cytology screening. *Cancer*, 81, 261-3.
- Ku, N. N. (1999). Automated Papanicolaou smear analysis as a screening tool for female lower genital tract malignancies. *Current Opinion in Obstetrics & Gynecology*, 11, 41-3.

- Lemay, C., & Meisels, A. (1999). 100% rapid (partial) rescreening for quality assurance. *Acta Cytologica*, 43, 86-8.
- Lerma, E., Colomo, L., Carreras, A., Esteva, E., Quilez, M., & Prat, J. (1998). Rescreening of atypical cervicovaginal smears using PAPNET. *Cancer*, 84, 361-5.
- Linder, J. (1997). Automation of the Papanicolaou smear: a technology assessment perspective. *Archives of Pathology & Laboratory Medicine*, 121, 282-6.
- Linder, J., & Zahniser, D. (1997). The ThinPrep Pap test. A review of clinical studies. *Acta Cytologica*, 41, 30-8.
- Losell, K., & Dejmek, A. (1999). Comparison of papnet-assisted and manual screening of cervical smears. *Diagnostic Cytopathology*, 21, 296-9.
- Maksem, J. A. (1999). Liquid-based cytology: where do we go from here?. *Diagnostic Cytopathology*, 21, 79-80.
- Mango, L. J. (1997). Clinical validation of interactive cytologic screening. Automating the search, not the interpretation. *Acta Cytologica*, 41, 93-7.
- Mango, L. J. (1998). Neural network-assisted cervical cancer screening. *Journal of Clinical Ligand Assay*, 21, 203-7.
- Mango, L. J., & Radensky, P. W. (1998a). Interactive neural network-assisted screening: a clinical assessment. *Acta Cytologica*, 42, 233-45.
- Mango, L. J., & Radensky, P. W. (1998b). Rescreening of cervical Papanicolaou smears using PAPNET. *JAMA*, 279, 1786-7.
- Mango, L. J., & Valente, P. T. (1998). Neural-network-assisted analysis and microscopic rescreening in presumed negative cervical cytologic smears. A comparison. *Acta Cytologica*, 42, 227-32.
- Masood, S., Cajulis, R. S., Cibas, E. S., Wilbur, D. C., & Bedrossian, C. W. (1998). Automation in cytology: a survey conducted by the New Technology Task Force, Papanicolaou Society of Cytopathology. *Diagnostic Cytopathology*, 18, 47-55.
- McGoogan, E. (1997). Advantages and limitations of automated screening systems in developing and developed countries. In E. Franco & J. Monsonego (Eds.), *New developments in cervical cancer screening and prevention* (pp. 317-22). Oxford: Blackwell Science.
- McNeil, C. (1997). New pap test technologies embark on shifting seas. *Journal of the National Cancer Institute*, 89, 410-2.
- Meijer, C. J. L. M., & Walboomers, J. M. M. (2000). Cervical cytology after 2000: where to go? *Journal of Clinical Pathology*, 53, 41-3.
- Melnikow, J., & Nuovo, J. (1999). Reducing mortality due to cervical cancer. PAPNET fails the test. *Archives of Family Medicine*, 8, 56-7.
- Michelow, P. M., Hlongwane, N. F., & Leiman, G. (1997). Simulation of primary cervical cancer screening by the PAPNET system in an unscreened, high-risk community. *Acta Cytologica*, 41, 88-92.
- Miller, A. B. (1992). The cost effectiveness of cervical cancer screening. *Annals of Internal Medicine*, 117, 529-30.
- Mitchell, H., & Medley, G. (1998a). Detection of laboratory false negative smears by the PAPNET cytologic screening system. *Acta Cytologica*, 42, 265-70.

- Mitchell, H., & Medley, G. (1998b). Detection of unsuspected abnormalities by PAPNET-assisted review. *Acta Cytologica*, 42, 260-4.
- Mitchell, H., & Medley, G. (1998c). Differences between false-negative and true-positive Papanicolaou smears on Papnet-assisted review. *Diagnostic Cytopathology*, 19, 138-40.
- Mody, D. R. (1999). Agonizing over AGUS. *Cancer Cytopathology*, 87, 243-4.
- Noorani, H. Z., Arratoon, C., & Hall, A. (1997). Assessment of techniques for cervical cancer screening. Ottawa: Canadian Coordinating Office for Health technology Assessment (CCOHTA).
- O'Leary, T. J., Buckner, S.-B., & Ollayos, C. W. (1998a). Rescreening of cervical Papanicolaou smears using PAPNET. *JAMA*, 279, 1787-8.
- O'Leary, T. J., Tellado, M., Buckner, S. B., Ali, I. S., Stevens, A., & Ollayos, C. W. (1998b). PAPNET-assisted rescreening of cervical smears: cost and accuracy compared with a 100% manual rescreening strategy. *JAMA*, 279, 235-7.
- Paul, C. (2000). Internal and external morality of medicine: lessons from New Zealand. *BMJ*, 320, 499-503.
- Reid, R. I. (1998). Conventional pap smears are unreliable. *Medical Journal of Australia*, 168, 252.
- Roberts, J. M., Bowditch, R., Gurley, A. M., Laverty, C. R. A., & Thurloe, J. K. (1998). Women have a right to the most accurate results. *Medical Journal of Australia*, 168, 252-3.
- Rosenthal, D. L. (1997). Computerized scanning devices for Pap smear screening: current status and critical review. *Clinics in Laboratory Medicine*, 17, 263-84.
- Ryan, M. R., Stasny, J. F., Remmers, R., Pedigo, M. A., Cahill, L. A., & Frable, W. J. (1996). PAPNET-directed rescreening of cervicovaginal smears: a study of 101 cases of atypical squamous cells of undetermined significance. *American Journal of Clinical Pathology*, 105, 711-8.
- Sawaya, G. F., & Washington, A. E. (1999). Cervical cancer screening: which techniques should be used and why? *Clinical Obstetrics and Gynecology*, 42, 922-38.
- Schechter, C. B. (1998). Rescreening of cervical Papanicolaou smears using PAPNET. *JAMA*, 279, 1787.
- Sherman, M. E. (1997). A comparison of automated and manual screening: theoretical considerations. In E. Franco & J. Monsonogo (Eds.), *New developments in cervical cancer screening and prevention* (pp. 306-10). Oxford: Blackwell Science.
- Sherman, M. E., Schiffman, M. H., Mango, L. J., Kelly, D., Acosta, D., Cason, Z., Elgert, P. et al. (1997). Evaluation of PAPNET testing as an ancillary tool to clarify the status of the "atypical" cervical smear. *Modern Pathology*, 10, 564-71.
- Smith, W. J. (1999). The cost-effectiveness of cervical screening. *Current Opinion in Obstetrics & Gynecology*, 11, 83-5.
- Solomon, H. M., & Frist, S. (1998). PAPNET testing for HSILs. The few cell/small cell challenge. *Acta Cytologica*, 42, 253-9.
- Spitzer, M. (1998). Cervical screening adjuncts: Recent advances. *Current Problems in Obstetrics, Gynecology & Fertility*, 21, 176-190.

- Stanley, M. W. (1998). Automated primary screening for gynecologic cytology: the time has not yet come. *American Journal of Clinical Pathology*, 109, 6-9.
- Stastny, J. F., Remmers, R. E., London, W. B., Pedigo, M. A., Cahill, L. A., Ryan, M., & Frable, W. J. (1997). Atypical squamous cells of undetermined significance: a comparative review of original and automated rescreen diagnosis of cervicovaginal smears with long term follow-up. *Cancer*, 81, 348-53.
- Sturgis, C. D., Isoe, C., McNeal, N. E., Yu, G. H., & DeFrias, D. V. (1998). PAPNET computer-aided rescreening for detection of benign and malignant glandular elements in cervicovaginal smears: a review of 61 cases. *Diagnostic Cytopathology*, 18, 307-11.
- Takahashi, M., Kimura, M., Akagi, A., & Naitoh, M. (1998). AutoCyte SCREEN interactive automated primary cytology screening system. A preliminary evaluation. *Acta Cytologica*, 42, 185-8.
- United States. Center for Devices and Radiological Health (1999). FDA classification of medical devices. Rockville, MD: CDRH Available from:
<http://www.fda.gov/cdrh/dsma/dsmaclas.html>.
- van Ballegooijen, M., Beck, S., Boon, M. E., Boer, R., & Habbema, J. D. (1998). Rescreen effect in conventional and PAPNET screening: observed in a study using material enriched with positive smears. *Acta Cytologica*, 42, 1133-8.
- Veneti, S., Papaefthimiou, M., Symiakaki, H., & Ioannidou-Mouzaka, L. (1999). PAPNET for cervical cytology screening. Experience in Greece. *Acta Cytologica*, 43, 30-3.
- Vooijs, G. P. (1998). On the performance and cost-effectiveness of semiautomated Papanicolaou smear screening. *Cancer*, 84, 269-72.
- Wain, G. V. (1998). What cost will society accept?. *Medical Journal of Australia*, 168, 253.
- Walsh, J. M. (1998). Cervical cancer: developments in screening and evaluation of the abnormal Pap smear. *Western Journal of Medicine*, 169, 304-10.
- Wright, F. G. (1998). PAPNET superior to rapid rescreening. *Medical Journal of Australia*, 168, 253.
- Zardawi, I. (1998). Increase pap smear uptake. *Medical Journal of Australia*, 168, 252.